Contributions to unsupervised domain adaptation : Similarity functions, optimal transport and theoretical guarantees

Sofiane Dhouib

N°d'ordre NN
2020LYSEI117

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
L'Institut National des Sciences Appliquées de Lyon

Ecole Doctorale N° 160
Electronique, Electrotechnique, Automatique (EEA)

Spécialité/ discipline de doctorat :
Traitement du signal et de l'image

Soutenue publiquement le 23/11/2020, par :
Sofiane Dhouib

Contributions to unsupervised domain adaptation: similarity functions, optimal transport and theoretical guarantees

---

Devant le jury composé de :

Ben-David, Shai	Professeur des Universités	Université de Waterloo	Rapporteur
Flamary, Rémi	Maître de Conférences HDR	Ecole Polytechnique	Rapporteur
Sebag, Michèle	Directrice de Recherche	CNRS	Examinatrice
Lartizien, Carole	Directrice de Recherche	CNRS	Directrice de thèse
Redko, Ievgen	Maître de Conférences	Université Jean Monnet	Co-directeur de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr** INSA : R. GOURDON	**M. Stéphane DANIELE** Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec. : M.C. HAVGOUDOUKIAN **ecole-doctorale.eea@ec-lyon.fr**	**M. Gérard SCORLETTI** École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 **gerard.scorletti@ec-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES **secretariat.e2m2@univ-lyon1.fr**	**M. Philippe NORMAND** UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://www.ediss-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE **secretariat.ediss@univ-lyon1.fr**	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Curien - 3ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tel : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 **infomaths@univ-lyon1.fr**	**M. Hamamache KHEDDOUCI** Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tel : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **ed.materiaux@insa-lyon.fr**	**M. Jean-Yves BUFFIÈRE** INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 **jean-yves.buffiere@insa-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **mega@insa-lyon.fr**	**M. Jocelyn BONJOUR** INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **http://ed483.univ-lyon2.fr** Sec. : Véronique GUICHARD INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 **veronique.cervantes@univ-lyon2.fr**	**M. Christian MONTES** Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

To the memory of my beloved mother.

Acknowledgements

The accomplishment of this work would not have been possible without many people to whom I am thankful.

I sincerely thank my two Ph.D. referees, Prof. Shai Ben-David and Prof. Rémi Flamary, as well as my PhD inspectors Prof. Michèle Sebag and Dr. Basura Fernando, for spending a remarkable effort in evaluating this work.

Of course, the mentoring I received from both my supervisors, Prof. Carol Lartizien and Dr. Ievgen Redko, was crucial for the fulfillment of this work. I am grateful to Carol for accepting my approach to research during this thesis and for Ievgen, who let me explore any research ideas I had in mind and did all he could to enrich them, leading to our contributions. Moreover, I owe it to both of them as they unhesitatingly supported me morally during the challenging times of this thesis.

I also thank Prof. Marc Sebban for being a member of my PhD monitoring committee and with whom I collaborated for the fifth chapter of this manuscript, along with Prof. Rémi Emonet and PhD student Tanguy Kerdoncuff.

Besides, I thank all of my colleagues, especially Mariem, Denis, Yufei, Kenny, Gia-Thuy, Susanne, Audrey, Daria, and Antonio, for contributing to a fantastic work atmosphere. This latter is also the result of the interesting discussions I had been having with Juan, Emmanuel and Odyssee. I thank Pierre and Fabrice too for always being helpful.

Beyond the professional circle, my feelings of gratefulness go to my beloved family members, to whom I am infinitely indebted: my mother Emna, my father Abderazek, and my sisters Mariem and Syrine, who are continuously believing in me, unconditionally showing their love and support despite the distance and encouraging me to pursue this adventure.

Finally, my special thanks go to my special one, Wissal, as she never quitted standing by my side during countless hardships that I went through.

Abstract

The surge in the quantity of data produced nowadays made of *Machine Learning*, a subfield of Artificial Intelligence, a vital tool used to extract valuable patterns from them and allowed it to be integrated into almost every aspect of our everyday activities. Concretely, a machine learning algorithm learns such patterns after being trained on a dataset called the *training set*, and its performance is assessed on a different set called the *testing set*.

Domain Adaptation is an active research area of machine learning, in which the training and testing sets are not assumed to stem from the same probability distribution, as opposed to *Supervised Learning*. In this case, the two distributions generating the training and testing data correspond respectively to the *source* and *target* domains. While the supervised learning theory relies on the convergence of the empirical distribution of the observed data to its true counterpart for establishing generalization guarantees, these latter are hindered by the shift between distributions and by the lack of labels in the testing set in the case of domain adaptation. Therefore, additional relatedness assumptions between the domains are inevitable in order for the learning process to succeed.

Our contributions focus on three theoretical aspects related to domain adaptation for classification tasks. The first one is learning with similarity functions, which deals with classification algorithms based on comparing an instance to other examples in order to decide its class. The second is large-margin classification, which concerns learning classifiers that maximize the separation between classes. The third is Optimal Transport that formalizes the principle of least effort for transporting probability masses between two distributions.

At the beginning of the thesis, we were interested in learning with so-called (ϵ, γ, τ)-good similarity functions in the domain adaptation framework, since these functions have been introduced in the literature in the classical framework of supervised learning. This is the subject of our first contribution in which we theoretically study the performance of a similarity function on a target distribution, given it is suitable for the source one. Then, we tackle the more general topic of large-margin classification in domain adaptation, with weaker assumptions than those adopted in the first contribution. In this context, we proposed a new theoretical study and a domain adaptation algorithm, which is our second contribution. We derive novel bounds taking the classification margin on the target domain into account, that we convexify by leveraging the appealing *Optimal Transport* theory, in order to derive a domain adaptation algorithm with an adversarial variation of the classic Kantorovich problem. Finally, after noticing that our adversarial formulation can be generalized to include several other cases of interest, we dedicate our last contribution to adversarial or minimax variations of the optimal transport problem, where we demonstrate the versatility of our approach.

Contents

List of Figures

List of Tables

Introduction

With the evolution of information technologies, data nowadays is being generated at a continuously increasing pace. As a matter of example, one can think of the millions of emails, website subscriptions, companies' transactions, and hospitals' medical information of patients stored all around the world every single day. There is an undeniable appeal in extracting valuable patterns from these data: an e-commerce website can propose better recommendations to its customers based on their previous activities, and a hospital can leverage the experiences of present patients to improve those of the future ones. What unifies these different examples is wanting to transform experience gained from data to expertise (Shalev-Shwartz and Ben-David, 2014), which is the main aim of the *Machine Learning* field. In this vast sub-field of the general area of *Artificial Intelligence*, numerous branches exist depending on the possible characteristics given, for instance, by the form of the data, the way it is accessed, the underlying process generating it, and what pattern one is looking seeking.

The work presented in this manuscript concerns offline learning, meaning that data is accessed in one time as a fixed batch (as opposed to online learning) of input-output pairs, and the pattern to search for is a rule that generates the output given the input. Moreover, we focus on the case where the outputs' possible values are finite and represent *labels* or *classes* to which inputs (also called instances) belong. To approximate such a relation, a *hypothesis* or a *classifier* is learned from the available data, *i.e.* it is refined until it matches the observations. The goal behind such a procedure is to be able to make correct predictions on new data not used in the learning process. A fundamental assumption in the theoretical study of the above-mentioned setting is that training and testing data stem from some unknown probability distributions. The question of whether the two sets of data are generated by the same distribution defines two disciplines of machine learning: assuming the same distribution is the case of *Supervised Learning*, whereas allowing the distribution of the testing data to be different defines the more challenging *Domain Adaptation* setting. Supervised learning has been intensively theoretically analyzed in the context of the statistical learning theory, where the main concern is the generalization of a hypothesis learned from a finite sample to the whole generating distribution. The study of domain adaptation is more recent and is motivated by real-world situations where the data generating processes are subject to change, making the same distribution assumption unreasonable, and where the labeling process is time-consuming or costly. In this case, the training and testing distributions correspond respectively to a *source* and a *target* domains. Typically, domain adaptation is convenient for situations where one has access to an unlabeled newly generated test dataset that is much larger than the previously available labeled source data sample. In this case, a part of the unlabeled target data can help the learning process along with the labeled source data. Domain adaptation is an active research area, and the scope of this work is its special case considering only one distribution for the source domain (as opposed to *Multi-source domain adaptation*). The discrepancy between the two domains' distributions, along with the lack of label information for the target data, make the domain adaptation problem much more difficult than supervised learning. Moreover, there is intuitively no hope in trying to learn from two completely unrelated domains: for example, a person who tries to learn a new language from a differ-

ent group than hers cannot succeed without supervision. Conversely, she can successfully guess the meaning of several words of a language related to her own based on the similarity between words. This intuition is reflected in the domain adaptation theory that aims to determine conditions reflecting the relatedness between the two domains and helping to learn despite the distribution's shift. Among these conditions, a low divergence between the source and target domains is the common ground of almost all of the literature of domain adaptation, with variations depending on the choice of this divergence. It is also the goal of the striking majority of domain adaptation algorithms, where the two domains are made close to each other via an alignment procedure, and the closeness, in this case, is contingent on the choice of the divergence measure. One choice that has recently gained in popularity is the *Wasserstein Distance* between probability distributions, associated with the *Optimal Transport* problem. This latter is a formalization of the principle of least efforts to the transport of probability mass between distributions.

Earlier in this introduction, we pointed out that we focus on classification tasks where the possible values of the output are finite. There are several approaches to solve these tasks, among which are those based on the "birds of a feather flock together" intuition, *i.e.* they rely on the similarity of an instance to the rest of the data to decide its class. Two popular algorithms in this regard are the *k Nearest Neighbors* and the *Support Vector Machine* algorithms: the former relies on distances and the latter on special functions called *kernels*, both reflecting resemblance between instances. The two algorithms make use of similarity functions, and these latter are fixed beforehand resulting in their potential failure to capture hidden patterns in the available data. The question of when does a similarity function suit the classification task at hand then arises naturally, and one of the lines of work that answer it is the theory of $(\epsilon, \gamma, \tau)-$good similarity functions. Roughly speaking, these latter require the existence of some distribution generating landmark points, such that most of the instances are more similar on average to landmarks having their label than to the ones with opposite labels. In this case, data can be mapped to a new space when the classes are separable with a large margin.

In this thesis, we address several limitations of the current state of research in domain adaptation for classification problems. First, despite the strong and appealing intuition behind learning with similarity functions, there is a lack of theoretical understanding of these latter in the context domain adaptation. We tackle this limitation in our first contribution where we provide novel results extending the (ϵ, γ, τ) framework to domain adaptation. Second, most of the theoretical results in domain adaptation do not consider the classification margin on the target domain. In fact, they essentially rely on the triangle inequality for the considered loss function, which is not verified for loss functions intended to maximize the classification margin. We address this problem in our second contribution and use our theoretical contribution to propose a sound domain adaptation algorithm where we compare the source and target distributions via an adversarial task-dependent optimal transport term. This latter is studied more generally in our last contribution, in which we solve several instances of the adversarial optimal transport problem and show its practical interest.

Outline

The rest of the manuscript is divided into three parts. The first part, Background, provides the reader with the current state of the art of the different topics that we address and is comprised of two chapters.

Chapter 1 is dedicated to supervised learning, with a focus on large margin binary classification. We briefly present some important results of the statistical learning theory, some reference algorithms and then we give an overview of learning with $(\epsilon, \gamma, \tau)-$good similarity functions.

Chapter 2 addresses the more general domain adaptation field, where we start by defining the specific setting in which we are interested. Then, we present the main assumptions of the domain adaptation theory, along with literature results showing their sufficiency and necessity for the success of adaptation. In particular, we cover several measures of divergence intended to compare the source and target domains.

The second part, Contributions, presents our work based on accepted submissions at several peer-reviewed international conferences. The proofs for the different theoretical results we provide are given either in their corresponding chapters when their length is at most half a page, or are postponed to the following Appendices part. In the latter case, we provide a short description of the proof's idea in the main chapter.

Chapter 3 corresponds to our publications Dhouib and Redko (2018a,b), where we study $(\epsilon, \gamma, \tau)-$good similarity functions in the domain adaptation setting. We establish theoretical results relating the performance of a similarity function on both the source and target domains. Then, we present a retrospective study in which we explain why we abandon the $(\epsilon, \gamma, \tau)-$good similarities framework in the rest of the thesis.

Chapter 4 is based on our publications Dhouib et al. (2019, 2020b), where we weaken the assumptions of the previous chapter and study the performance of a classifier on the target domain while focusing on the quality of separation between classes. We present theoretical results that generalize some former work from the literature and introduce a task-dependent variation of the *Optimal Transport* problem. We then specialize the study to linear classification and empirically show its benefits.

Chapter 5 represents the work leading to our publication Dhouib et al. (2020a). It deals with a min-max variation of the optimal transport problem that is a generalization of the divergence term we obtained in the previous chapter. We propose an optimization method to solve it and detail its variations according to different instances of the considered problem.

The last part is for appendices containing either some prerequisites for reading the manuscript or additional details on different parts of it.

Appendix A recalls some prerequisites that are necessary for reading this manuscript.

Appendices B, C and D provide more material for chapters 3, 4 and 5, respectively, including the proofs of different theoretical claims, as well as some additional details on the empirical evaluations.

Appendix E is based on our submission to the Machine Learning journal, in which we theoretically study learning an $(\epsilon, \gamma, \tau)-$good similarity function via regression, and which led us to the retrospective study presented in Chapter 3.

Appendix F encloses a summary of the current manuscript in French, as required by the doctoral school EEA.

Notations

For the sake of readability and to allow a quicker recognition of the different types of quantities used, we adopt the following notation conventions. The sets known mathematically as spaces (*e.g.* known vector spaces, hypothesis spaces...) are denoted in a `\mathbb` font. The bold font is exclusively used for vectors and matrices, which are respectively denoted by lower- and upper-case Latin letters. Scalars are denoted either by Latin or Greek lower-case letters. Moreover, probability distributions are denoted using only the

\mathcal font. These conventions, in addition to other notations, are summarized in the next table.

Part I

Background

Chapter 1

Supervised Learning

Abstract Supervised Learning is arguably one of the most known machine learning settings. In this chapter, we present this setting with a particular emphasis on the task of binary classification. We start by formalizing notions that are needed to define a supervised learning problem and the performance measure used to assess whether the learning in this context is successful. Then, we focus on supervised classification where the output can only take a finite number of values. In particular, we highlight the role of scoring functions and their associated classification margin to further review several generalization bounds linking the empirical and true performances of a classifier at hand. This is followed by a description of two famous classification algorithms, namely the Support Vector Machine and the k-nearest neighbors. Finally, after pointing out the role of similarity functions in both of the presented algorithms, we review the general theory of learning with $(\epsilon, \gamma, \tau)-$good similarity functions and link it to both kernel learning and metric learning frameworks.

Introduction

In this chapter, we introduce several notions related to supervised learning, a branch of machine learning field that formalizes the idea of learning by example. Such learning consists in inducing a relation between elements of a certain set and outputs associated to them, based on observing examples of input-output pairs. These latter pairs are in general the result of a certain experiment, where the outputs correspond to measurements and the input to their respective configurations. The relation is commonly referred to as a hypothesis, as one who observes such data puts forward a hypothesis on the mechanism that links an input to its output. The complexity and the difficulty of describing such a mechanism in certain cases of interest is what motivated the emergence of supervised learning. For example, it is easy for an average person to recognize a previously known object on a given image, even if the latter is encountered in a completely new configuration with a change of perspective, of lights, or of orientation. However, attempting to characterize any image of an object by a set of predetermined explicit rules would most likely be a very difficult endeavor. Likewise, the expertise of a house pricing agent is the result of years of experience, and it is not possible to predict the future house selling price using simple rules. Mimicking such behavior lies at the core of supervised learning, as it is of high interest to automate such decision-making processes using a set of previously gathered observations.

Before proceeding to the formalization of supervised learning, we note that in the first example, the goal is to identify to which class out of a finite set of possible classes a given image that may belong to, making it a *classification* task, whereas in the second one, the set of output values is infinite and continuous, corresponding to a *regression* task. In what follows, we describe a general framework unifying these two cases, then we restrict our

interest to classification, and more precisely binary classification, with which most of this manuscript is concerned. In the previous examples, the set of input-output observed pairs is called a dataset, and in the most simple setting the input examples are characterized by the same set of *descriptors* or *features*. As pointed out in Shalev-Shwartz and Ben-David (2014), it is more convenient to call it a *sequence* rather than a set to account for pairs that are observed more than once and we will use both terms interchangeably. Both inputs and outputs are in a vast majority of cases encoded as real-valued vectors. Learning the task associated to this data set consists in finding a hypothesis, or more precisely a function, that is able to predict the output of a given input as accurately as possible, even if the input at hand is not included in the provided data set, *i.e.* the hypothesis has not "observed" the example before. This last requirement is called *generalization*, as it assesses how the learned hypothesis generalizes to unseen data. In practice, since the observed data is finite, one can find an infinite number of hypotheses that perform well on it. The intuition of choosing the best one is in line with the principle of *Occam's razor* (Vapnik, 2006, Section 2.7.1) that suggests choosing the "simplest" one. This latter intuition, as we will see, is tightly linked to the generalization of a given hypothesis.

1.1 Theoretical Framework

In this section, we formalize the intuition presented previously by introducing a rigorous model of the process generating the observed data, the evaluation of a hypothesis that aims at explaining it, and how to learn it to generalize well to previously unseen observations.

1.1.1 Observed Data

The data pairs are modeled as a sample or sequence $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, where the inputs $\mathbf{x}_i$ and the outputs y_i are elements of sets $\mathbb{X}$ and $\mathbb{Y}$, respectively. In this setting, $\mathbb{X}$ is known as the input or feature set, and is generally given by a bounded subset of some metric space (Definition A.1.1). In the rest of the manuscript, we will consider the most common case $\mathbb{X} \subseteq \mathbb{R}^n$, where n is the number of features describing each example. As for the output values, $\mathbb{Y}$ will be a continuously infinite subset of $\mathbb{R}^K$ for regression where $K \in \mathbb{N}^*$ is the dimension of the output space, and $\mathbb{Y} = \{c_1, \cdots c_K\}$ for classification, where K is the number of candidate classes. Commonly, a class is encoded by either a number $c_k = k \in \mathbb{N}$, or a vector $c_k = \mathbf{e}_k$, where $\mathbf{e}_k$ is the k^{th} vector of $\mathbb{R}^K$'s canonical basis.

The data generating process is modeled by a joint probability distribution $\mathcal{D}$ of the couple $(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}$, *i.e.* $(\mathbf{x}, y) \sim \mathcal{D}$, and we write $S \sim \mathcal{D}^m$ for any sample S of m elements drawn independently from $\mathcal{D}$. In this case, the output value of y given $\mathbf{x}$ is not necessarily deterministic, and consequently represents possible measurement noise, or the fact that the considered features fail to completely determine the output. This latter scenario occurs due to the fact that when collecting the data one is often ignorant of the link between inputs and outputs (otherwise there's no need for learning). In practice, probability distribution $\mathcal{D}$ is observed only through the sample S, with an associated empirical discrete probability distribution $\hat{S}$ which is uniform over the different data points. Formally, it is defined as:

$$\hat{S} := \frac{1}{m} \sum_{i=1}^{m} \delta_{(\mathbf{x}_i, y_i)}, \tag{1.1}$$

where $\delta_{(\mathbf{x},y)}$ is the Dirac point measure associated with example $(\mathbf{x}, y)$. Taking the expectation with this probability distribution is exactly taking the empirical mean over sample S, *i.e.* for any measurable function f defined over $\mathbb{X} \times \mathbb{Y}$, one has:

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\hat{S}} [f(\mathbf{x}, y)] = \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{x}_i, y_i). \tag{1.2}$$

Additionally, we will consider the conditional distribution of y given an example $\mathbf{x}$ denoted by $\mathcal{D}_{\mathbb{Y}|\mathbf{x}}$, and the marginal distribution of inputs and outputs, respectively denoted by $\mathcal{D}_{\mathbb{X}}$ and $\mathcal{D}_{\mathbb{Y}}$.

1.1.2 Hypothesis Space

A *hypothesis* is a function $h : \mathbb{X} \to \mathbb{Y}$ that represents a possible deterministic rule of how the output values are generated from the input observations. A hypothesis is picked from a predetermined possible infinite set, called the *hypothesis space*, and denoted by $\mathbb{H}$. It encodes a certain structure of candidate hypotheses and reflects a prior knowledge, or similarly, a form of *inductive bias* (Shalev-Shwartz and Ben-David, 2014, Section 5.2), that one has about the problem at hand. Hypotheses are often taken to be functions with values in $\mathbb{R}^n$, whether the task is regression or classification with an *a posteriori* transformation applied to"discretize" their values in the latter case. We will detail this further in Section 1.2.3.

1.1.3 Task Performance

To assess the performance of a given hypothesis h, the classic approach is to use a function $l : \mathbb{Y}^2 \to \mathbb{R}_+$ quantifying the disagreement between the value of $h(\mathbf{x})$ and the observed output value y. In other words, this function models the loss incurred by h when predicting the value of y as $h(\mathbf{x})$, hence l is called a *loss function*. Aggregating the losses from individual examples to the whole data set is usually done by calculating its mean value over the data, or, more generally, its expectation over the generating probability distribution $\mathcal{D}$. The expectation of the loss l incurred by a hypothesis is known as the *$l-$risk* in the machine learning literature, as defined in the following definition.

Definition 1.1.1 ($l-$risk)**.** *Given a loss function l, the true*[1] *$l-$risk of a hypothesis h over a probability distribution $\mathcal{D}$ is*

$$\mathfrak{E}^l_{\mathcal{D}}(h) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [l(h(\mathbf{x}), y)] . \tag{1.3}$$

In the case of empirical distribution $\hat{S}$ associated to sample S, the $l-$risk of h is called the empirical $l-$risk over sample S, and is given by:

$$\mathfrak{E}^l_{\hat{S}}(h) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\hat{S}} [l(h(\mathbf{x}), y)] = \frac{1}{m}\sum_{i=1}^{m} l(h(\mathbf{x}_i), y_i). \tag{1.4}$$

The ability to measure the risk of a hypothesis is a basic requirement of learning strategies that aim at minimizing it, as we will detail in the next section.

1.1.4 Learning the Task

Ideally, solving a supervised learning problem over a generating distribution $\mathcal{D}$ consists in finding a hypothesis $h^* \in \mathbb{H}$ that disagrees the least with the observed outputs:

$$h^* \in \mathop{\arg\min}_{h\in\mathbb{H}} \mathfrak{E}^l_{\mathcal{D}}(h) . \tag{1.5}$$

As we only have access to the sample $S \sim \mathcal{D}^m$, several approaches to approximate $\min_{h\in\mathbb{H}} \mathfrak{E}^l_{\mathcal{D}}(h)$ empirically have been studied in the literature. We hereby present some common strategies employed to learn a hypothesis h_S that is expected to have a low $l-$risk over $\mathcal{D}$. Once it is learned, one can use it to predict, or infer, the output of a newly observed instance $\mathbf{x} \in \mathbb{X}$ during the so-called *inference step*.

[1] Calling it "true" risk is due to the fact that it is calculated as if one had full access to the unknown data generating probability distribution.

1.1.4.1 Empirical Risk Minimization (ERM)

According to the law of large numbers, the empirical risk converges to the true risk when the number of available instances tends to infinity and thus provides a good proxy for the latter. Consequently, it suggests a learning strategy that consists in searching for a hypothesis minimizing the empirical risk, *i.e.*

$$h_S \in \arg\min_{h \in \mathbb{H}} \mathfrak{E}^l_{\hat{S}}(h)\,. \tag{1.6}$$

The minimization of the empirical risk over the choice of h can be seen in some cases as a minimization of the negative log-likelihood of observing the data set S, where h represents the parameters of the probability distribution that we try to estimate. In this case, the introduced inductive bias corresponds to the parametric form of the considered probability distribution and we refer the interested reader to Bishop (2006) for more details concerning this viewpoint.

1.1.4.2 The Bias-Variance Trade-off

The empirical risk minimization approach is prone to *overfitting*: the resulting hypothesis h_S can perfectly fit the observed sample S while having poor performance on the underlying distribution $\mathcal{D}$. This phenomenon is characterized by a high $l-$risk of h_S on a sample $S' \neq S$ generated from $\mathcal{D}$ that was not used for learning, and can be explained by one of the two most common reasons. First, the sample S may be not representative enough of the unknown distribution $\mathcal{D}$: this may happen when S is not large enough or when the output values are noisy, and thus gathering more examples helps in tackling this issue. Second, even when S is large, overfitting can occur as a result of an excessive richness of $\mathbb{H}$ implying that a small variation in the data set due, for instance, to changing a few learning examples, may alter a lot the learned hypothesis. As a result, the performance of h_S varies significantly for different samples drawn from $\mathcal{D}$ suggesting that the performance on sample S cannot be used as a faithful indicator of the performance on the whole distribution. For this reason, overfitting is also referred to as a *high variance* or *high complexity* problem.

To avoid overfitting, one must gather enough data and somehow restrain the considered hypotheses from being too flexible. While the former point concerns the data collecting process itself, the latter can be achieved by imposing some restrictions on the flexibility of the considered hypothesis space. On the other hand, one must be careful when imposing such restrictions as they may prevent the learned hypothesis from capturing complex patterns in the data, resulting in poor performance on both the observed sample and the true generating distribution. This phenomenon is called *underfitting*, or *high bias*, as imposing restrictions on the considered hypotheses reflects the inductive bias that one has towards $\mathbb{H}$ before starting to learn.

Formally, the difference between the best observed value of the risk $\min_{h \in \mathbb{H}} \mathfrak{E}^l_{\hat{S}}(h)$ induced by S and $\mathbb{H}$, and the best achievable true risk $\min_{h \in \mathbb{Y}^{\mathbb{X}}} \mathfrak{E}^l_{\mathcal{D}}(h)$ by a hypothesis that is only required to be deterministic, can be decomposed as follows (Shalev-Shwartz and Ben-David, 2014, section 5.2):

$$\underbrace{\mathfrak{E}^l_{\hat{S}}(h_S)}_{\substack{\text{best empirical}\\ \text{biased performance}}} - \underbrace{\min_{h \in \mathbb{Y}^{\mathbb{X}}} \mathfrak{E}^l_{\mathcal{D}}(h)}_{\substack{\text{best true}\\ \text{unbiased performance}}} = \underbrace{\min_{h \in \mathbb{H}} \mathfrak{E}^l_{\hat{S}}(h) - \mathfrak{E}^l_{\mathcal{D}}(h^*)}_{\text{Estimation error}} + \underbrace{\mathfrak{E}^l_{\mathcal{D}}(h^*) - \min_{h \in \mathbb{Y}^{\mathbb{X}}} \mathfrak{E}^l_{\mathcal{D}}(h)}_{\text{Approximation error}}\,. \tag{1.7}$$

The first grouping bracket of the right-hand side (r.h.s) of Equation (1.7) reflects the deviation between the empirical and the true best risks, with an inductive bias given by the particular choice of hypothesis space $\mathbb{H}$, and will converge to 0 as m goes to infinity

under certain assumptions on $\mathbb{H}$ (Section 1.2.5). It is called the *estimation error* as the empirical risk is an estimation of the true one. The second term, called the *approximation error*, represents the capacity of the best hypothesis from $\mathbb{H}$ to approximate the function in $\mathbb{Y}^{\mathbb{X}}$ achieving the best possible performance over the whole distribution $\mathcal{D}$. While the estimation error can be made arbitrarily small by gathering more data, picking the right hypothesis space is a much more complex problem as one needs to make sure that it is rich enough to decrease the approximation error, but not at the expanse of increasing of the estimation error for a given sample S. The "richness" of a hypothesis space has been theoretically formalized via the notions of *Vapnik-Chervonenkis (VC) dimension* (Vapnik and Chervonenkis, 1971) and the *Rademacher complexity* (Koltchinskii and Panchenko, 2000), and plays a crucial role in controlling the convergence rate of the estimation error towards 0. For a visual illustration of the bias-variance trade-off, we refer the interested reader to Hastie et al. (2001, Figure 7.2).

1.1.4.3 Learning while Avoiding Overfitting

As we have seen in the previous section, one can tackle the overfitting problem by restricting the search space $\mathbb{H}$. This idea is the motivation behind two learning approaches: the *structural risk minimization* and the *regularized risk minimization*.

Structural Risk Minimization (SRM) Introduced in Vapnik (1992), the structural risk minimization consists in minimizing the risk, while penalizing the structure of the considered hypothesis space. Concretely, a (possibly infinite) set of nested hypothesis spaces

$$\{\mathbb{H}_i; i \in I; \mathbb{H}_i \subset \mathbb{H}_{i+1}\},$$

is fixed *a priori*, and a penalization $\text{pen}(\cdot)$ is applied to the i^{th} hypothesis space, expressed as a function of the VC dimension of $\mathbb{H}$ (Definition 1.2.4). This penalization, contrary to the minimum operator, is an increasing function for set inclusion, implying that $\text{pen}(\mathbb{H}_i) \leq \text{pen}(\mathbb{H}_{i+1})$. More formally, SRM consists in finding

$$h_S \in \underset{\substack{h \in \mathbb{H}_i \\ 1 \leq i \leq n}}{\arg\min}\, \mathfrak{E}^l_{\hat{S}}(h) + \text{pen}(\mathbb{H}_i). \tag{1.8}$$

Consequently, the minimum risk is no longer a decreasing function of the chosen hypothesis space (for set inclusion), and its choice relies on a trade-off between the complexity of $\mathbb{H}_i$ and the empirical risk value.

Regularized Risk Minimization (RRM) The idea is to penalize the searched hypothesis at the learning time, via a *regularizer* (regularization function) $R(\cdot)$ which penalizes excessively flexible hypotheses:

$$h_S \in \underset{h \in \mathbb{H}}{\arg\min}\, \mathfrak{E}^l_{\hat{S}}(h) + \lambda R(h), \tag{1.9}$$

where λ is a positive parameter controlling the trade-off between the empirical risk minimization and the regularization strength. Using the notion of Lagrange multipliers (Karush, 1939; Kuhn and Tucker, 1951), one can show that such procedure is equivalent to empirical risk minimization over a subset of $\mathbb{H}$, hence strengthening the bias in the hope of reducing the variance (Hastie et al., 2001, Section 7.3). RRM is arguably the most used learning strategy in supervised learning and can be interpreted as a *maximum a posteriori* (MAP) estimation (Bishop, 2006).

1.2 Supervised Classification

From this section on, and unless stated otherwise, we will address the task of supervised classification, for which the set of possible output values $\mathbb{Y} = (c_1, \ldots, c_K)$ is finite. We will use a terminology proper to classification: for an input $\mathbf{x}$, its class will be called its *label* and the hypotheses will be called *classifiers*. We now introduce several other notions used throughout the manuscript.

1.2.1 Decision Boundaries

Any classifier $h : \mathbb{X} \to \mathbb{Y}$ defines a partition of the input space $\mathbb{X}$ into regions $\mathbb{X}_k$ corresponding to the inverse images by classifier h of the different classes in $\mathbb{Y}$. Formally, we have:

$$\mathbb{X} = \bigcup_{k=1}^{K} \mathbb{X}_k, \quad \text{where} \quad \mathbb{X}_k := h^{-1}(\{c_k\}). \tag{1.10}$$

The boundaries between these regions are called the *decision boundaries*, as crossing them changes the decision made about the class that should be predicted for a given instance. The shape of these boundaries reflects the structure of the hypothesis space $\mathbb{H}$: for linear classifiers, they are hyperplanes, while for nonlinear ones they are more complex manifolds that can "bend" to better respect the instances' labels.

1.2.2 Assessing Classification Performance

Intuitively, the most straightforward way of evaluating a classifier h on a given data set is to count the fraction of times it misclassifies data points, *i.e.* by computing the following quantity:

$$\frac{1}{m} \sum_{i=1}^{m} [h(\mathbf{x}_i) \neq y_i]. \tag{1.11}$$

The latter is exactly the empirical risk $\mathfrak{E}_{\hat{S}}^{01}(h)$ associated with the loss function l_{01} defined as

$$l_{01}(y', y) := [y' \neq y], \tag{1.12}$$

and called the *misclassification loss* or the *0-1 loss*. Of course, we can also consider the true risk associated to l_{01}, which is equal to:

$$\mathfrak{E}_{\mathcal{D}}^{01}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [l_{01}(h(\mathbf{x}), y)] = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}} [h(\mathbf{x}) \neq y], \tag{1.13}$$

i.e. the probability of misclassification. For this particular loss, the best error achievable by a measurable function from $\mathbb{Y}^{\mathbb{X}}$ (last term in Equation (1.7)) is known as the *Bayes error*, and the classifier achieving it is called the *Bayes classifier* defined as follows:

$$h_{\text{Bayes}}(\mathbf{x}) = c_{k^*}, \quad \text{where} \quad k^* \in \arg\max \mathop{\mathbb{P}}_{y\sim\mathcal{D}_{\mathbb{Y}|\mathbf{x}}} [y = c_k]. \tag{1.14}$$

Despite the simplicity behind the definition of l_{01}, finding a hypothesis h that minimizes its associated risk is an NP-hard problem due to two main reasons. First, the hypothesis takes values in $\mathbb{Y}$ which is a finite set, making it discontinuous on $\mathbb{X}$. Second, the l_{01} loss is, in its turn, also discontinuous. It was shown that even for continuous hypotheses, the minimization of the l_{01} risk is an NP-hard problem (Arora et al., 1997).

To make such optimization procedure tractable, the 0-1 loss is often replaced by a surrogate that is convex in $h(\mathbf{x})$, or more generally in the parameters defining h, in order to benefit from efficient convex optimization techniques (Boyd and Vandenberghe, 2004; Bubeck, 2015). We will return to this point when discussing binary classification. Also, instead of searching for a hypothesis with a finite set of possible values, this assumption can be relaxed, leading to the notion of scoring functions that we present below.

1.2.3 Scoring Functions

The classic approach in classification is not to directly search for a classifier with values in $\mathbb{Y}$, but rather for a function with values in $\mathbb{R}^K$, called a *scoring function*. The continuous range of these functions' values (as in regression) allows to encode more information when compared to a traditional classifier. Indeed, if $\vec{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}))$, then $f_i(\mathbf{x})$ usually represents the confidence attributed by $\vec{f}$ for $\mathbf{x}$ to belong to class c_k. What these components exactly represent depends on the considered learning algorithm. For instance, they can be equal to an estimation of the conditional probabilities of the labels, *i.e.*

$$(f(\mathbf{x}))_k = \mathop{\mathbb{P}}_{y \sim \mathcal{D}^{\theta}_{\mathbb{Y}|\mathbf{x}}} [y = c_k]$$

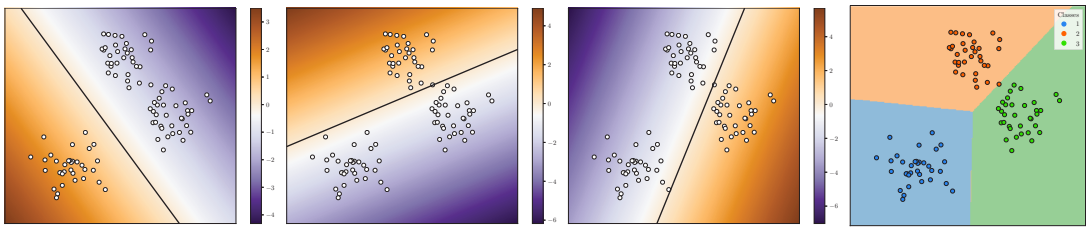
where θ is the parameters of the considered probability model. Or, they can be defined by the distance to the decision boundary separating it from the closest different class, also called the *classification margin* (Koltchinskii and Panchenko, 2002).

Formally, the classifier h_f associated with a scoring function f is given by the following rule:

$$h_f(\mathbf{x}) = c_{k^*} \quad \text{with} \quad k^* \in \mathop{\arg\max}_{1 \leq k \leq K} f_k(\mathbf{x}), \tag{1.15}$$

implying that it predicts the class for a given data point which the scoring function is most confident about, similar to the rule behind the Bayes classifier defined in Equation (1.14).

In addition to the rich information they enclose, scoring functions benefit from appealing properties for the minimization algorithms, as they are smoother than finite-valued classifiers.



(a) $f_1(\mathbf{x})$. (b) $f_2(\mathbf{x})$. (c) $f_3(\mathbf{x})$. (d) classifier $h_f(\mathbf{x})$.

Figure 1.1: From scores to classes: every component of scoring function $\vec{f} : \mathbb{X} \to \mathbb{R}^3$ represents the confidence accorded to its corresponding class. The color bars indicate the value of $f_k(\mathbf{x})$ for $k \in \{1, 2, 3\}$.

Remark In what follows and when clear from the context, we will also refer to scoring functions as classifiers, even if they have a continuous set of values. Consequently, we extend the domain of loss functions to $\mathbb{R}^K \times \mathbb{Y}$, allowing the first argument to be a scoring function.

1.2.4 Binary Classification

When $|\mathbb{Y}| = 2$, we fall into the setting of binary classification. Despite its apparent simplicity, binary classification is the task that corresponds to many challenging real-world applications of machine learning. It is also the main setting considered in the rest of this manuscript.

1.2.4.1 Encoding the Labels

Since the number of classes is $K = 2$, the scoring function $\vec{f}$ takes its values in $\mathbb{R}^2$, whereas h_f takes its values in $\{c_1, c_2\}$. According to the decision rule given in Equation (1.15), the

binary classifier is given by

$$h_f(\mathbf{x}) = \begin{cases} c_1, & \text{if} \quad f_1(\mathbf{x}) > f_2(\mathbf{x}) \\ c_2, & \text{otherwise.} \end{cases} \tag{1.16}$$

If we use the labelling $c_1 = 1, c_2 = 0$, then $h(\mathbf{x}) = [f_1(\mathbf{x}) > f_0(\mathbf{x})]$. In the particular case where f_1 and f_2 represent the conditional probabilities of two classes, the previous criterion reduces to $[f_1(\mathbf{x}) > \frac{1}{2}]$. It is then sufficient to keep only the component f_1 to make the prediction. Another interesting encoding is $c_1 = 1, c_2 = -1$ for which the terminology *negative* and *positive* classes is used. In this case, $h(\mathbf{x}) = \text{sgn}\,(f_1(\mathbf{x}) - f_2(\mathbf{x}))$ and one can define the scoring function $f = f_1 - f_2$ instead of keeping the two components. These different choices are summarized in Table 1.1.

Label encoding	**Scoring function's codomain**	**Decision rule**
$\{0, 1\}$	$\mathbb{R}^2$	$[f_1(\mathbf{x}) > f_2(\mathbf{x})]$
$\{0, 1\}$	$[0, 1]$	$[f_1(\mathbf{x}) > \frac{1}{2}]$
$\{-1, 1\}$	$\mathbb{R}$	$\text{sgn}\,(f_1(\mathbf{x}) - f_2(\mathbf{x}))$

Table 1.1: Different encodings and decision rules for binary classification.

In what follows, we adopt the encoding $\mathbb{Y} = \{-1, 1\}$. As a result, we consider scoring functions having values in $\mathbb{R}$ instead of $\mathbb{R}^2$. The value of the scoring function then represents the difference $f_1 - f_2$ and we denote it by f. Moreover, whether the prediction agrees with the observed label boils down to whether $y \cdot f(\mathbf{x}) > 0$, *i.e.*

$$\text{sgn}\,(f(\mathbf{x})) = y \quad \Leftrightarrow \quad y \cdot f(\mathbf{x}) > 0. \tag{1.17}$$

1.2.4.2 Classification Margin

The scoring function $f = f_1 - f_2$ is the difference of confidences in the available classes, and while comparing the two components (*i.e.* , comparing the value of f to 0) allows to decide the class of an instance, one can also benefit from the information contained in the continuous range of values of f_1 and f_2 to assess the confidence in the prediction. This can be done by comparing this difference to a positive constant $\rho > 0$ representing a margin parameter: stating that f_1 is greater than f_2 with a margin ρ sheds more light on the confidence that one attributes to the made prediction. This intuition can be captured in two different ways by introducing the notions of signed and absolute margins.

Signed margin We saw in Equation (1.17) that it is sufficient to compare $y \cdot f(\mathbf{x})$ to 0 to see if f prediction for $\mathbf{x}$ agrees with label y. The quantity $y \cdot f(\mathbf{x})$ is called the *signed margin* of f at $\mathbf{x}$: if $y = 1$, then it measures how far f_1 is from f_2, and inversely for $y = -1$.

Absolute margin This notion characterizes the confidence of the scoring function f in its own prediction, regardless of whether it is correct or not, and it is defined by $|f(\mathbf{x})|$, *i.e.* the absolute value of the signed margin.

For a generalization of margin theory to the multi-class case, *i.e.* for $K > 2$, we refer the interested reader to Koltchinskii and Panchenko (2002).

1.2.4.3 Loss Functions

For $\mathbb{Y} = \{-1, 1\}$, the commonly used loss functions have the following form

$$l(h(\mathbf{x}), y) = \ell(y \cdot h(\mathbf{x})), \tag{1.18}$$

where ℓ is a non increasing function verifying $\ell(t) \xrightarrow{t\to\infty} 0$. This choice reflects the idea that the larger is the signed margin of an example $(\mathbf{x}, y)$, the less h is penalized. In practice, when one seeks to find the best classifier for a given dataset via a risk minimization strategy, a convex ℓ is chosen. In Table 1.2, we list some popular loss functions having the above-mentioned form and we illustrate them in Figure 1.2.

Loss name	**Notation**	$\ell(t)$	**Convex?**
Misclassification/0-1	l_{01}	$[t < 0]$	No
Margin violation	–	$[t < \rho]$ (where $\rho > 0$)	No
Ramp	–	$\min\left(1, \left(1 - \frac{t}{\beta}\right)_+\right)$ (where $\beta > 0$)	No
Hinge	l_+	$(1-t)_+$	Yes
Softplus	l_{soft}	$\log\left(1 + e^{-t}\right)$	Yes
Exponential	–	e^{-t}	Yes

Table 1.2: Common loss functions for binary classification.

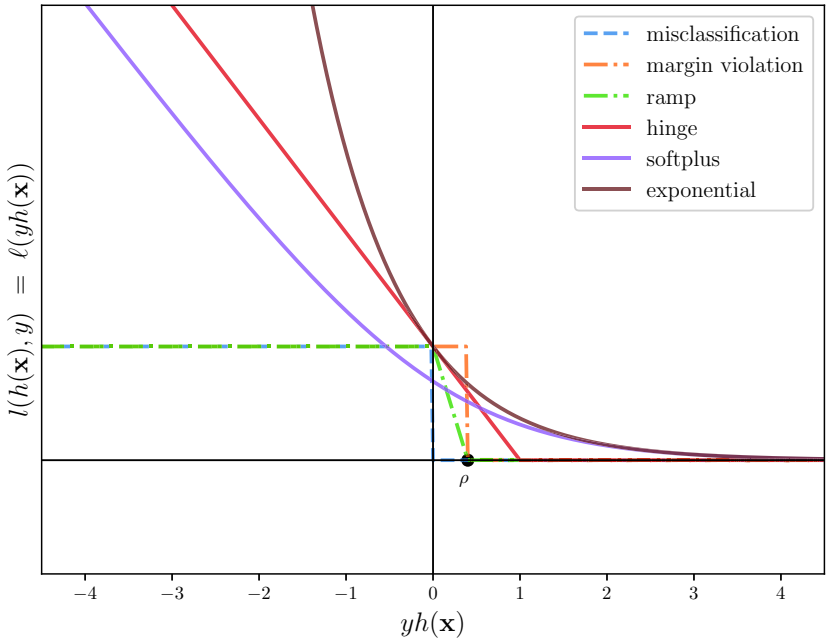


Figure 1.2: Common loss functions used in the binary classification setting where $\mathbb{Y} = \{-1, 1\}$.

1.2.5 Generalization Guarantees

In this section, we formalize several notions mentioned in Section 1.1.4.2 by introducing the rigorous definitions of the richness and complexity measures of hypothesis spaces and by reviewing the existing convergence rates of the estimation error towards zero.

1.2.5.1 Probably Approximately Correct (PAC) Learning

Probably Approximately Correct (PAC) learning, introduced in Valiant (1984), is a theoretical framework defining what it means for a hypothesis space $\mathbb{H}$ to be learnable. It formalizes the idea that given a large enough learning sample $S \sim \mathcal{D}^m$, the true risk associated to the hypothesis h_S learned from S is arbitrarily close to the minimum achievable

true risk (*i.e.* the second term of the estimation error in Section 1.1.4.2). Since even large samples can be non-representative of the distribution $\mathcal{D}$ (for example, imagine a very large sample with elements having all the same label), PAC learnability takes the probability of this latter event into account.

Definition 1.2.1 (PAC learnability)**.** *A hypothesis space $\mathbb{H}$ is PAC-learnable if for any $\epsilon, \delta \in (0,1)$, there exists $m(\epsilon,\delta) \in \mathbb{N}$ such that if $m \geq m(\epsilon,\delta)$, then for any probability distribution $\mathcal{D}$ over $\mathbb{X} \times \mathbb{Y}$, with a probability at least $1-\delta$ over the draw of a sample $S \sim \mathcal{D}^m$, we have:*

$$\mathfrak{E}^l_{\mathcal{D}}(h_S) \leq \min_{h \in \mathbb{H}} \mathfrak{E}^l_{\mathcal{D}}(h) + \epsilon. \tag{1.19}$$

In other words, by learning a hypothesis h_S from a sample S we will *probably* (with a confidence level $1-\delta$) succeed in obtaining an *approximately* (up to $\epsilon > 0$) *correct* hypothesis (minimizing the true risk over $\mathbb{H}$). The minimum $m(\epsilon,\delta)$ verifying the former requirement is called the *sample complexity*.

To prove PAC-learnability, a stronger requirement, called the *uniform convergence property*, is commonly used. It requires that uniformly over all of the hypothesis space $\mathbb{H}$, the empirical and true risks of a given hypothesis are close to each other with a high probability over the draw of sample S, as formalized by the following definition.

Definition 1.2.2 (Uniform convergence property)**.** *A hypothesis space $\mathbb{H}$ has the uniform convergence property if for any $\epsilon, \delta \in (0,1)$, there exists $m(\epsilon,\delta) \in \mathbb{N}$ such that if $m \geq m(\epsilon,\delta)$, then for any probability distribution $\mathcal{D}$ over $\mathbb{X} \times \mathbb{Y}$, with a probability at least $1-\delta$ over the draw of $S \sim \mathcal{D}^m$,*

$$\sup_{h \in \mathbb{H}} \left| \mathfrak{E}^l_{\mathcal{D}}(h) - \mathfrak{E}^l_{\hat{S}}(h) \right| \leq \epsilon. \tag{1.20}$$

The left hand side of Equation (1.20) reflects how much using sample S succeeds in approximating the true risk of a hypothesis $h \in \mathbb{H}$ uniformly over the choice of this latter. It is then a measure of the ability of S to represent $\mathcal{D}$ w.r.t. learning a hypothesis from $\mathbb{H}$, the reason for which it is called *representativeness* of sample S (Shalev-Shwartz and Ben-David, 2014, Section 26.1).

1.2.5.2 Uniform Generalization Bounds

Until now, the two parameters ϵ and δ were chosen arbitrarily, and it is the sample complexity $m(\epsilon,\delta)$ that depended on them. Another option is to let m and δ free and to express the approximability parameter ϵ as a function of the two. Additionally, some results in the literature are not uniform over the choice of the underlying distribution $\mathcal{D}$, thus ϵ may depend on it as well. With this re-parametrization, we give in the next definition a general form of uniform generalization inequalities, which are a consequence of the uniform convergence property (Definition 1.2.2).

Definition 1.2.3 (Uniform generalization bound)**.** *Given a hypothesis space $\mathbb{H}$ and a probability distribution $\mathcal{D}$, a uniform generalization bound has the following form:*

For any $\delta \in (0,1)$, with a probability at least $1-\delta$ over the draw of sample $S \sim \mathcal{D}^m$, the following holds

$$\forall h \in \mathbb{H}, \quad \mathfrak{E}^l_{\mathcal{D}}(h) \leq \mathfrak{E}^l_{\hat{S}}(h) + \epsilon(\mathcal{D}, \mathbb{H}, \delta, m), \tag{1.21}$$

where $\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) \xrightarrow{m \to \infty} 0$.

In Equation (1.21), the gap between the true and the empirical risks of any hypothesis $h \in \mathbb{H}$ is controlled by $\epsilon(\mathcal{D}, \mathbb{H}, \delta, m)$ that vanishes when one gathers more data. The dependence on $\mathbb{H}$ is expressed via a complexity measure of this latter: the more complex are the hypotheses from $\mathbb{H}$, the larger is this term for a fixed m and the slower is the

convergence towards the true risk. As for the dependence on $\mathcal{D}$, it is introduced to cover a general form of generalization bounds.

In what follows, we consider two particular cases of uniform generalization bounds depending on the considered measure of complexity of $\mathbb{H}$ given by the *Vapnik-Chervonenkis dimension* (Vapnik, 1992) and the *Rademacher complexity* (Koltchinskii and Panchenko, 2000).

Definition 1.2.4 (VC dimension)**.** *The Vapnik-Chervonenkis (VC) dimension of a binary hypothesis space $\mathbb{H}$ is the size of the largest sample of elements from $\mathbb{X}$ that can be labeled in all of the possible ways by hypotheses from $\mathbb{H}$:*

$$VC(\mathbb{H}) := \max\{|A|;\ |A| < \infty;\ A \subset \mathbb{X};\ |\mathbb{H}(A)| = 2^{|A|}\}, \tag{1.22}$$

where

$$\mathbb{H}(A) := \{h(\mathbf{x});\ \mathbf{x} \in A;\ h \in \mathbb{H}\} \tag{1.23}$$

is the set of all possible labelings of A by elements from $\mathbb{H}$.

The VC dimension is a measure of the richness of the hypothesis space $\mathbb{H}$ and captures from which sample size a hypothesis space $\mathbb{H}$ stops behaving like functions from $\mathbb{Y}^{\mathbb{X}}$, as these latter can label any finite sample $A \subset \mathbb{X}$ in all of the possible $2^{|A|}$ ways. It is independent of the probability distribution $\mathcal{D}$ generating the data, which is not the case for the next complexity measure.

Definition 1.2.5 (Rademacher complexity)**.** *Let $r_1, ..., r_m$ be Rademacher random variables,* i.e.

$$\mathbb{P}[r_i = 1] = \mathbb{P}[r_i = -1] = \frac{1}{2}, \quad \forall 1 \leq i \leq m.$$

1. *The empirical Rademacher complexity of a hypothesis space $\mathbb{H}$ associated to a finite sample $S \subset \mathbb{X}$ is*

$$\mathrm{Rad}_S(\mathbb{H}) := \mathop{\mathbb{E}}_{r_1,...,r_m \sim r}\left[\sup_{h \in \mathbb{H}} \frac{1}{m}\sum_{i=1}^{m} r_i h(\mathbf{x}_i)\right]. \tag{1.24}$$

2. *The Rademacher complexity of a hypothesis space $\mathbb{H}$ associated to a sample size m is*

$$\mathrm{Rad}_m(\mathbb{H}) := \mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\mathbb{X}}^m}\left[\mathrm{Rad}_S(\mathbb{H})\right]. \tag{1.25}$$

For a sample S, the empirical Rademacher complexity measures the ability of hypotheses from $\mathbb{H}$ to correlate with random noise defined by the Rademacher random variables. If the correlation is high, then the hypotheses are too flexible and may lead to overfitting.

Both introduced complexity measures help in quantifying the deviation of the true $l-$risk from the empirical one, as stated by the following theorem.

Theorem 1.2.1. *Given a binary hypothesis space $\mathbb{H}$ and the misclassification loss l_{01}, generalization bound of Equation* (1.21) *holds with $\epsilon(\mathcal{D}, \mathbb{H}, \delta, m)$ defined:*

- *with the VC dimension*

$$\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) = 2\sqrt{\frac{1}{m}\left(VC(\mathbb{H})\log\left(\frac{2em}{VC(\mathbb{H})}\right) + \log\frac{4}{\delta}\right)}. \tag{1.26}$$

- *with the empirical Rademacher complexity*

$$\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) = \mathrm{Rad}_S(\mathbb{H}) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{1.27}$$

- *with the Rademacher complexity*

$$\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) = \mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\mathbb{X}}^m} \left[\mathrm{Rad}_S\left(\mathbb{H}\right)\right] + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{1.28}$$

We note that in the first case, the generalization bound is independent of the probability distribution $\mathcal{D}$, hence it holds uniformly over all the considered distributions in addition to the hypotheses from $\mathbb{H}$.

While the previous result concerns the 0-1 loss, different generalization bounds can be derived for other loss functions, especially for those verifying the Lipschitz property (Definition A.1.2). Moreover, the Rademacher complexity and VC dimension are related, hence one can establish Rademacher bounds and then deduce VC bounds. We refer the interested reader to Mohri et al. (2018, Chapter 3) for more details on these aspects.

1.2.5.3 Other Generalization Bounds

The uniform generalization bounds from the previous section do not take into account the learning algorithm, as they hold uniformly over all of the hypotheses considered. By algorithm, we mean any rule that takes a sample $S \sim \mathcal{D}^m$ and a hypothesis space $\mathbb{H}$ and outputs a hypothesis h_S, as in the case of ERM, RRM and SRM (Section 1.1.4).

Recent lines of work provided generalization guarantees that take into account the algorithm used to produce a hypothesis in order to study its generalization. Among these contributions, we cite:

Stability theory Introduced by Bousquet and Elisseeff (2002), this framework formalizes the idea that if the loss associated to the output hypothesis of an algorithm does not change too much under the removal of one element of the training sample S, then the learned hypothesis has good generalization properties. More precisely, they define *uniform stability* of a learning algorithm as follows.

Definition 1.2.6 (Uniform stability)**.** *A learning algorithm $\mathfrak{A}$ has uniform stability $\beta > 0$ w.r.t. a loss function l if for any finite sequence $S \subset \mathbb{X} \times \mathbb{Y}$, and for any $i \in \{1, \ldots, |S|\}$, we have*

$$\sup_{(\mathbf{x},y) \in S} |l(h_S(\mathbf{x}), y) - l(h_{S^{\setminus i}}, y)| \leq \beta, \tag{1.29}$$

where $S^{\setminus i} := S \setminus \{(\mathbf{x}_i, y_i)\}$ and h_S and $h_{S^{\setminus i}}$ are learned by $\mathfrak{A}$ from S and $S^{\setminus i}$ respectively.

If an algorithm verifies the previous definition, then its output hypotheses enjoying the following generalization guarantee.

Theorem 1.2.2. *Let $\mathfrak{A}$ be an algorithm having uniform stability β and assume there exists $M > 0$ such that for any $S \sim \mathcal{D}^m$ and any $(\mathbf{x}, y) \in S$, $l(h(\mathbf{x}), y) \leq M$. Then for any $\delta \in (0, 1)$, with a probability $1 - \delta$ over the draw of $S \sim \mathcal{D}^m$, we have:*

$$\mathfrak{E}_{\mathcal{D}}^l\left(h_S\right) \leq \mathfrak{E}_S^l\left(h_S\right) + 2\beta + (4m\beta + M)\sqrt{\frac{\log \frac{1}{\delta}}{2m}}, \tag{1.30}$$

where h_S is learned by $\mathfrak{A}$ from S.

The authors show that this requirement is met by RRM (Section 1.1.4.3) algorithms for loss functions verifying the Lipschitz property w.r.t. the first argument and when the regularizer is a squared norm in a *Reproducing Kernel Hilbert Space* (RKHS, Section 1.2.6.1). They also prove that for several classic algorithms, the stability factor β is inversely proportional to the number of samples, hence the generalization bound enjoys an overall decay rate of $\frac{1}{\sqrt{m}}$.

Robustness theory Introduced in Xu and Mannor (2010, 2012), the main requirement for an algorithm to enjoy generalization guarantees is that its output hypothesis should have similar performance on instances that fall in the same set from an partition of the $\mathbb{X}$ fixed beforehand, as formalized by the following definition.

Definition 1.2.7 (Algoritmic robustness)**.** *Given $M \in \mathbb{N}$ and $\epsilon : (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}_+$, an algorithm $\mathfrak{A}$ is $(M, \epsilon(\cdot))-$robust on $\mathcal{D}$ w.r.t. a loss function l if it is possible to partition $\mathbb{X} \times \mathbb{Y}$ into M subsets $\{\mathcal{Z}_k\}_{k=1}^M$ such that for all $S \sim \mathcal{D}^m$ and all $(\mathbf{x}, y) \in S, (\mathbf{x}', y') \sim \mathcal{D}$ and $1 \leq k \leq M$, we have:*

$$(\mathbf{x}, y), (\mathbf{x}', y') \in \mathcal{Z}_k \Rightarrow \left| l(h_S(\mathbf{x}), y) - l(h_S(\mathbf{x}'), y') \right| \leq \epsilon(S), \tag{1.31}$$

where h_S is the hypothesis learned by $\mathfrak{A}$ from S.

The robustness-based generalization guarantee is then stated as follows.

Theorem 1.2.3. *Let $\mathfrak{A}$ be an $(M, \epsilon(\cdot))-$ robust algorithm on $\mathcal{D}$ w.r.t. a loss function l. Assume that for some constant $B > 0$, $l(h_S(\mathbf{x}), y) \leq B$ for all $(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}$. Then, for any $\delta \in (0, 1)$, with a probability $1 - \delta$ over the draw of a sample $S \sim \mathcal{D}^m$, we have*

$$\mathfrak{E}^l_{\mathcal{D}}(h_S) \leq \mathfrak{E}^l_S(h_S) + \epsilon(S) + B\sqrt{\frac{2\left(\log 2M + \log \frac{2}{\delta}\right)}{m}}, \tag{1.32}$$

where h_S is learned by $\mathfrak{A}$ from S.

According to Xu and Mannor (2012), the SVM algorithm (Section 1.2.6.1), majority voting-based algorithms and feed-forward neural networks (under Lipschitzness assumptions) benefit from algorithmic robustness.

PAC Bayesian theory This framework addresses the learning problem from a fundamentally different point of view, by expressing a prior belief about the best hypothesis for the learning problem at hand, formalized by a prior probability distribution $\mathcal{P}_0$ over h. The goal then is to learn a posterior probability distribution $\mathcal{P}$ over $\mathbb{H}$. Below, we present a generalization bound for the l_{01} loss due to Catoni (2007), involving the *Kullback-Leibler* divergence (Definition A.2.8).

Theorem 1.2.4. *Let $\mathcal{P}_0$ be a prior distribution over $\mathbb{H}$, let $\delta \in (0, 1)$ and $c > 0$. Then, with a probability $1 - \delta$ over the draw of $S \sim \mathcal{D}^m$, we have for any probability distribution $\mathcal{P}$ over $\mathbb{H}$:*

$$\mathop{\mathbb{E}}_{h \sim \mathcal{P}} \left[\mathfrak{E}^{01}_{\mathcal{D}}(h) \right] \leq \frac{c}{1 - e^{-c}} \left(\mathop{\mathbb{E}}_{h \sim \mathcal{P}} \left[\mathfrak{E}^{01}_{S}(h) \right] + \frac{\mathrm{KL}(\mathcal{P} \| \mathcal{P}_0) + \log \frac{1}{\delta}}{cm} \right). \tag{1.33}$$

In addition to the fact that this bound concerns the $\mathcal{P}-$expectation of the $l-$risk over $\mathbb{H}$, the parameter c reflects a trade-off between the $\mathcal{P}-$expected empirical risk $\mathop{\mathbb{E}}_{h \sim \mathcal{P}} \left[\mathfrak{E}^{01}_{S}(h) \right]$ and a complexity term $\frac{\mathrm{KL}(\mathcal{P} \| \mathcal{P}_0)}{m}$. This trade-off has tight connections with the MAP estimation interpretation of the RRM learning rule (in fact, connections to the SVM algorithm were shown in Germain et al. (2009)). Moreover, choosing $c = \frac{1}{\sqrt{m}}$ allow the r.h.s. to converge towards the empirical risk.

1.2.6 Some Notable Algorithms

In this section, we present two popular supervised classification algorithms: the *support vector machine* and the *k-nearest neighbors* algorithms.

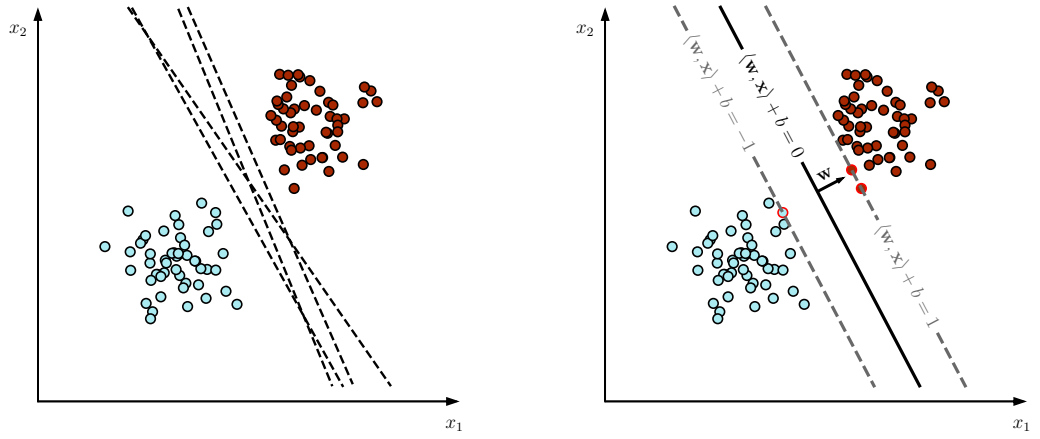
1.2.6.1 Support Vector Machine (SVM)

Introduced in Boser et al. (1992); Cortes and Vapnik (1995), the SVM is a binary classification algorithm (although it can be extended the multiclass setting) based on the idea of finding a hyperplane that separates the instances of the two classes, while staying as far as possible from them. Formally, the SVM aims to find an affine hypothesis from the set

$$\mathbb{H} = \{h : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b; \mathbf{w} \in \mathbb{R}^n, \|\mathbf{w}\|_2 = 1, b \in \mathbb{R}\}, \tag{1.34}$$

that maximizes the minimal signed margin on the observed dataset S. The idea is portrayed in Figure 1.3, and consists in solving the following optimization problem:

$$\max_{h \in \mathbb{H}} \left(\min_{1 \leq i \leq m} y_i \cdot h(\mathbf{x}_i) \right). \tag{1.35}$$



(a) Several candidate classifiers successfully separate the two classes. The dashed lines indicate the decision boundaries.

(b) The candidate maximizing the margin is chosen. The support vectors have red edges, and correspond to saturated constraints.

Figure 1.3: Illustration of the idea behind the SVM algorithm, where the brown and blue colors represent the two classes.

The previous problem is equivalent to the following quadratic programmming problem (QP):

$$\begin{aligned} \min_{\substack{\mathbf{w} \in \mathbb{R}^n \\ b \in \mathbb{R}}} \quad & \|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \qquad \forall\ 1 \leq i \leq m. \end{aligned} \tag{1.36}$$

The feasible set of the previous problem is non empty if and only if the two classes are linearly separable[2], which is not always the case in practice. To tackle this limitation, the constraints are relaxed (Cortes and Vapnik, 1995) via the introduction of non-negative slack variables ξ_i:

$$\begin{aligned} \min_{\substack{\mathbf{w} \in \mathbb{R}^n \\ b \in \mathbb{R}}} \quad & C \sum_{i=1}^{m} \xi_i + \|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & \xi_i \geq 0, \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \qquad \forall 1 \leq i \leq m, \end{aligned} \tag{1.37}$$

[2]In fact, the constraints define a separating hyperplane between elements of the two classes.

where C controls a trade-off between minimizing the violation allowed by slack variables and the norm of the classifier. The previous formulation is further equivalent to the following problem:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \quad C\sum_{i=1}^{m} l_+(\langle \mathbf{w}, \mathbf{x}_i\rangle + b, y_i) + \|\mathbf{w}\|^2. \tag{1.38}$$

The latter formulation is an instance of RRM (Section 1.1.4.3), and the SVM algorithm then boils down to solving a convex problem which is nonsmooth due to the non differentiability of the hinge loss.

Dual formulation Arguably, the most popular method for finding the SVM classifier is switching to its dual formulation that can be shown to have the following form:

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in[0,C]^m} \quad & \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}^T\mathbf{y} = 0, \end{aligned} \tag{1.39}$$

where $\mathbf{K} \in \mathbb{R}^{m\times m}$ with $\mathbf{K}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j\rangle$ and $\mathbf{y} = (y_1, \ldots, y_m)^T$.

The dual formulation of the SVM problem is a QP problem, and can be solved using standard convex optimization software. In practice, the *Sequential Minimal Optimization* (SMO) algorithm (Platt, 1998) is widely used by SVM libraries.

The optimal vector $\mathbf{w}^*$ has the following closed form:

$$\mathbf{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \mathbf{x}_i, \tag{1.40}$$

where $\boldsymbol{\alpha}^*$, the solution of the dual problem, is often sparse in practice, meaning that most of its components are zero. This implies that the classifier $\mathbf{w}^*$ is a combination of only a subset of the data points lying exactly on the margin boundary and commonly called *support vectors*. These points can be used to compute the bias term b as well by noticing that, since $y^2 = 1$, a saturated constraint can be equivalently written $\mathbf{w}^T\mathbf{x}_i + b = y$, and theoretically any point at which the constraints are saturated is sufficient to determine b. In practice, and for numerical stability, b is computed as an average over the set of p support vectors (Bishop, 2006, Section 7.1.1):

$$b^* = \frac{1}{p}\sum_{i:\alpha_i>0} (y_i - \langle \mathbf{w}^*, \mathbf{x}_i\rangle). \tag{1.41}$$

Kernel trick As mentioned above, the first formulation of SVM, called the hard-margin SVM, admits a solution only when data is linearly separable. This limitation can be dealt with by introducing the slack variables leading to a soft-margin SVM formulation. However, even solving the relaxed problem can still result in poor performance when using affine classifiers. To address this drawback, a common approach in machine learning is to use a new data representation that can be given, for instance, by a mapping of the data to a new feature space where the classes may hopefully become linearly separable. This idea turns out to be particularly suitable in the context of the SVM algorithm since its dual formulation only depends on the inner products between data instances used to define the matrix $\mathbf{K}$. This observation is what motivated the so-called *kernel trick* (Boser et al., 1992) allowing the SVM dual to be defined for instances in high-dimensional and even infinite-dimensional spaces, as long as it is possible to calculate the inner product between two instances in such space. In other words, apart from the instances' labels, the inner products between instances is the only information needed to solve the SVM problem.

The idea of such inner products is well-suited for the case of *Reproducing Kernel Hilbert Spaces* (RKHS) (Aronszajn, 1950), which are Hilbert spaces of real-valued functions over

$\mathbb{X}$, where for any $\mathbf{x} \in \mathbb{X}$ the evaluation mapping $\eta_{\mathbf{x}} : f \mapsto f(\mathbf{x})$ is continuous. By invoking the Riesz representation theorem (Riesz, 1914), it can be represented by an inner product . In other words, for an RKHS $\mathbb{V}$, we have

$$\forall \mathbf{x} \in \mathbb{X}, \exists k_{\mathbf{x}} \in \mathbb{V} \quad \text{such that} \quad \forall f \in \mathbb{V}, \eta_{\mathbf{x}}(f) = f(\mathbf{x}) = \langle k_{\mathbf{x}}, f \rangle_{\mathbb{V}} . \tag{1.42}$$

The idea behind the kernel trick is to map the data points from $\mathbb{X}$ to a certain RKHS $\mathbb{V}$ via the mapping $\mathbf{x} \mapsto k_{\mathbf{x}}$ and then to consider an inner product in $\mathbb{V}$, known as a *kernel* function formally defined as follows.

Definition 1.2.8 (Kernel function)**.** *A function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is called a kernel, if there exists a reproducing kernel Hilbert space (RKHS) $(\mathbb{V}, \langle ., . \rangle_{\mathbb{V}})$ and a mapping $\phi : \mathbb{X} \to \mathbb{V}$ such that:*

$$K(\mathbf{x}, \mathbf{x}') = \left\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \right\rangle_{\mathbb{V}} . \tag{1.43}$$

According to Definition 1.2.8, K is characterized by an inner product after mapping the data points from $\mathbb{X}$ to $\mathbb{V}$ via ϕ. However, such a definition requires knowing both $\mathbb{V}$ and ϕ, where the latter may not be possible to represent numerically when $\mathbb{V}$ is an infinite-dimensional space. This problem is solved by *Mercer's theorem*, which characterizes functions that are kernels without needing to know $\mathbb{V}$ or ϕ.

Theorem 1.2.5. *(Mercer, 1909) A function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a kernel, if and only for any vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{X}$, the matrix $\mathbf{K}$ defined by $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ is symmetric and positive semi-definite (PSD).*

We provide examples of some popular kernels in Table 1.3.

Name	$K(\mathbf{x}, \mathbf{x}')$	**Parameters**
Linear	$\langle \mathbf{x}, \mathbf{x}' \rangle$	
Polynomial	$(1 + \langle \mathbf{x}, \mathbf{x}' \rangle)^b$	degree b
Radial basis function (RBF)/Gaussian	$e^{-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2}$	$\gamma > 0$
Laplace	$e^{-\gamma \|\mathbf{x} - \mathbf{x}'\|_1}$	$\gamma > 0$

Table 1.3: Examples of some commonly used kernel functions.

Although replacing $\mathbf{x}_i$ by $\phi(\mathbf{x}_i)$ in Equation (1.40) does not always allow to express $\mathbf{w}^*$ explicitely, it still can be done for a given instance at the inference step (Section 1.1.4):

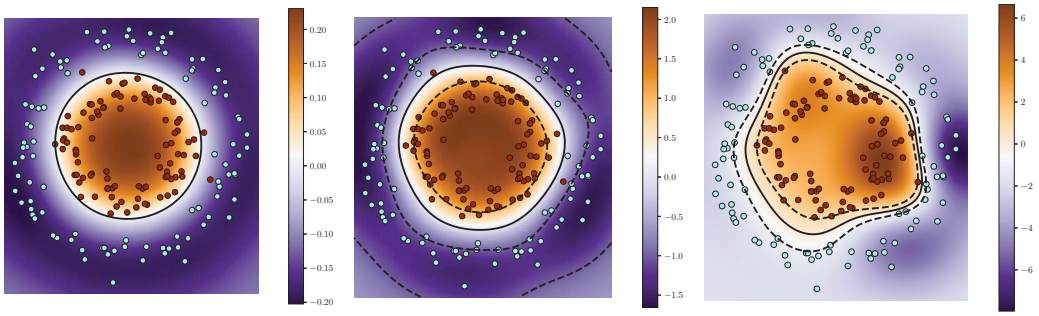
$$\langle \mathbf{w}^*, \phi(\mathbf{x}) \rangle_{\mathbb{V}} = \sum_{i=1}^{m} \alpha_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle_{\mathbb{V}} = \sum_{i=1}^{m} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}). \tag{1.44}$$

Bias-variance trade-off for SVMs According to the value of the parameter C in the SVM's RRM formulation (Equation (1.38)), one can continuously move between the underfitting and overfitting regimes as illustrated in Figure 1.4:

- For large values of C, one focuses on minimizing the slack variables ξ (represented in the RRM formulation via the hinge loss), at the expense of allowing the norm of the vector $\mathbf{w}$ to be very large and, thus decreasing the separation margin. In this case, the learned classifier might incur considerable changes by adding new data points that violate the margin leading to a high variance problem.
- For small values of C, it is the minimization of the classifier's norm, and consequently the maximization of the separation margin, that is privileged. The larger is the demanded separation margin, the more restricted is the choice of candidate classifiers.

Aside from C, when using a non-linear kernel, other parameters can be directly related to overfitting. For example, using a polynomial kernel with a high degree b leads to an excessive flexibility and can induce a high variance problem.

From a theoretical point of view, the SVM algorithm has appealing generalization properties, as the VC dimension of the class of affine classifiers in $\mathbb{R}^n$ is finite and is equal to $p + 1$ (Shalev-Shwartz and Ben-David, 2014, Theorem 9.2). In the case of the ℓ_2 norm, corresponding to the most common formulation of the SVM, one can establish a generalization bound that is independent of the input space's dimension n by bounding the Rademacher complexity in Equation (1.28). For more details on generalization bounds for the formulation of the SVM with the ℓ_2 norm, we refer the interested reader to Shalev-Shwartz and Ben-David (2014, Section 26.3) and Bartlett and Shawe-Taylor (1999); Bartlett and Mendelson (2002), and for other norms to Shalev-Shwartz and Ben-David (2014, Section 26.4) and Kakade et al. (2009).



$C = 10^{-2}$. $C = 1$. $C = 10^{2}$.

Figure 1.4: Illustration of the impact of hyperparameter C for a non-linear SVM with RBF kernel having $\gamma = 4$. As C increases, the decision boundary becomes more flexible.

Other variations Using the RRM formulation of the SVM (Equation (1.38)) as a basis, several other variations of it have been proposed in the literature. For instance, the hinge loss can be replaced by a squared hinge loss $(1 - .)_+^2$ (Lee and Lin, 2013) to make the problem smoother and more suitable for gradient-based optimization procedures, especially in the case of low-dimensional and abundant data. Other modifications include adding a more sophisticated regularization term (Zhu et al., 2004; Wang et al., 2006). Furthermore, the SVM algorithm was extended to settings different from supervised learning, such as transductive SVM for semi-supervised learning (Joachims, 1999) and one-class SVM for anomaly detection (Schölkopf et al., 2000).

1.2.6.2 k-Nearest Neighbors (k-NN)

The k-NN algorithm (Cover and Hart, 1967) is arguably one of the most intuitive supervised learning algorithms. It formalizes the idea that when instances are spatially close to each other in the sense of some metric $d : \mathbb{X}^2 \to \mathbb{R}_+$ (Definition A.1.1), then they should belong to the same class. More formally, given a test instance $\mathbf{x} \in \mathbb{X}$, let $N_k(\mathbf{x})$ be the set of the k closest instances to $\mathbf{x}$ from the training sample in the sense of metric d. We define a scoring function $\vec{f}$ with components:

$$f_k(\mathbf{x}) = \frac{1}{k} \sum_{x' \in N_k(\mathbf{x})} [y(\mathbf{x}') = c_k]. \tag{1.45}$$

Then, the class of $\mathbf{x}$ is given by:

$$h(\mathbf{x}) := c_{k^*} \quad \text{where} \quad k^* = \underset{1 \leq k \leq K}{\arg\max}\, f_k(\mathbf{x}). \tag{1.46}$$

Hence, given the scoring function $\vec{f}$, deciding the class of an element consists in finding the class with the most of instances within the k neighbors. We immediately see that the nearest neighbors classifier is not the result of a minimization procedure, and is rather "learned" by the mere storage of the training data followed by a majority voting. Note that from a theoretical point of view, the nearest neighbors scoring function (Equation (1.45)) can be seen as an empirical estimate of the class conditional probabilities $\underset{\mathbf{x}\sim\mathcal{D}_{\mathbb{Y}|\mathbf{x}}}{\mathbb{P}}[y = c_k]$ (Bishop, 2006, Section 2.5.2), if these probabilities are constant in the neighborhood of $\mathbf{x}$.

Bias-variance trade-off for k-NN classifiers The number of neighbors k controls the bias-variance trade-off for the k-NN algorithm. In fact, when using one neighbor to decide an instance $\mathbf{x}$'s class, one takes more risk of potentially picking an example that does not represent the dominant class in $\mathbf{x}$'s neighborhood in the input space. As k grows, there's less risk that the closest data to $\mathbf{x}$ are from different classes, leading to a better estimation of $\mathbf{x}$'s class membership. Finally, in the extreme case where $k = m$, we see that $f_k(\mathbf{x})$ in Equation (1.46) becomes the empirical estimation of $\underset{y\sim\mathcal{D}_{\mathbb{Y}}}{\mathbb{P}}[y = c_k]$, *i.e.* the resulting classifier is constant, associating to a testing point $\mathbf{x}$ the class with most of instances in the training set. In this case, the flexibility of the classifier is lost leading to a high bias problem. From a theoretical point of view, generalization bounds for the nearest neighbors classifier were derived in several lines of work (Cover and Hart, 1967; Gottlieb et al., 2010; Shalev-Shwartz et al., 2010).

1.3 Model Selection

In most cases of interest, finding a classifier by solving an optimization problem requires some parameters to be fixed beforehand. For instance, for algorithms following the RRM rule (Section 1.1.4.3), the regularization multiplier λ has to be chosen, while for the kernel-based formulation of the SVM and the k-NN algorithms one has to select the kernel parameter (Table 1.3) and the number of the k neighbors, respectively. Such parameters are often called *hyperparameters*, and they reflect some prior knowledge one has about the problem at hand. They can also be viewed as additional degrees of freedom allowing to fine-tune the used model to obtain its best possible performance. For example, by increasing the regularization multiplier, one shrinks the search space towards a prior classifier (equal to zero in most of the cases).

For any learning algorithm, a given configuration of its hyperparameters leads to a different instance of the underlying optimization problem, and selecting the "good" configuration w.r.t. to some performance measure is an instance of the *model selection* problem. The most commonly used approach is *cross-validation* (Stone, 1974) that consists in dividing the available data into training, validation and testing sets. Then the first set is used to learn a hypothesis, and this latter's performance is assessed on the second set as a guide in the selection of the hyperparameters. The final evaluation of such an approach is done on the testing set which has been used neither for training, nor for the selection of hyperparameters. When data is scarce, and when no particular order is assumed on the training set, other model selection methods make the most of the available data by considering different training-validation splits after isolating a testing set. For example, the *k-fold* procedure (for $k \in \mathbb{N}^*$) is widely used and consists in dividing the rest of the data into k sets: each fold consists in selecting the k^{th} set as a validation set and training on the union of the remaining $k - 1$ sets. The hyperparameters are selected w.r.t. the average validation set performance over the k folds. For more details on model selection, we refer the interested reader to Hastie et al. (2001, Chapter 7) and Arlot et al. (2010).

1.4 Learning with $(\epsilon, \gamma, \tau)-$Good Similarity Functions

In this section, we present the general theory of learning with similarity functions, proposed in Balcan et al. (2008b,a).

1.4.1 From kernels to similarity functions

In Section 1.2.6.1, we saw that the dual formulation of the SVM allows to induce the class membership of a given instance, by calculating a linear combination of the inner products between this instance and the rest of the training data, eventually after mapping the data to a new space via a mapping $\phi : \mathbb{X} \to \mathbb{V}$, where $\mathbb{V}$ is some RKHS. Let us denote by I_+ (resp. I_-) the set of indices of points having class $y = 1$ (resp. $y = -1$), then Equation (1.44) can be re-written as:

$$\langle \mathbf{w}^*, \phi(\mathbf{x}) \rangle = \underbrace{\sum_{i \in I_+} \alpha_i K(\mathbf{x}_i, \mathbf{x})}_{\substack{\text{similarity to instances} \\ \text{with a positive label}}} - \underbrace{\sum_{i \in I_-} \alpha_i K(\mathbf{x}_i, \mathbf{x})}_{\substack{\text{similarity to instances} \\ \text{with a negative label}}}. \tag{1.47}$$

Hence, the scoring function at example $\mathbf{x}$ takes the form of a difference between two weighted sums of inner products in the RKHS: the first inner products are between $\mathbf{x}$ and instances labeled as 1, whereas the second ones are for those with the opposite label. But since any inner product can be seen as a measure of similarity[3], the expression above takes a more intuitive interpretation that suggests inferring the class of an example by comparing its similarity to instances with positive labels with its similarity to those with negative labels, where the similarity to each group of instances is aggregated as a weighted sum.

It is this last intuition that lies at the heart of the seminal papers of Balcan et al. (2008b,a), which formalize the idea of deciding an instance's class via comparison to other instances, using similarity functions that need not to be kernels. Although the two mentioned papers contain several mathematical formalizations of this idea, we only present those introduced in the most recent version provided in Balcan et al. (2008a). To this end, let us consider as a similarity function any function defined over $\mathbb{X} \times \mathbb{X}$ with values in $[-1, 1]$. This definition is clearly weaker than that of a kernel (Definition 1.2.8) as, for instance, it does not impose PSD constraints (Theorem 1.2.5). The definition of an intuitively good similarity function involves the margin violation loss and the convex hinge loss, that we now detail.

1.4.2 Goodness in Margin Violation Loss

We start with a definition of goodness in margin violation loss.

Definition 1.4.1. *(Balcan et al., 2008a, Definition 6) A similarity function K is (ϵ, γ, τ)-good for distribution $\mathcal{D}$ if there exists a (probabilistic) indicator function R of a set of "reasonable points" such that:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}} \left[y \cdot g(\mathbf{x}) < \gamma \right] \leq \epsilon, \tag{1.48}$$

$$\mathbb{P}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}} \left[R(\mathbf{x}') = 1 \right] \geq \tau, \tag{1.49}$$

where $g(\mathbf{x}) = \mathbb{E}_{(\mathbf{x}',y')\sim\mathcal{D}} [y' K(\mathbf{x}, \mathbf{x}') | R(\mathbf{x}') = 1]$.

[3]The inner product between any two vectors is maximized when they are co-linear with a positive co-linearity coefficient, as suggested by the Cauchy-Schwarz inequality.

For a given $\mathbf{x} \in \mathbb{X}$, the condition inside the probability in Equation (1.48) can be also re-written after conditioning on class y of $\mathbf{x}$. Denoting $p_y := \mathop{\mathbb{P}}_{(\mathbf{x}',y')\sim\mathcal{D}}[y' = y \wedge R(\mathbf{x}')]$ the probability of reasonable points that have label y, the statement of Equation (1.48) is equivalent to saying that at least a $1-\epsilon$ fraction of instances verify

$$\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}}\left[K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}'),y'=y\right]p_y \geq \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}}\left[K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}'),y'\neq y\right](1-p_y)+\gamma. \quad (1.50)$$

If the landmarks are evenly distributed between classes, *i.e.* $p_y = \frac{1}{2}$, then Equation (1.50) is equivalent to the following condition

$$\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}}\left[K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}'),y'=y\right] \geq \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}}\left[K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}'),y'\neq y\right]+2\gamma. \quad (1.51)$$

Simply put, for almost all of the instances (at least a $1-\epsilon$ fraction), an instance $\mathbf{x}$ is on average more similar (in the sense of K) to reasonable instances having the same label, than to those of opposite label, by a large margin (at least 2γ). Hence, the above definition rigorously captures the intuition presented at the beginning of this section. Another view for this definition is that ϵ is an upper bound for the expected margin violation loss of classifier g given at the end of Definition 1.4.1.

The next result shows how good similarity functions can be used for linear classification in a manner that is reminiscent of the kernel trick.

Theorem 1.4.1. *(Balcan et al., 2008a, Theorem 8) Let K be an $(\epsilon,\gamma,\tau)-$good similarity function for a distribution $\mathcal{D}$. For any $\delta > 0$, let $L = \{\mathbf{x}'_1,...,\mathbf{x}'_{n'}\} \sim \mathcal{D}_{\mathbb{X}}^{n'}$ be an unlabeled sample of size $n' = \frac{2}{\tau}\log\left(\frac{2}{\delta}\right)\left(1+\frac{8}{\gamma^2}\right)$ of landmarks. Consider the mapping:*

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

Then with a probability at least $1-\delta$ over the draw of L, there exists $\mathbf{w} \in \mathbb{R}^{n'}$ such that $\|\mathbf{w}\|_1 = 1$ and

$$\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y\left\langle\mathbf{w},\phi^L(\mathbf{x})\right\rangle < \frac{\gamma}{2}\right] \leq \epsilon+\delta. \quad (1.52)$$

In other words, the data distribution induced by ϕ^L has a linear separator achieving margin violation risk at most $\epsilon+\delta$ at margin $\frac{\gamma}{2}$.

What the previous theorem states informally is that, given a good similarity function, one can draw a sufficiently large sample L and map the data into a new space where their features are similarities to the points of L. Then with a high probability over the draw of L, the new distribution induced by ϕ^L has a linear separator achieving an error at most $\epsilon+\delta$ at margin γ. The required number of drawn points is inversely proportional to the fraction of landmarks τ, which is an intuitive behavior. Indeed, the previous result considers that one ignores the landmarks distribution. Consequently, the lower τ is, the rarer they become, and the more instances one needs to draw.
The result of the previous theorem is reminiscent of the kernel trick, in the sense that the data is mapped to a new space where the classes hopefully become linearly separable. However, all it requires is an intuitive goodness condition rather than the additional requirement for it to be a kernel. Moreover, even if the proportion of the landmarks is unknown, the theorem guarantees that when enough of them are drawn, the classification risk of the induced classifier is small.

1.4.3 Convexification with Hinge Loss

From a practical point of view, having access to a sample S, we can draw $L \sim \mathcal{D}_{\mathbb{X}}^{n'}$ as a subset of S and then look for a classifier which minimizes the empirical probability of margin violation in the ϕ^L space as follows

$$\min_{\mathbf{w}\in\mathbb{R}^{n'}} \frac{1}{m}\sum_{i=1}^{m}[y_i \left\langle \mathbf{w}, \phi^L(\mathbf{x}_i)\right\rangle < \gamma]. \tag{1.53}$$

However, as we mentioned in Section 1.2.2, this minimization problem is NP-hard to solve. To this end, more algorithm-friendly notions of similarity goodness have been introduced in Balcan et al. (2008a), in which the hinge loss is considered as a convex surrogate to the margin violation loss.

Definition 1.4.2. *(Balcan et al., 2008a, Definition 7) A similarity function K is (ϵ, γ, τ)-good in hinge loss for problem (distribution) $\mathcal{D}$ if there exists a (probabilistic) indicator function R of a set of "reasonable points" such that:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[l_+\left(\frac{g(\mathbf{x})}{\gamma}, y\right)\right] \leq \epsilon, \tag{1.54}$$

$$\mathop{\mathbb{P}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}}\left[R(\mathbf{x}') = 1\right] \geq \tau, \tag{1.55}$$

where $g(\mathbf{x}) = \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}}[y'K(\mathbf{x}, \mathbf{x}')|R(\mathbf{x}') = 1]$.

In line with Theorem 1.4.1, the following theorem guarantees the existence of a classifier given a good similarity function in hinge loss.

Theorem 1.4.2. *(Balcan et al., 2008a, Theorem 11) Let K be an $(\epsilon, \gamma, \tau)-$good similarity function in hinge loss for a learning problem $\mathcal{D}$. For any $\epsilon_1 > 0$ and $0 < \delta < \frac{\gamma\epsilon_1}{4}$, let $L = \{\mathbf{x}'_1, ..., \mathbf{x}'_{n'}\} \sim \mathcal{D}_{\mathbb{X}}^{n'}$ be a (potentially unlabeled) sample of size $n' = \frac{2}{\tau}\log\left(\frac{2}{\delta}\right)\left(1 + \frac{16}{(\epsilon_1\gamma)^2}\right)$ of landmarks. Consider the mapping ϕ^L from Theorem 1.4.1. Then with a probability at least $1-\delta$ over the draw of L, there exists $\mathbf{w} \in \mathbb{R}^{n'}$ such that $\|\mathbf{w}\|_1 = 1$ and:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[l_+\left(\frac{\left\langle \mathbf{w}, \phi^L(\mathbf{x})\right\rangle}{\gamma}, y\right)\right] \leq \epsilon + \epsilon_1. \tag{1.56}$$

In other words, the distribution induced by ϕ^L has a linear separator achieving hinge risk at most $\epsilon + \epsilon_1$ at margin γ.

Given an $(\epsilon, \gamma, \tau)-$good similarity function K in hinge loss, and a drawn sample $L \sim \mathcal{D}_{\mathbb{X}}^{n'}$, the previous theorem motivates an algorithm for linear classification in the similarity induced space defined by mapping ϕ^L:

$$\min_{\mathbf{w}\in\mathbb{R}^{n'}} \quad \sum_{i=1}^{m} l_+\left(\left\langle \mathbf{w}, \phi^L(\mathbf{x}_i)\right\rangle, y_i\right) \tag{1.57}$$

$$\text{subject to} \quad \|\mathbf{w}\|_1 \leq \frac{1}{\gamma}. \tag{1.58}$$

This is a linear program and the $1-$norm constraint ensures that the solution is sparse, meaning that in order to infer a test example's class, one needs to compare it to few training examples. One can see here an analogy with the support vectors of the SVM algorithm Section 1.2.6.1.

1.4.4 Connection to Similarity Learning

Both kernels and metrics, used respectively by the SVM and the k-NN algorithms, reflect a certain similarity measure between instances. For kernels, they are similarity measures as we explained in Section 1.4.1, whereas for metrics they are dissimilarity measures. Nevertheless, we will consider both as similarity measures. These latter are traditionally fixed beforehand and might be unrepresentative of the geometry of the considered classification problem. To tackle this limitation, metric learning and kernel learning have emerged as fields with the aim of constructing a similarity measure that make instances of the same class close to each other, while separating those which are not. We now briefly describe some main ideas from kernel learning and metric learning approaches for classification.

1.4.4.1 Kernel Learning

Our brief discussion here concerns *Kernel Learning* (Abbasnejad et al., 2012) and *Multiple Kernel Learning* (Gönen and Alpaydın, 2011). For binary classification, these approaches consider the *ideal kernel* matrix $\mathbf{y}\mathbf{y}^T \in \mathbb{R}^{m\times m}$, where $\mathbf{y}^T := (y_1, \ldots, y_m)$ is the vector of data labels. The term "ideal" here is due to the fact that such a matrix perfectly reflects similarity between instances of the same label and dissimilarity in the opposite case, *i.e.* for $y, y' \in \{-1, 1\}$, $yy' = 1$ if $y = y'$, and $yy' = -1$ otherwise. In this case, the learned kernel represented by the matrix $\mathbf{K}$ is obtained by aligning it with the ideal kernel $\mathbf{y}\mathbf{y}^T$, for example by maximizing the *kernel alignment measure* defined as

$$\frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^T \rangle}{\|\mathbf{K}\|_2 \|\mathbf{y}\mathbf{y}^T\|_2},$$

while ensuring PSD constraints. Other criteria relating a combination (not necessarily linear) of predefined kernel matrices and the ideal kernel matrix are reviewed in Gönen and Alpaydın (2011). The goal of such approaches is to ensure that the learned similarity function reflects, in the best way possible, the relationships between the instances and their labels. However, the criteria that they optimize and the constraints that they enforce are different from the ones concerned by $(\epsilon, \gamma, \tau)-$good similarity functions, as the latter are based on the classification margin and do not require positive-semidefiniteness.

1.4.4.2 Metric Learning

The idea of *Metric Learning* was first introduced in Xing et al. (2003), and has gained the attention of the machine learning community in the following years, as detailed in the surveys Kulis (2013); Bellet et al. (2013). The goal of metric learning is to learn a pseudo-metric tailored to the data at hand. This is done under constraints that can be divided into 3 main groups:

Must-link constraints reflect the requirement for instances of the same class to be close to each other. They concern pairs of data and are represented by a set

$$C = \{(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{X}^2; \quad \mathbf{x}_1 \text{ and } \mathbf{x}_2 \text{ are close to each other}\}.$$

Such constraints are forced, for example, by including a sum of the quantities $d(\mathbf{x}_i, \mathbf{x}_j)$ to the objective function to minimize.

Cannot link constraints enforce the separability between the different classes via a set

$$F = \{\ (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{X}^2; \quad \mathbf{x}_1 \text{ and } \mathbf{x}_2 \text{ are far from each other}\}.$$

Similarly to the previous type of constraints, such constraints can be forced by maximizing a sum of the quantities $d(\mathbf{x}_i, \mathbf{x}_j)$.

Relative constraints relate the closeness of an instance to those of its own class and those of opposite classes. Hence, unlike the previous two groups, they are represented by triplets of data:

$$R = \{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \in \mathbb{X}^3, \quad \mathbf{x}_i \text{ is closer to } \mathbf{x}_j \text{ than to } \mathbf{x}_k\}.$$

One can introduce these constraints by imposing that $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k)$, or a stronger condition involving a margin $\rho > 0$: $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) - \rho$.

A metric learning algorithm outputs a function d that satisfies some of these constraints. For example, the Large Margin Nearest Neighbors (LMNN) (Weinberger et al., 2006) algorithm is formulated to satisfy the first and third groups of constraints, whereas in the pioneering work of Xing et al. (2003) and in Information Theoretic Metric Learning (Davis et al., 2007), rather the two first groups are used.

We observe a link between $(\epsilon, \gamma, \tau)-$good similarities and the relative constraints in their form involving a margin. In fact, if d is a metric, then $K_d := -d$ can be seen as a similarity function, and the margin condition is re-written:

$$K_d(\mathbf{x}_i, \mathbf{x}_j) \geq K_d(\mathbf{x}_i, \mathbf{x}_k) + \rho.$$

This requirement is analogous in form to the one in Equation (1.51), with the exception that this latter holds only on average, hence reducing the number of constraints.
A considerable number of metric learning algorithms rely on learning a Mahalanobis pseudo-metric $d_{\mathbf{M}}$ (Mahalanobis, 1936; Ishikawa et al., 1998) defined as:

$$\begin{aligned} d_{\mathbf{M}} :& \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+ \\ & (\mathbf{x}, \mathbf{x}') \mapsto \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')}, \end{aligned} \tag{1.59}$$

where $\mathbf{M}$ is a PSD matrix. Learning such a pseudo-metric entails PSD constraints, which can be intractable for large scale problems as the projection of a symmetric matrix onto the PSD cone requires an eigenvalue decomposition (Higham, 1988) with a $\mathcal{O}(n^3)$ cost. This issue is not inherent for $(\epsilon, \gamma, \tau)-$good similarity functions, as they do not require such a condition. Finally, the predominant majority of metric learning algorithms are designed to be used with k-NN classifier, whereas $(\epsilon, \gamma, \tau)-$good similarities are naturally fit for linear classification as pointed out in Theorem 1.4.1 and Theorem 1.4.2.

1.4.5 Some Notable Works Based on $(\epsilon, \gamma, \tau)-$Good Similarities

In this section, we present several works that built upon the seminal papers of Balcan et al. (2008b,a).

Bilinear similarity learning for sparse linear classification In Bellet et al. (2012), the authors present an approach that consists in learning an $(\epsilon, \gamma, \tau)-$good bilinear similarity function and then learning a linear classifier in its induced similarity space. To learn the similarity function, they solve the following problem:

$$\min_{\mathbf{A} \in \mathbb{R}^{n \times n}} \sum_{i=1}^{m} l_+ \left(\sum_{l=1}^{L} y_l K_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_l), y_i \right) + \beta \|\mathbf{A}\|_2^2, \tag{1.60}$$

where $\beta > 0$ is a regularization parameter. The authors use stability theory (Section 1.2.5.3) in order to derive generalization guarantees on the goodness of the obtained similarity function. They further argue that the obtained classifier is sparse requiring a comparison to only a few data points to infer the test point's label.

Guaranteed regularized classification Guo and Ying (2014) provided a theoretical analysis for a problem similar to that of Bellet et al. (2012), where the regularization matrix norm is not restricted only to the Frobenius case, and the matrix parameter $\mathbf{A}$ of the bilinear similarity is assumed to be symmetric. This is formalized in the following objective function:

$$\min_{\mathbf{A}\in\mathbb{S}^{n\times n}} \sum_{i=1}^{m} l_+ \left(\sum_{l=1}^{L} y_l K_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_l), y_i \right) + \beta \|\mathbf{A}\|. \tag{1.61}$$

They derive generalization bounds based on the Rademacher complexity for both the similarity function's risk and for its induced classifier. In particular, they show that the generalization error of the similarity learning gives an upper bound for the generalization error of its induced linear classifier.

Jointly learning the similarity function and induced classifier The context of this work, presented in Nicolae et al. (2015), is semi-supervised classification. The latter a branch of machine learning where contrary to the supervised setting, the training set is only partially labeled. The authors derive an algorithm for jointly learning a similarity function and its associated classifier via alternating optimization by solving

$$\begin{aligned} \min_{\substack{\mathbf{w}\in\mathbb{R}^L \\ \mathbf{A}\in\mathbb{R}^{n\times n}}} \quad & \sum_{i=1}^{m} l_+ \left(\sum_{l=1}^{L} w_l K_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_l), y_i \right) + \lambda \|\mathbf{A} - \mathbf{R}\| \\ \text{subject to} \quad & \|\mathbf{w}\|_1 \leq \frac{1}{\gamma} \\ & \mathbf{A} \text{ diagonal}, \quad |(\mathbf{A})_{kk}| \leq 1 \quad \forall 1 \leq k \leq d, \end{aligned} \tag{1.62}$$

where $\lambda > 0$ is a regularization hyperparameter, and m and m' are respectively the sizes of labeled and unlabeled samples at hand. The authors also derived generalization guarantees on the performance of both the bilinear similarity parameter $\mathbf{A}$ and the vector $\mathbf{w}$ based on the Rademacher complexity.

Conclusion

In this chapter, we reviewed the supervised learning setting with an emphasis on large margin binary classification. We started by formalizing the supervised learning problem in general, before providing more details for the task of classification, including a review of various loss functions, the notions of scoring functions, classification margin and generalization bounds. We further presented the SVM and the k-NN algorithms with an emphasis on the former due to its tight link with large margin classification. We then linked it to learning with $(\epsilon, \gamma, \tau)-$good similarity functions, presented the latter, and highlighted their connection to the metric and kernel learning fields.

Chapter 2

Domain Adaptation

Abstract The previous chapter concerned one of the most considered scenarios in machine learning, namely supervised learning. Nevertheless, the theoretical foundations of the latter do not cover some real-world problems where data-generating processes differ between the training and testing sets, corresponding to the so-called source and target domains. The latter case is addressed by the domain adaptation (DA) field that we review in this chapter. We start by the theoretical formalization of the domain adaptation problem in the particular case where there is only one source domain, when the feature and output spaces remain the same for both domains, and when the target set's labels are totally unavailable at the training time. We further cover sufficient and necessary conditions for the success of adaptation. This is followed by a non-exhaustive presentation of DA learning algorithms that were proposed in the literature.

Introduction

In the previous chapter, the training and testing data related to learning a given task were supposed to come from the same probability distribution over the same feature space. In such a case, learning a good hypothesis is a matter of choosing the "right" hypothesis space and eventually increasing the size of the learning sample (Section 1.1.4.2) when the latter is not representative enough of the underlying distribution. However, in many real-world applications, the data-generating process if often subject to change, as illustrated by the following examples:

Visual recognition In the healthcare sector, machine learning algorithms often are trained on scanner images manually labeled by human experts in order to help doctors in detecting certain diseases. However, imaging technologies improve over time in terms of resolution, contrast, and other characteristics of the acquired images. As a result, hospitals may buy a new scanner over the course of years. In this case, the trained algorithms may not work well on the newly generated images and may require costly manual labeling all over again. Yet, a better solution would be to find a way to leverage the already available related labeled data. Similarly to images, one can also think about systems that recognize movements from video streams, for which the performance can worsen if deployed in noisy real-world environments, after being trained in controlled laboratory ones.

Sentiment analysis Administrators of an e-commerce website specializing in book sales managed to develop efficient classification algorithms to help to categorize the customers' product reviews into negative or positive ones. Such an algorithm would not have come into existence if it weren't for the effort that the administrators have put into labeling the reviews to constitute training data. Aiming at expanding their

business, they add the new DVD category to their products and would like to automatize the classification process for this family of products too. Alas, using the classifier trained only on book reviews results in poor performance on DVD ones.

User specific applications Several software applications leverage information provided by the user to work efficiently, such as spam filtering or speech recognition applications. The variety of user-profiles inevitably causes a difference between the data they generate: For example, the spam e-mails that target a professional address differ from those targeting personal ones. Likewise, a speech recognition application may need to adapt to the different voices and accents of the users. Without an adaption of the underlying learning algorithm to each user, one cannot hope for a good performance given the significant disparities between user profiles.

In all of the previous scenarios, it would not be an exaggeration to qualify the same distribution assumption for the training and testing data as unrealistic. As a matter of fact, the change of measuring instruments, data acquisition environment, or even the sampling method (sample selection bias (Storkey, 2009; Moreno-Torres et al., 2012)) induce a change on the joint distribution of the inputs and the outputs, and can even change their respective input and output spaces.

Given such a setting, two solutions are to consider: either to manually label the newly acquired data at the expense of time and resources or to notice that, despite the change of the data-generating distributions, they may remain related. After all, a human expert can easily classify medical images even when they have slightly different visual characteristics. In the product review example, some words can express contentment or dissatisfaction regardless of the product type. As for the third example, commonalities between different user profiles may be leveraged once correctly identified. Overall, these examples suggest that making use of the relatedness between the different distributions, in order to transfer knowledge acquired on the training data to the testing data, seems to be a much more intelligent solution when compared to learning each task from scratch.

Transferring the knowledge extracted from a given domain for it to serve in another one lies at the heart of the field of *transfer learning*, where the training and testing data, associated respectively with *source* and *target* domains, are assumed to be drawn from different probability distributions. For the sake of clarity, we recall the formalization of the terms domain, task, and transfer learning given by Pan and Yang (2010) below:

Domain A pair $(\mathbb{X}, \mathcal{D}_{\mathbb{X}})$ where $\mathbb{X}$ is an input space and $\mathcal{D}_{\mathbb{X}}$ is a probability distribution over $\mathbb{X}$.

Task A pair $(\mathbb{Y}, \mathcal{D}_{\mathbb{Y}|\mathbf{x}})$ where $\mathbb{Y}$ is an output space and $\mathcal{D}_{\mathbb{Y}|\mathbf{x}}$ is the conditional probability of outputs $y \in \mathbb{Y}$ given an input instance $\mathbf{x}$.

Transfer learning Given a source domain $(\mathbb{X}_S, \mathcal{S}_{\mathbb{X}})$ and a target domain $(\mathbb{X}_T, \mathcal{T}_{\mathbb{X}})$, with respective associated tasks $(\mathbb{Y}_S, \mathcal{S}_{\mathbb{Y}|\mathbf{x}})$ and $(\mathbb{Y}_T, \mathcal{T}_{\mathbb{Y}|\mathbf{x}})$, the goal of transfer learning is to learn the target task (possibly with a hypothesis space $\mathbb{H}$) using the available empirical knowledge related to the previous domains and tasks.

We note that we slightly modified the transfer learning definition from (Pan and Yang, 2010). In the latter, the goal is to learn the target task using only the source information, whereas, as we will see later, target domain information can also sometimes be leveraged, for example, through an unlabeled target sample.

In this manuscript, we consider a particular case of transfer learning defined by the following conditions:

1. Both tasks are related and share the same output space, *i.e.* $\mathbb{Y}_S = \mathbb{Y}_T = \mathbb{Y}$. This defines the classic[1] *domain adaptation* (DA) setting.
2. The labels are observed only for the source domain. This corresponds to the *unsupervised* [2] *DA* setting.
3. The source and target instances are represented in the same feature space, *i.e.* $\mathbb{X}_S = \mathbb{X}_T = \mathbb{X}$. This assumption is what we refer to as *homogeneous* DA.
4. We consider only one source domain, *i.e.* *single-source* DA, as opposed to *multi-source* DA (Sun et al., 2015; Zhao et al., 2020).

To summarize, we consider *single-source homogeneous unsupervised DA*.

Remark Unless specified otherwise, we simply use the term "DA" to describe the restricted DA setting satisfying the assumptions described above. We also use the "domain" terminology to refer to a joint distribution over $\mathbb{X} \times \mathbb{Y}$, as opposed to the definition from Pan and Yang (2010). For a taxonomy of the different transfer learning settings, we refer the interested reader to Pan and Yang (2010); Zhuang et al. (2019); Redko et al. (2020).

2.1 Theoretical Framework

We formalize the DA setting that we consider in the rest of the manuscript. We suppose that the source and target domains have joint distributions $\mathcal{S}$ and $\mathcal{T}$ over $\mathbb{X} \times \mathbb{Y}$, where $\mathbb{X}$ and $\mathbb{Y}$ are the same as in Chapter 1. The only information we have about these distributions is given by the observed samples that they generate: $S = \{(\mathbf{x}_{s,i}, y_{s,i})\}_{i=1}^{m_s} \sim \mathcal{S}^{m_s}$ and $T_u = \{\mathbf{x}_{t,j}\}_{j=1}^{m_t} \sim \mathcal{T}_{\mathbb{X}}{}^{m_t}$, where the latter target sample is unlabeled, hence it is sampled from the marginal distribution $\mathcal{T}_{\mathbb{X}}$. Given a hypothesis class $\mathbb{H}$ and a loss function l, the goal is to achieve a low $l - risk$ on the target distribution in spite of the absence of labeled data needed to learn a good classifier in the supervised setting (Chapter 1).

Intuitively, even if the source and target distributions are different, they should be somehow related for adaptation to be successful. Several questions are then to be answered:

- What does it mean for two domains to be similar or related? Can we measure such relatedness? If so, how to measure it?
- Since we lack access to the target distribution labels, is it sufficient to align the marginal distributions and to somehow leverage the source labels information?

In what follows, we address the previous questions by formalizing the notion of relatedness between the domains. In particular, we will detail what it means for the source and the target domains to be related, where relatedness is expressed via some assumptions on the marginals and the conditional probabilities of the labels. Then, we will review some results showing the sufficiency or the necessity of some of these assumptions for successful adaptation, *i.e.* for finding a model with a low risk on the target domain.

2.1.1 Domain Relatedness Assumptions

We discuss some commonly made assumptions that formalize the relatedness of the source and target domains, and that are considered as those contributing to the success of adaptation. They can be divided into two groups: the first concerns the feature marginal distributions $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$, whereas the second concerns the output conditional distributions $\mathcal{S}_{\mathbb{Y}|\mathbf{x}}$ and $\mathcal{T}_{\mathbb{Y}|\mathbf{x}}$ for a given instance $\mathbf{x} \in \mathbb{X}$.

[1] Recently, variants where $\mathbb{Y}_S \neq \mathbb{Y}_T$ were considered, as in *open set DA* (Panareda Busto and Gall, 2017).

[2] Different from unsupervised transfer learning where labels are unavailable in both domains.

2.1.1.1 Relating Input Marginal Distributions

Dominance of the target by the source This assumption can be formalized as $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$, *i.e.* $\mathcal{T}_{\mathbb{X}}$ is absolutely continuous w.r.t. $\mathcal{S}_{\mathbb{X}}$ (Definition A.2.6). In this case, one can define the *Radon-Nikodym* (Nikodym, 1930) derivative[3] $\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}$ (Theorem A.2.2) studied, for example, in Cortes et al. (2010), in which the authors consider two cases depending on whether $\sup_{\mathbf{x}\in\mathbb{X}} \frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) = \infty$ or not. This assumption is at the heart of instance re-weighting domain adaptation approaches covered further in Section 2.2.1.1.

Bounded-weight ratio In the previous assumption, the case where $\sup_{\mathbf{x}\in\mathbb{X}} \frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) < \infty$ is equivalent, as long as we also have $\mathcal{S}_{\mathbb{X}} \ll \mathcal{T}_{\mathbb{X}}$, to $\inf_{\mathbf{x}\in\mathbb{X}} \frac{\mathrm{d}\mathcal{S}_{\mathbb{X}}}{\mathrm{d}\mathcal{T}_{\mathbb{X}}}(\mathbf{x}) > 0$. The latter condition implies that

$$\inf_{\substack{b\in\mathfrak{F}\\ \mathcal{T}_{\mathbb{X}}(b)>0}} \frac{\mathcal{S}_{\mathbb{X}}(b)}{\mathcal{T}_{\mathbb{X}}(b)} > 0,$$

where $\mathfrak{F}$ is the collection of subsets of $\mathbb{X}$ that are measurable w.r.t. $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$. A relaxation of this assumption was studied in (Ben-David and Urner, 2012, 2014) by restricting the infimum to a predefined family of subsets $\mathfrak{B}$, leading to the *weight ratio* definition given as follows:

Definition 2.1.1 (Weight ratio)**.** *Let $\mathfrak{B} \subset 2^{\mathbb{X}}$ be a collection of subsets of $\mathbb{X}$, measurable w.r.t $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$. The $\eta-$weight ratio $C_{\mathfrak{B},\eta}$ for some $\eta > 0$ and the weight ratio $C_{\mathfrak{B}}$ between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ are defined as:*

$$C_{\mathfrak{B},\eta} := \inf_{\substack{b\in\mathfrak{B}\\ \mathcal{T}_{\mathbb{X}}(b)\geq\eta}} \frac{\mathcal{S}_{\mathbb{X}}(b)}{\mathcal{T}_{\mathbb{X}}(b)}, \qquad C_{\mathfrak{B}} := \inf_{\substack{b\in\mathfrak{B}\\ \mathcal{T}_{\mathbb{X}}(b)>0}} \frac{\mathcal{S}_{\mathbb{X}}(b)}{\mathcal{T}_{\mathbb{X}}(b)}. \tag{2.1}$$

The bounded-weight ratio assumption then becomes $C_{\mathfrak{B},\eta} > 0$ or $C_{\mathfrak{B}} > 0$. This ratio has the virtue of being estimable from finite samples, if it corresponds to sets that are supports of binary hypotheses of a hypothesis space with a finite VC dimension, as pointed out in Ben-David and Urner (2012).

Similarity Different from only assuming a certain dominance relation between distributions, in this case, the domain marginals are supposed to be similar in the sense of some divergence measure. Two main families of dissimilarity measures to compare probability distributions are *$\phi-$divergences* (Csiszár, 1967) and *Integral Probability Metrics* (IPM) (Zolotarev, 1984). While the first supposes that one distribution dominates the other in order for its values to be finite, the IPM does not need such an assumption, and is a popular choice in establishing DA theoretical bounds as well as in proposing new algorithms. We will detail some IPMs in Section 2.1.2.

2.1.1.2 Relating Output Conditional Distributions

Covariate shift Any probability $\mathcal{D}$ over $\mathbb{X} \times \mathbb{Y}$ can be decomposed as $\mathcal{D} = \mathcal{D}_{\mathbb{Y}|\mathbf{x}}\mathcal{D}_{\mathbb{X}}$. Hence, the condition $\mathcal{S} \neq \mathcal{T}$ implies that either $\mathcal{S}_{\mathbb{X}} \neq \mathcal{T}_{\mathbb{X}}$ or $\mathcal{S}_{\mathbb{Y}|\mathbf{x}} \neq \mathcal{T}_{\mathbb{Y}|\mathbf{x}}$. The covariate shift assumption corresponds to the first case with the equality of the conditional distributions. In other words, the domain shift is only due to a shift in the distribution of the covariates (the features) as described by the following conditions:

$$\mathcal{S}_{\mathbb{X}} \neq \mathcal{T}_{\mathbb{X}} \quad \text{and} \quad \mathcal{S}_{\mathbb{Y}|\mathbf{x}} = \mathcal{T}_{\mathbb{Y}|\mathbf{x}}. \tag{2.2}$$

[3]For probabilities with density functions, this is the density ratio and for probabilities over discrete domains, this is the ratio of probability masses.

This assumption is related to the sample selection bias problem (Storkey, 2009; Moreno-Torres et al., 2012) and, when combined with an additional assumption of dominance of the target by the source, suggests approaches that align the source and target domains via instance re-weighting (Section 2.2.1.1).

$\lambda-$shift This assumption was introduced in Mansour and Schain (2014) based on linking the labels' conditional distributions in the two domains as follows.

Definition 2.1.2. *Let $\mathbb{Y} = \{c_1, \dots, c_K\}$ and let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two probability distributions over $\mathbb{Y}$. $\mathcal{D}_1$ is said to be $\lambda-$shift w.r.t. $\mathcal{D}_2$ if for all $1 \leq k \leq K$,*

$$(1-\lambda) \mathop{\mathbb{P}}_{y\sim\mathcal{D}_2}[y = c_k] \leq \mathop{\mathbb{P}}_{y\sim\mathcal{D}_1}[y = c_k] \leq \mathop{\mathbb{P}}_{y\sim\mathcal{D}_2}[y = c_k] + \lambda(1 - \mathop{\mathbb{P}}_{y\sim\mathcal{D}_2}[y = c_k]). \quad (2.3)$$

For $\lambda = 0$, this implies that $\mathcal{D}_1 = \mathcal{D}_2$, while for $\lambda = 1$, it reduces to a trivial inequality on the range of probability values. If these two probabilities are chosen to be the conditional label distributions $\mathcal{S}_{\mathbb{Y}|\mathbf{x}}$ and $\mathcal{T}_{\mathbb{Y}|\mathbf{x}}$, then Definition 2.1.2 reduces to the covariate shift assumption. Overall, the $\lambda-$shift assumption is a relaxation of the covariate shift assumption that was used in Mansour and Schain (2014) to bound the risk on the target domain in the framework of algorithmic robustness (Section 1.2.5.3).

Low ideal joint error Another weak requirement of the similarity between labeling functions is the existence of a hypothesis h^*, called the *ideal joint hypothesis*, performing well on both domains. It was introduced in Ben-David et al. (2007), and is defined as:

$$h^* \in \arg\min_{h\in\mathbb{H}} \mathfrak{E}^{01}_{\mathcal{S}}(h) + \mathfrak{E}^{01}_{\mathcal{T}}(h), \quad (2.4)$$

where $\mathbb{H}$ is a binary hypothesis space. The authors argue that this term is non-estimable as it requires target domain labels, and it should be small to hope for the adaptation to be successful. The exact role of this assumption will become clearer in Theorem 2.1.6.

Low disagreement between best hypotheses A somewhat different relaxation of the covariate shift assumption, it consists in requiring from the labeling functions[4] to be close to each other, with closeness being expressed via the disagreement between the two functions, as defined below.

Definition 2.1.3 (Mean $l-$disagreement between two hypotheses)**.** *Given two hypotheses h_1 and h_2, their $l-$disagremment over a distribution $\mathcal{D}$ defined on $\mathbb{X} \times \mathbb{Y}$ is:*

$$\mathfrak{E}^l_{\mathcal{D}}(h_1, h_2) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[l(h_1(\mathbf{x}), h_2(\mathbf{x}))]. \quad (2.5)$$

We note that the mean $l-$disagreement is an extension of the $l-$risk definition Definition 1.1.1: if a domain $\mathcal{D}$ has a deterministic labeling function $f_{\mathcal{D}}$, then for any $h \in \mathbb{H}$, we have

$$\mathfrak{E}^l_{\mathcal{D}}(h, f_{\mathcal{D}}) = \mathfrak{E}^l_{\mathcal{D}}(h). \quad (2.6)$$

The low $l-$disagreement assumption was proven to be useful for adaptation in Mansour et al. (2009b), more exactly due to Theorem 2.1.7 later in this chapter.

Probabilistic Lipschitzness This relaxation of classic deterministic Lipschitzness of a function was theoretically studied in Urner et al. (2011) for semi-supervised learning and in Ben-David and Urner (2014) for DA. A modified version of this notion was also used in Courty et al. (2017). We give the definition used in both Ben-David and Urner (2014) and Courty et al. (2017) below.

[4] By the labeling function we generally mean hypotheses minimizing the misclassification risk in the corresponding domain.

Definition 2.1.4 (Probabilistic Lipschitzness)**.** *Given a metric space $(\mathbb{X}, d)$ and a function $\phi : \mathbb{R} \to [0, 1]$,*

1. *We say that $f : \mathbb{X} \to \mathbb{R}$ is ϕ-Lipschitz w.r.t. a distribution $\mathcal{P}$ over $\mathbb{X}$ if, for all $\lambda > 0$, we have*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{P}} \left[\exists \mathbf{x}' \in \mathbb{X}; \left|f(\mathbf{x}) - f(\mathbf{x}')\right| > \lambda d(\mathbf{x}, \mathbf{x}')\right] \leq \phi(\lambda). \tag{2.7}$$

2. *We say that $f : \mathbb{X} \to \mathbb{R}$ is ϕ-Lipschitz w.r.t. a distribution $\mathcal{P}$ over $\mathbb{X}^2$ and we write $f \in \mathrm{PTL}_\phi(\mathcal{P})$ if, for all $\lambda > 0$, we have*

$$\mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}} \left[\left|f(\mathbf{x}) - f(\mathbf{x}')\right| > \lambda d(\mathbf{x}, \mathbf{x}')\right] \leq \phi(\lambda). \tag{2.8}$$

Deterministic $L-$Lipschitzness is a particular case of the definition provided above as it corresponds to setting $\phi(\lambda) = [\lambda < L]$ and the generalization follows from choosing ϕ to be decreasing. A consequence of this property, when applied to a hypothesis h with continuous values, is that it tends to have the same behavior in high-density regions, thus having the same output in such regions with high probability. In fact, for binary hypotheses, deterministic $\lambda-$Lipschitzness implies that two points that are at most $\frac{1}{\lambda}$ away from each other must have the same label, whereas probabilistic Lipschitzness relaxes this requirement.

2.1.2 Assessing Divergence between the Feature Marginals

The similarity between the source and target domains can be assessed by comparing their probability distributions. As one does not have access to the labels on the target domain, their marginals $\mathcal{S}_\mathbb{X}$ and $\mathcal{T}_\mathbb{X}$ are usually compared. Several ways of comparing probability distributions have been proposed in the literature (Gibbs and Su, 2002) with a vast majority of such measures used in DA taking the form of an *Integral Probability Metric* (Zolotarev, 1984) defined below.

Definition 2.1.5 (Integral Probability Metric)**.** *Let $\mathbb{F}$ be a family of bounded functions that are measurable w.r.t $\mathcal{S}_\mathbb{X}$ and $\mathcal{T}_\mathbb{X}$. The Integral Probability Metric (IPM) associated to $\mathbb{F}$ between $\mathcal{S}_\mathbb{X}$ and $\mathcal{T}_\mathbb{X}$ is defined by*

$$\mathrm{IPM}_\mathbb{F}(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X}) = \sup_{f\in\mathbb{F}} \left| \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbb{X}} [f(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_\mathbb{X}} [f(\mathbf{x})] \right|.$$

For any $\mathbb{F}$ as in the definition given above, $\mathrm{IPM}_\mathbb{F}$ is a symmetric function of $\mathcal{S}_\mathbb{X}$ and $\mathcal{T}_\mathbb{X}$ and obeys the triangle inequality due to the sub-additivity of the supremum, making it a pseudo-metric (Definition A.1.1) over the set of probability distributions over $\mathbb{X}$. For some common choices of $\mathbb{F}$ (Sections 2.1.2.3 and 2.1.2.4), $\mathrm{IPM}_\mathbb{F}$ becomes a metric.

In what follows, we summarize the IPMs commonly used in DA in Table 2.1 for four common choices of $\mathbb{F}$. We highlight some of their properties and provide the generalization inequalities based on them.

Name	Function space $\mathbb{F}$	Assumptions
$\mathbb{H}$-divergence	$\mathbb{H}$: hypothesis space of binary classifiers	$\mathbb{H}$ has a finite VC dimension, labels in {0,1}
$l-$discrepancy	$\{\mathbf{x} \mapsto l(h(\mathbf{x}), h'(\mathbf{x})); h, h' \in \mathbb{H}\}$	$l : \mathbb{X} \to \mathbb{R}_+$ symmetric, triangle inequality
MMD	$\{f : \mathbb{X} \to \mathbb{R}; f \in \mathbb{V}; \|f\|_\mathbb{V} \leq 1\}$	$\mathbb{V}$ is an RKHS of a universal kernel
Wasserstein-1	$\{f : \mathbb{X} \to \mathbb{R}; f$ is 1-Lipchitz$\}$	$(\mathbb{X}, d)$ is a metric space

Table 2.1: Some notable Integral Probability Metrics used in DA.

2.1.2.1 $\mathbb{H}-$divergence

Introduced in one of the earliest theoretical studies of DA provided by Ben-David et al. (2010), the $\mathbb{H}-$divergence is a pseudo-metric between probability distributions $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$, defined w.r.t. a binary hypothesis space $\mathbb{H}$ with $\mathbb{Y} = \{0, 1\}$. Its definition is given below.

Definition 2.1.6 ($\mathbb{H}-$divergence)**.** *Given a binary hypothesis space $\mathbb{H}$, the $\mathbb{H}-$divergence between two probability distributions $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ over $\mathbb{X}$ is*

$$d_{\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) = 2 \sup_{h \in \mathbb{H}} \left| \mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{S}_{\mathbb{X}}} [h(\mathbf{x}) = 1] - \mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{T}_{\mathbb{X}}} [h(\mathbf{x}) = 1] \right|. \tag{2.9}$$

We have that $d_{\mathbb{H}} = 2\,\mathrm{IPM}_{\mathbb{H}}$ because

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{D}_{\mathbb{X}}} [h(\mathbf{x}) = 1] = \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbb{X}}} [[h(\mathbf{x}) = 1]] = \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbb{X}}} [h(\mathbf{x})]\,,$$

for any distribution $\mathcal{D}_{\mathbb{X}}$ over $\mathbb{X}$, where the first equality follows from the definition of a probability and the second is due to the fact that $\mathbb{H}$'s hypotheses take binary values. This quantity is estimable from finite samples as long as $\mathbb{H}$ has a finite VC dimension (Definition 1.2.4), as shown by the following theorem.

Theorem 2.1.1. *(Ben-David et al., 2010, Lemma 1) Let $\mathbb{H}$ be a binary hypothesis space with a finite VC dimension and let $\delta \in (0, 1)$. Then with a probability at least $1 - \delta$ over the draw of $S_u \sim {\mathcal{S}_{\mathbb{X}}}^m$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^m$, we have:*

$$\left| d_{\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) - d_{\mathbb{H}}(\hat{S}_u, \hat{T}_u) \right| \leq 4 \sqrt{\frac{VC(\mathbb{H}) \log(2m) + \log \frac{2}{\delta}}{m}}. \tag{2.10}$$

Interestingly, the $\mathbb{H}-$divergence has an appealing interpretation: it is the accuracy of the best classifier that tries to distinguish between the source and target instances. This intuition is formalized in the following proposition.

Proposition 2.1.1. *(Ben-David et al., 2010, Lemma 2) Let $\mathbb{H}$ be a symmetric binary hypothesis space,* i.e. $\forall h \in \mathbb{H}, 1 - h \in \mathbb{H}$*, and let $S_u \sim {\mathcal{S}_{\mathbb{X}}}^{m_s}$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^{m_t}$, then:*

$$\frac{1}{2} d(\hat{S}_u, \hat{T}_u) = 1 - \min_{h \in \mathbb{H}} \left(\frac{1}{m_s} \sum_{\mathbf{x}: h(\mathbf{x}) = 0} [\mathbf{x} \in S_u] + \frac{1}{m_s} \sum_{\mathbf{x}: h(\mathbf{x}) = 1} [\mathbf{x} \in T_u] \right). \tag{2.11}$$

The minimum term in the r.h.s. of Equation (2.11) is proportional to the best misclassification rate of a binary classifier from $\mathbb{H}$ that decides whether a given instance comes from S_u or T_u. The more accurate this classifier is, the easier it is to distinguish between the two domains, hence the more dissimilar they are. Conversely, if the best classifier trying to distinguish between the two domains fails, *i.e.* has a performance that is close to random guessing, then the domains are expected to be similar in a certain sense.

2.1.2.2 $l-$discrepancy

While the $d_{\mathbb{H}}$ divergence was one of the first IPMs used in the theoretical analysis of DA, it remains restricted to binary classification only. The $l-$discrepancy, introduced in Mansour et al. (2009b), addresses this issue and is defined w.r.t. an arbitrary loss function that is symmetric and respects the triangle inequality, *i.e.*

$$\forall y, y', y'' \in \mathbb{Y} : l(y, y') = l(y', y) \quad \text{and} \quad l(y, y') \leq l(y, y'') + l(y'', y').$$

Such loss functions include l_{01} loss, as well as loss functions used in regression, such as $l_q : (y, y') \mapsto |y - y'|^q$ for $0 < q \leq 1$. The $l-$discrepancy is then defined in terms of the $l-$disagreement (Definition 2.1.3) as follows.

Definition 2.1.7 ($l-$discrepancy)**.** *Given a loss function l and a hypothesis space $\mathbb{H}$, the $l-$discrepancy between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ is:*

$$\mathrm{disc}_l(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) = \sup_{h,h' \in \mathbb{H}} \left| \mathfrak{E}^l_{\mathcal{S}_{\mathbb{X}}}\left(h, h'\right) - \mathfrak{E}^l_{\mathcal{T}_{\mathbb{X}}}\left(h, h'\right) \right|. \tag{2.12}$$

The $l-$discrepancy is a pseudo-metric that takes into account the learning task at hand via hypothesis space $\mathbb{H}$, a property that it shares with the $\mathbb{H}-$divergence. It is estimable from finite samples as specified by the following theorem.

Theorem 2.1.2. *(Mansour et al., 2009b, Corollary 7) Let $l_q : \mathbb{Y}^2 \to \mathbb{R}$ be a loss function defined by $l_q(y, y') := |y - y'|^q$ for some $q > 0$. Assume there exists $M > 0$ such that for all $h, h \in \mathbb{H}$ and all $\mathbf{x} \in \mathbb{X}$, $l(h(\mathbf{x}), h'(\mathbf{x})) \leq M$. Then, for any $\delta \in (0, 1)$, we have with a probability at least $1 - \delta$ over the draw of two samples $S_u \sim {\mathcal{S}_{\mathbb{X}}}^{m_s}$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^{m_t}$:*

$$\left| \mathrm{disc}_{l_q}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) - \mathrm{disc}_{l_q}(\hat{S}_u, \hat{T}_u) \right| \leq 4q(\mathrm{Rad}_{S_u}(\mathbb{H}) + \mathrm{Rad}_{T_u}(\mathbb{H})) + 3M\sqrt{\log \tfrac{4}{\delta}} \left(\tfrac{1}{\sqrt{m_s}} + \tfrac{1}{\sqrt{m_t}} \right). \tag{2.13}$$

This generalization bound involves the Rademacher complexity which depends on the distributions generating the data, resulting in bounds that are tighter than the ones involving the VC dimension.

2.1.2.3 Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) is an IPM introduced in Gretton et al. (2009a, 2012) that relies on mapping the data to an RKHS, as with the kernel trick in the context of kernel SVMs (Section 1.2.6.1). In fact, for MMD the supremum in the definition of the IPM is taken over a unit ball in an RKHS $\mathbb{V}$ defined by a universal kernel, meaning that $\mathbb{V}$ is dense in the space of continuous functions from $\mathbb{X}$ to $\mathbb{R}$ (Steinwart, 2001; Micchelli et al., 2006). The MMD is formally defined as follows.

Definition 2.1.8 (Maximum Mean Discrepancy)**.** *Given an RKHS $\mathbb{V}$ associated to a universal kernel k, the maximum mean discrepancy* $\mathrm{MMD}_{\mathbb{V}}$ *between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ is an IPM for*

$$\mathbb{F} = \{f \in \mathbb{V}; \|f\|_{\mathbb{V}} \leq 1\}. \tag{2.14}$$

Recalling the reproducing kernel property due to Riesz's representation theorem, we have $f(\mathbf{x}) = \langle f, k_{\mathbf{x}} \rangle_{\mathbb{V}}$, where $k_{\mathbf{x}}$ is the image of $\mathbf{x}$ in $\mathbb{V}$ by some mapping. This allows to "separate" f from $\mathbf{x}$ in the IPM's definition, and to use the autoduality[5] of the RKHS norm in order to express the MMD solely in terms of the kernel function k's values, as stated by the following proposition.

Proposition 2.1.2. *(Gretton et al., 2012, Lemma 6) Given an RKHS $\mathbb{V}$ induced by a universal kernel k, the squared MMD between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ verifies:*

$$\mathrm{MMD}^2_{\mathbb{V}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) = \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}' \sim \mathcal{S}_{\mathbb{X}}} \left[k(\mathbf{x}, \mathbf{x}')\right] + \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}' \sim \mathcal{T}_{\mathbb{X}}} \left[k(\mathbf{x}, \mathbf{x}')\right] - 2 \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim \mathcal{S}_{\mathbb{X}} \\ \mathbf{x}' \sim \mathcal{T}_{\mathbb{X}}}} \left[k(\mathbf{x}, \mathbf{x}')\right]. \tag{2.15}$$

This latter form is user friendly as it yields the supremum defining the MMD in a closed form, and allows for its efficient empirical estimation enjoying the following guarantee.

Theorem 2.1.3. *(Gretton et al., 2012, Theorem 7, reformulated [6]) Let $\mathbb{V}$ be an RKHS associated to a universal kerel k such that $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{X}, k(\mathbf{x}, \mathbf{x}') \leq K$ for some $K > 0$.*

[5] $\forall f \in \mathbb{V}, \quad \sup_{\|g\|_{\mathbb{V}} \leq 1} \langle f, g \rangle_{\mathbb{V}} = \|f\|_{\mathbb{V}}$.

[6] We reformulated the last theorem's statement so that it matches the common form used later in the generalization bounds (with $1 - \delta$ characterizing the confidence of the bound).

Then, for any $\delta \in (0,1)$, with a probability at least $1-\delta$ over the draw of $S_u \sim {\mathcal{S}_{\mathbb{X}}}^{m_s}$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^{m_t}$, we have:

$$\left|\mathrm{MMD}_{\mathbb{V}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) - \mathrm{MMD}_{\mathbb{V}}(\hat{S}_u, \hat{T}_u)\right| \leq \sqrt{K}\left(\frac{2}{\sqrt{m_s}} + \frac{2}{\sqrt{m_t}} + \sqrt{2\log\frac{2}{\delta}\left(\frac{1}{m_s}+\frac{1}{m_t}\right)}\right). \tag{2.16}$$

Note that several empirical estimators for the MMD can be used in practice, such as the unbiased or linear time MMD estimator, as explained in Gretton et al. (2012). The MMD is a popular choice in domain adaptation algorithms as we will see in Section 2.2.

2.1.2.4 Wasserstein Distance

The Wasserstein distance between probability distributions is tightly related to the optimal transport problem (Monge, 1781; Kantorovich, 1942) widely studied in the operations research field. It is an IPM over the space of functions that have the Lipschitz property w.r.t. a given metric $d : \mathbb{X}^2 \to \mathbb{R}_+$ (Definition A.1.1) and is formally defined as follows.

Definition 2.1.9. *Given a metric $d : \mathbb{X}^2 \to \mathbb{R}_+$, the Wasserstein distance W_1 is the IPM over the space $\mathbb{F}$ of functions verifying the $1-$Lipchitz property,* i.e.

$$\mathbb{F} = \{f : \mathbb{X} \to \mathbb{R};\ \forall \mathbf{x}, \mathbf{x}' \in \mathbb{X}, \left|f(\mathbf{x}) - f(\mathbf{x}')\right| \leq d(\mathbf{x}, \mathbf{x}')\}. \tag{2.17}$$

The Wasserstein distance naturally captures the geometry of the underlying feature space as induced by its metric d. It can be estimated empirically from finite data as justified by the following theorem.

Theorem 2.1.4. *Given a metric $d : \mathbb{X}^2 \to \mathbb{R}_+$, there exists $c' > 0$ and $m_0 \in \mathbb{N}^*$ such that for all $m_s, m_t \geq m_0$, and for all $\delta \in (0,1)$, with a probability $1-\delta$ over the draw of $S_u \sim {\mathcal{S}_{\mathbb{X}}}^{m_s}$ and $T \sim \mathcal{T}^{m_t}$ we have:*

$$\left|W_1(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) - W_1(\hat{S}_u, \hat{T}_u)\right| \leq \sqrt{\frac{2}{c'}\log\frac{2}{\delta}}\left(\frac{1}{\sqrt{m_s}} + \frac{1}{\sqrt{m_t}}\right). \tag{2.18}$$

This theorem is an intermediate result used in the proof of Courty et al. (2017, Theorem 3.1), and is based on (Bolley et al., 2007, Theorem 1.1). A fundamental difference in this result compared to the previous IPMs is that it holds starting from a threshold $m_0 \in \mathbb{N}^*$.

So far, the expression of the Wasserstein distance that we presented above is given by its dual form, whereas its primal form is a formulation of the optimal transport problem as stated in Kantorovich (1942). In order to present this result, we first need to introduce the set of transport plans between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$, which is the set of probability distributions over $\mathbb{X}^2$ having marginals $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ as stated in the following definition.

Definition 2.1.10 (Set of transport plans)**.** *Let $\mathfrak{P}(\mathbb{X}^2)$ be the set of probability distributions over $\mathbb{X}^2$, and let $\pi_1 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_1$ and $\pi_2 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_2$. The set of transport plans between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$, denoted $\Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$ is:*

$$\Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) := \{\mathcal{P} \in \mathfrak{P}(\mathbb{X}^2); \pi_1 \# \mathcal{P} = \mathcal{S}_{\mathbb{X}} \text{ and } \pi_2 \# \mathcal{P} = \mathcal{T}_{\mathbb{X}}\}. \tag{2.19}$$

In particular, for empirical distributions $\hat{S}_u$ and $\hat{T}_u$ associated to two samples $S_u \sim {\mathcal{S}_{\mathbb{X}}}^{m_s}$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^{m_t}$, $\Pi(\hat{S}, \hat{T})$ is called the set of transport matrices defined as:

$$\Pi(\hat{S}_u, \hat{T}_u) := \left\{\mathbf{P} \in \mathbb{R}_+^{m_s \times m_t}; \mathbf{P}\mathbf{1}_{m_t} = \frac{\mathbf{1}_{m_s}}{m_s}; \mathbf{P}^T\mathbf{1}_{m_s} = \frac{\mathbf{1}_{m_t}}{m_t}\right\}. \tag{2.20}$$

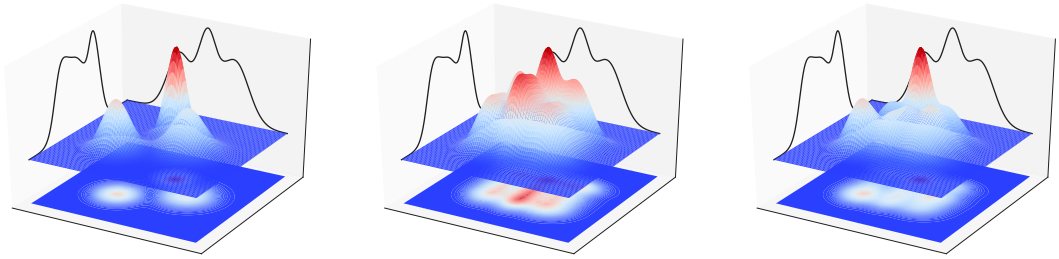


Figure 2.1: Illustration of the set of transport plans. The surface represents the density of a mixture of 2D Gaussians, and the marginal distributions, which are the same for the 3 figures, are represented by the two black curves. These latter were normalized for the sake of illustration.

Intuitively, the elements from the set $\Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$, as illustrated in Figure 2.1, represent all possible probability mass transports between instances drawn from $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$.
The next result gives the primal form of the Wasserstein distance, expressed as an infimum of the transport cost over the set of transport plans. The equality between the primal and dual forms is known as the Kantorovich-Rubinstein theorem and can be found in (Dudley, 2002, Theorem 11.8.2). We recall this result in the following theorem.

Theorem 2.1.5 (Kantorovich-Rubinstein duality)**.** *The primal form of the the Wasserstein distance W_1 is*

$$W_1(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) = \inf_{\mathcal{P} \in \Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})} \mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} [d(\mathbf{x}_s, \mathbf{x}_t)] \,. \tag{2.21}$$

For empirical distributions associated to two samples $S_u \sim {\mathcal{S}_{\mathbb{X}}}^{m_s}$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^{m_t}$, we have:

$$W_1(\hat{S}_u, \hat{T}_u) = \inf_{\mathbf{P} \in \Pi(\hat{S}_u, \hat{T}_u)} \sum_{i=1}^{m_s} \sum_{j=1}^{m_t} (\mathbf{P})_{ij} d(\mathbf{x}_{s,i}, \mathbf{x}_{t,j}) = \inf_{\mathbf{P} \in \Pi(\hat{S}_u, \hat{T}_u)} \langle \mathbf{P}, \mathbf{D} \rangle \,, \tag{2.22}$$

where $(\mathbf{D})_{ij} := d(\mathbf{x}_{s,i}, \mathbf{x}_{t,j})$

In addition to being a metric, the distance $d(\mathbf{x}_s, \mathbf{x}_t)$ between $\mathbf{x}_s \sim \mathcal{S}_{\mathbb{X}}$ and $\mathbf{x}_t \sim \mathcal{T}_{\mathbb{X}}$ can be seen as the cost of transporting one unit of probability mass from $\mathbf{x}_s$ to $\mathbf{x}_t$, whereas the probability density $p_{\mathcal{P}}(\mathbf{x}_s, \mathbf{x}_t)$ is the proportion of the mass of $\mathbf{x}_s$ that is transported to $\mathbf{x}_t$. The same holds in the discrete case concerning probability $(\mathbf{P})_{ij}$ and instances $\mathbf{x}_{s,i}$ and $\mathbf{x}_{t,j}$. Hence, under the primal form, the optimal transport terminology takes all its meaning: we are planning to transport the probability mass from $\mathcal{S}_{\mathbb{X}}$ to $\mathcal{T}_{\mathbb{X}}$ in a way that minimizes the overall transport cost $\mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} [d(\mathbf{x}_s, \mathbf{x}_t)]$.

Remark We presented the optimal transport problem as the primal form of the Wasserstein distance, with a cost function that is a metric. This is a restriction that we took for the sake of coherence with the current section on IPMs. In its most general form, the optimal transport problem is defined w.r.t a cost function $c : \mathbb{X}^2 \to \mathbb{R}_+$ that is only required to be lower semi-continuous (Santambrogio, 2015, Theorem 1.7) for the optimal transport plan to exist. In this case the dual form no longer has the form of an IPM, and can be found in Santambrogio (2015, Section 1.6).

2.1.2.5 Other Dissimilarity Measures between Distributions

Divergence measures other than IPMs were considered in several lines of work in order to relate the two domains. We review some of them below.

Weight ratio divergences This family of divergences include the Kullback-Leibler divergence (Definition A.2.8) used in (Sugiyama and Müller, 2005; Sugiyama et al., 2007) and the Rényi divergence (Definition A.2.9) considered in Cortes et al. (2010). In general, they require one probability distribution to be absolutely continuous w.r.t. the other in order for the divergence value to be finite. As mentioned before, they were employed in importance weighting approaches for DA that we will cover in more detail in Section 2.2.1.1.

$\mathcal{P}-$disagreement has been studied in the PAC-Bayesian setting (Section 1.2.5.3) in (Germain et al., 2013, 2016b,a). In their analysis, the authors introduced the $\mathcal{P}-$disagreement measure defined as follows.

Definition 2.1.11. *Given a probability distribution $\mathcal{P}$ over $\mathbb{H}$, the $\mathcal{P}-$disagreement* $\mathrm{dis}_{\mathcal{P}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$ *between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ is*

$$\mathrm{dis}_{\mathcal{P}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) := \left| \mathop{\mathbb{E}}_{h,h' \sim \mathcal{P}} \left[\mathfrak{E}^{01}_{\mathcal{T}_{\mathbb{X}}}\left(h, h'\right) - \mathfrak{E}^{01}_{\mathcal{S}_{\mathbb{X}}}\left(h, h'\right) \right] \right|. \tag{2.23}$$

This quantity is similar to an IPM but with an expectation over hypotheses in $\mathbb{H}$ instead of a supremum. When presenting the DA learning bounds, we will show that it can be less strict than the $\mathbb{H}-$divergence (Ben-David et al., 2010). The $\mathcal{P}-$disagreement is further estimable from finite samples as shown in (Germain et al., 2013, Theorem 3).

Margin Disparity Discrepancy (MDD) Recently introduced in Zhang et al. (2019) in the context of multi-class classification with $\mathbb{Y} = \{1, \ldots, K\}$, the MDD is formulated in terms of the ramp loss (Table 1.2) and for which we recall the expression below:

$$\ell_\beta(t) := \min\left(1, \left(1 - \frac{t}{\beta}\right)_+\right).$$

Given a scoring function $\vec{h} : \mathbb{X} \to \mathbb{R}^K$ (Section 1.2.3), function ℓ_β allows to define the following loss function:

$$l_\beta(h(\mathbf{x}), y) = \frac{1}{2}\ell_\beta\left(h_y(\mathbf{x}) - \max_{y' \neq y} h_{y'}(\mathbf{x})\right), \tag{2.24}$$

where the argument of ℓ_β is the generalization of the classification margin for the multiclass case, introduced in Koltchinskii and Panchenko (2002). With these definitions and by letting $y(\vec{h})(\mathbf{x}) := \arg\max_{1 \leq y \leq K} h_y(\mathbf{x})$, we can now give the MDD's definition below.

Definition 2.1.12 (Margin Disparity Discrepancy)**.** *Given a hypothesis space $\mathbb{H}$ of scoring functions and $\vec{h} \in \mathbb{H}$, the Margin Disparity Discrepancy is defined by:*

$$d^{(\beta)}_{h,\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) := \sup_{\vec{h}' \in \mathbb{H}} \left(\mathfrak{E}^{(\beta)}_{\mathcal{T}_{\mathbb{X}}}\left(\vec{h}', y(\vec{h})\right) - \mathfrak{E}^{(\beta)}_{\mathcal{S}_{\mathbb{X}}}\left(\vec{h}', y(\vec{h})\right) \right). \tag{2.25}$$

As it is the case with the $\mathbb{H}-$divergence and the $l-$discrepancy, the MDD is defined as a supremum over the hypothesis space at hand. However, this supremum is taken over one hypothesis instead of two, thus making the MDD dependent on $\vec{h}$ and tighter than the $\mathbb{H}-$divergence for $\beta = 0$, corresponding to the misclassication loss. For binary classification with labels encoded as $\{-1, 1\}$, its expression reduces [7] to:

$$d^{(\beta)}_{h,\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) := \sup_{h' \in \mathbb{H}} \left(\mathfrak{E}^{(\beta)}_{\mathcal{T}_{\mathbb{X}}}\left(h', \mathrm{sgn}\,(h)\right) - \mathfrak{E}^{(\beta)}_{\mathcal{S}_{\mathbb{X}}}\left(h', \mathrm{sgn}\,(h)\right) \right). \tag{2.26}$$

The authors show in Zhang et al. (2019, Lemma 3.6) that the MDD is estimable from finite samples with guarantees expressed in terms of the Rademacher complexity (Definition 1.2.5) and the margin parameter β.

[7] We provide a proof for the bound of Zhang et al. (2019) when $\mathbb{Y} = \{-1, 1\}$ in Section A.3.

2.1.3 Sufficiency: Bounding the Target Risk

Since the labels are not available for the sample $T_u \sim \mathcal{T}_{\mathbb{X}}$, studying the performance of a model on this latter must be done in terms of the available quantities, which are the unlabeled target data and the labeled source data. Several works in the literature provide bounds on a risk in the target domain, having the following generic form for a given hypothesis $h \in \mathbb{H}$:

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{S}}(h) + \text{divergence}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + t(\mathcal{S}, \mathcal{T}). \tag{2.27}$$

Hence, in order to have a good performance on the target domain, it is sufficient for all of the three terms, *i.e.* the source domain $l-$risk, the divergence between the marginals and the non-estimable term, to be small. We now discuss the role of each of these three terms.

Source domain l-risk is estimable from finite samples and can be minimized by learning h from the available source labeled data to reflect in the best way possible the relation between inputs $\mathbb{X}$ and outputs $\mathbb{Y}$.

Divergence between the marginals is a measure of the dissimilarity between the distributions given generally by an IPM (Section 2.1.2). It is estimable from the observed data and is expected to be small if the two domains are similar (Section 2.1.1.1).

Non-estimable term requires having access to the labels in the target domain and thus, contrary to two previous terms, cannot be assessed in practice. In general, this term is assumed to be small for a successful adaptation to be possible. This latter assumption boils down to assuming that the conditional probabilities $\mathcal{T}_{\mathbb{Y}|\mathbf{x}}$ and $\mathcal{S}_{\mathbb{Y}|\mathbf{x}}$ or the labeling functions for the two domains are somehow related (Section 2.1.1.2).

In what follows, we present several DA bounds having the general form presented above and being expressed in terms of quantities defined in Sections 2.1.1 and 2.1.2.

2.1.3.1 Based on the $\mathbb{H}\Delta\mathbb{H}-$divergence

With the binary classification setting of Section 2.1.2.1, the authors of Ben-David et al. (2010) introduce the *symmetric difference hypothesis space* $\mathbb{H}\Delta\mathbb{H}$ defined as follows.

Definition 2.1.13 (Symmetric difference hypothesis space)**.** *Given a set $\mathbb{H}$ of binary hypotheses taking their values in $\{0,1\}$, the symmetric difference hypothesis space $\mathbb{H}\Delta\mathbb{H}$ is*

$$\mathbb{H}\Delta\mathbb{H} := \{h \oplus h'; h, h' \in \mathbb{H}\} = \{\left|h - h'\right|; h, h' \in \mathbb{H}\} = \{[h \neq h']; h, h' \in \mathbb{H}\}. \tag{2.28}$$

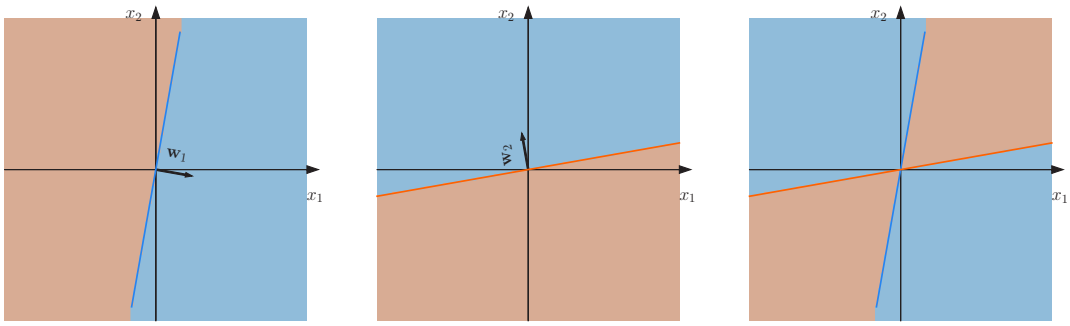
Put differently, a hypothesis g belongs to $\mathbb{H}\Delta\mathbb{H}$ if and only if (iff) it is written as a disagreement between two hypotheses h and h' from $\mathbb{H}$ (Figure 2.2). With the $\mathbb{H}\Delta\mathbb{H}-$divergence (Definition 2.1.6), Ben-David et al. (2007) proved a bound on the target misclassification rate given in the following theorem.

Theorem 2.1.6. *(Ben-David et al., 2007, Theorem 2) Given a binary hypothesis space $\mathbb{H}$ with values in $\mathbb{Y} = \{0, 1\}$, we have for any $h \in \mathbb{H}$:*

$$\mathfrak{E}^{01}_{\mathcal{T}}(h) \leq \mathfrak{E}^{01}_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathbb{H}\Delta\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda_{\mathbb{H}}(\mathcal{S}, \mathcal{T}), \tag{2.29}$$

where

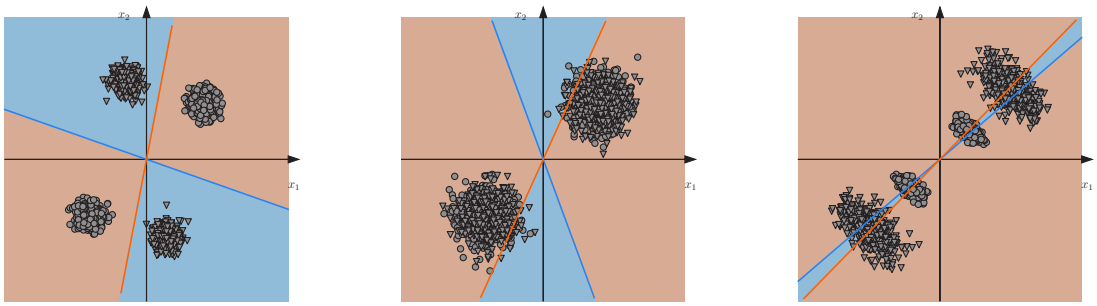
$$\lambda_{\mathbb{H}}(\mathcal{S}, \mathcal{T}) = \min_{h \in \mathbb{H}} \mathfrak{E}^{01}_{\mathcal{S}}(h) + \mathfrak{E}^{01}_{\mathcal{T}}(h). \tag{2.30}$$



(a) $h_1(\mathbf{x}) = [\mathbf{w}_1^T\mathbf{x} > 0]$. (b) $h_2(\mathbf{x}) = [\mathbf{w}_2^T\mathbf{x} > 0]$. (c) $(h_1 \oplus h_2)(\mathbf{x})$.

Figure 2.2: Illustration of the symmetric difference between two classifiers. The blue and brown colors respectively indicate the positive and negative classes. Classifier $h_1 \oplus h_2$ indicates where h_1 and h_2 disagree.

This bound shows that a good performance on the source domain, similar marginals in terms of the $\mathbb{H}\Delta\mathbb{H}-$divergence illustrated in Figure 2.3) and the existence of a low error ideal hypothesis (Section 2.1.1) are sufficient for successful adaptation. Moreover, we have $VC(\mathbb{H}\Delta\mathbb{H}) \leq 2VC(\mathbb{H})$ (Ben-David et al., 2010), hence $d_{\mathbb{H}\Delta\mathbb{H}}$ is estimable from finite samples as long as $\mathbb{H}$ has a finite VC dimension (Theorem 2.1.1). As for the last term, the ideal joint hypothesis error is expected to be small. This is the case, for example, when the labeling functions of the two domains are similar.



(a) $\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}(\hat{S}_u, \hat{T}_u) = 1$: $\hat{S}_u$ and $\hat{T}_u$ are separable by $h \in \mathbb{H}\Delta\mathbb{H}$.

(b) $\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}(\hat{S}_u, \hat{T}_u) = 0.07$, for similar domains.

(c) $\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}(\hat{S}_u, \hat{T}_u) = 0.07$, even when the domains are different.

Figure 2.3: Illustration of the $\mathbb{H}\Delta\mathbb{H}-$divergence with $\mathbb{H}$ the space of linear classifiers, where the source and target instances are represented by circles and triangles. We numerically computed the empirical $\mathbb{H}\Delta\mathbb{H}-$divergence for each case. For Figure 2.3c, the target domain is constructed from the source via a homothety.

2.1.3.2 Based on the $l-$discrepancy

Mansour et al. (2009b) provided a bound based on $l-$discrepancy that enjoys the properties discussed above and generalizes, in some sense, the seminal result provided in Theorem 2.1.6. This bound is given in the following theorem.

Theorem 2.1.7. *(Mansour et al., 2009b, Theorem 8) Given a hypothesis space $\mathbb{H}$ and a symmetric loss function l verifying the triangle inequality, we have for any $h \in \mathbb{H}$:*

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{S}}(h, h_{\mathcal{S}}) + \mathrm{disc}_l(\mathcal{S}, \mathcal{T}) + \mathfrak{E}^l_{\mathcal{S}}(h_{\mathcal{S}}, h_{\mathcal{T}}) + \mathfrak{E}^l_{\mathcal{T}}(h_{\mathcal{T}}), \tag{2.31}$$

where

$$h_{\mathcal{S}} \in \arg\min_{h\in\mathbb{H}} \mathfrak{E}^l_{\mathcal{S}}(h), \quad h_{\mathcal{T}} \in \arg\min_{h\in\mathbb{H}} \mathfrak{E}^l_{\mathcal{T}}(h). \tag{2.32}$$

We further note that:

$$\mathrm{disc}_{l_{01}}(\mathcal{S},\mathcal{T}) = \frac{1}{2} d_{\mathbb{H}\Delta\mathbb{H}}(\mathcal{S},\mathcal{T}), \tag{2.33}$$

hence the $l-$discrepancy generalizes the $\mathbb{H}\Delta\mathbb{H}-$discrepancy and concerns a broader scope of learning tasks including regression. However, this bound has two notable distinctions. First, the source related term is not the risk of h but rather its disagreement with the best hypothesis in $\mathbb{H}$. Second, the non-estimable term is a sum of the $l-$disagreement (Definition 2.1.3) between $h_{\mathcal{S}}$ and $h_{\mathcal{T}}$ and the $l-$risk of $h_{\mathcal{T}}$ on the target domain. This latter must of course be small (otherwise, does it even make sense to look for a hypothesis in $\mathbb{H}$?) and, as discussed before, requires the labeling functions to be somehow similar.

2.1.3.3 Based on the Maximum Mean Discrepancy

The MMD (Section 2.1.2.3) has been actively used to derive DA algorithms (Huang et al., 2007; Pan et al., 2008, 2011; Gong et al., 2013) before its first theoretical study in the context of DA provided in Redko (2015). This latter considers a hypothesis space $\mathbb{H}$ that is an RKHS and a loss function $l : (y, y') \mapsto |y - y'|^q$ for some $q > 0$ such that the function $\ell^{h,f} : \mathbf{x} \mapsto l(h(\mathbf{x}), f(\mathbf{x}))$ is convex for any $h, f \in \mathbb{H}$. This latter then belongs to an RKHS $\mathbb{H}_q$ according to Saitoh (1997). The authors then derive a bound that is analogous to the one in Theorem 2.1.6 in the following theorem.

Theorem 2.1.8. *(Redko, 2015, Theorem 6.10, reformulated) Let $\mathbb{H}$ be an RKHS. Let l the loss function defined by $l_q(y, y') = |y - y'|^q$ for some $q > 0$ assumed to verify the triangle inequality, such that $\ell^{h,h'} : \mathbf{x} \mapsto l(h(\mathbf{x}), h'(\mathbf{x}))$ is convex for all $h, h' \in \mathbb{H}$. Let $\mathbb{H}_q$ be the RKHS such that $\ell^{h,f} \in \mathbb{H}_q$. Then for all $h \in \mathbb{H}$:*

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{S}}(h) + \mathrm{MMD}_{\mathbb{H}_q}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda(\mathcal{S},\mathcal{T}), \tag{2.34}$$

where

$$\lambda_{\mathbb{H}}(\mathcal{S},\mathcal{T}) = \min_{h\in\mathbb{H}} \mathfrak{E}^l_{\mathcal{S}}(h) + \mathfrak{E}^l_{\mathcal{T}}(h). \tag{2.35}$$

This result shows that a low MMD between domains' marginals is one of the sufficient conditions for the success of adaptation, in addition to a low joint error. Also, as mentioned in Section 2.1.2.3, the MMD benefits from the closed form of Proposition 2.1.2 and from the estimation guarantees (Theorem 2.1.3) that do not depend on the dimensionality of the feature space, contrary to the analoguous bound of Theorem 2.1.6.

2.1.3.4 Based on the Wasserstein Distance

Compared to the MMD, the Wasserstein distance has become common in DA only recently (Courty et al., 2016, 2017; Shen et al., 2018). To theoretically analyze the DA problem with the Wasserstein distances, Redko et al. (2017) provided a bound in the spirit of Theorem 2.1.8 that we present in the theorem below.

Theorem 2.1.9. *(Redko et al., 2017, Theorem 2, modified) With the assumptions of Theorem 2.1.8, let d be the metric defined as $d : (\mathbf{x}, \mathbf{x}') \mapsto \sqrt{k(\mathbf{x},\mathbf{x}) + k(\mathbf{x}',\mathbf{x}') - 2k(\mathbf{x},\mathbf{x}')}$ where k is the kernel function associated to $\mathbb{H}_q$, verifying $0 \leq k(\mathbf{x},\mathbf{x}') \leq K$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$, and let W_1 be the Wasserstein distance induced by d. Then for any $h \in \mathbb{H}$ we have:*

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{D}}(h) + W_1(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda(\mathcal{S},\mathcal{T}), \tag{2.36}$$

where

$$\lambda_{\mathbb{H}}(\mathcal{S},\mathcal{T}) = \min_{h\in\mathbb{H}} \mathfrak{E}^l_{\mathcal{S}}(h) + \mathfrak{E}^l_{\mathcal{T}}(h). \tag{2.37}$$

The previous bound concerns an RKHS space, which is considered in several learning algorithms, and is the first theoretical justification for the use of the Wasserstein metric for DA. Another bound in terms of that metric for the absolute value loss was given in Shen et al. (2018).

Theorem 2.1.10. *(Shen et al., 2018, Theorem 1) Let l the loss function defined by $l(y, y') = |y - y'|$, and let $\mathbb{H}$ be hypothesis space of functions verifying the L-Lipschitz continuity for some $L > 0$ with respect to some metric d over $\mathbb{X}$. With this latter inducing the Wasserstein metric W_1, we have for any $h \in \mathbb{H}$:*

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{S}}(h) + 2LW_1(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda(\mathcal{S}, \mathcal{T}), \tag{2.38}$$

where

$$\lambda(\mathcal{S}, \mathcal{T}) = \min_{h \in \mathbb{H}} \mathfrak{E}^l_{\mathcal{S}}(h) + \mathfrak{E}^l_{\mathcal{T}}(h). \tag{2.39}$$

This bound relies on the Lipschitz property which holds for several choices of $\mathbb{H}$ as a space of scoring functions. However, it is restricted to one particular loss function, whereas the bound from Theorem 2.1.9 holds for a family of loss functions indexed by a parameter $q > 0$ and concerns RKHS hypothesis spaces. Both theorems justify the sufficiency of a small Wasserstein distance along with a low ideal joint error for a successful adaptation.

2.1.3.5 Other Bounds

Below we cover some other bounds that do not verify the learning setting we considered so far, or that do not have the generic bound form from Equation (2.27). These include the following settings.

PAC-Bayes DA was addressed in the PAC-Bayesian framework in (Germain et al., 2013, 2016a), where a prior $\mathcal{P}_0$ over a space of binary hypotheses $\mathbb{H}$ was considered. In the first paper, the following result using the $\mathcal{P}-$disagreement (Definition 2.1.11) was proven.

Theorem 2.1.11. *(Germain et al., 2016b, Theorem 9) For any posterior distribution $\mathcal{P}$ over $\mathbb{H}$, we have*

$$\mathop{\mathbb{E}}_{h \sim \mathcal{P}} \left[\mathfrak{E}^{01}_{\mathcal{T}}(h)\right] \leq \mathop{\mathbb{E}}_{h \sim \mathcal{P}} \left[\mathfrak{E}^{01}_{\mathcal{S}}(h)\right] + \frac{1}{2} \mathrm{dis}_{\mathcal{P}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda_{\mathcal{P}}(\mathcal{S}, \mathcal{T}), \tag{2.40}$$

where

$$\lambda_{\mathcal{P}}(\mathcal{S}, \mathcal{T}) := \left| \mathop{\mathbb{E}}_{\substack{(\mathbf{x},y) \sim \mathcal{T} \\ h, h' \sim \mathcal{P}}} [[h(\mathbf{x}) \neq y][h'(\mathbf{x}) \neq y]] - \mathop{\mathbb{E}}_{\substack{(\mathbf{x},y) \sim \mathcal{S} \\ h, h' \sim \mathcal{P}}} [[h(\mathbf{x}) \neq y][h'(\mathbf{x}) \neq y]] \right|. \tag{2.41}$$

This result looks analogous to Equation (2.27) with an important distinction that consists in replacing the supremum in the domain dissimilarity term with an expectation. In this case, the authors of Germain et al. (2013) proved that $\frac{1}{2} d_{\mathbb{H}\Delta\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) \geq \mathrm{dis}_{\mathcal{P}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$, making their disagreement term less strict when used as a divergence. Also, we note that the posterior distribution $\mathcal{P}$ intervenes even in the non-estimable term $\lambda_{\mathcal{P}}(\mathcal{S}, \mathcal{T})$, which is not the case for all of the previously presented bounds.

While the previous bound has an additive dependence between the domain divergence term and a supervised source term (the first two terms of the r.h.s.), the bound proved Germain et al. (2016a) is drastically different and introduces a multiplicative dependence between them as follows.

Theorem 2.1.12. *(Germain et al., 2016a, Theorem 3) For any $q > 0$ and any probability $\mathcal{P}$ over $\mathbb{H}$, we have:*

$$\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\mathfrak{E}^{01}_{\mathcal{T}}(h)\right] \leq \frac{1}{2}\mathop{\mathbb{E}}_{h,h'\sim\mathcal{P}}\left[\mathfrak{E}^{01}_{\mathcal{T}}\left(h,h'\right)\right] + \beta_q(\mathcal{T}\|\mathcal{S})\mathop{\mathbb{E}}_{\substack{h,h'\sim\mathcal{P}\\(\mathbf{x},y)\sim\mathcal{S}}}\left[l_{01}(h(\mathbf{x}),y).l_{01}(h'(\mathbf{x}),y)\right]^{1-\frac{1}{q}} + \eta_{\mathcal{T}\setminus\mathcal{S}}, \tag{2.42}$$

where

$$\beta_q(\mathcal{T}\|\mathcal{S}) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mathcal{S}}(\mathbf{x},y)\right)^q\right]^{\frac{1}{q}},$$
$$\eta_{\mathcal{T}\setminus\mathcal{S}} := \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{T}}[(\mathbf{x},y)\notin\operatorname{supp}\mathcal{S}]\sup_{h\in\mathbb{H}}\mathfrak{E}^{01}_{\mathcal{T}}(h).$$

The first distinction of this bound is that the domain dissimilarity term $\beta_q(\mathcal{S},\mathcal{T})$ multiplies a source domain term, as opposed to the additive dependence in the previously presented results. The source term is not a classic risk but rather what the authors call the *expected joint error*. We note also that the first term is only related to the target domain, is unsupervised and is tightly linked to the cluster assumption (Section 2.1.1). Finally, the last term $\eta_{\mathcal{T}\setminus\mathcal{S}}$ is non-estimable, and is the price to pay for considering the quantity $\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mathcal{S}}$ as it is properly defined only for regions of $\mathbb{X}\times\mathbb{Y}$ verifying $\mathcal{S}\gg\mathcal{T}$.

Wasserstein distance between joint probabilities In this contribution of (Courty et al., 2017), the authors consider a pseudo-labeled version of the target domain, denoted $\mathcal{T}^h$ for $h\in\mathbb{H}$. Their bound involves transport plans (Definition 2.1.10) defining the Wasserstein distance (Theorem 2.1.5), and is given below.

Theorem 2.1.13. *(Courty et al., 2017, Theorem 3.1) Let l be an $L-$Lipschitz loss function for some $L>0$, assumed to verify the triangle inequality. Let $h\in\mathbb{H}$ and let $\mathcal{T}^h$ be the target domain distribution with labels predicted by h. For $\alpha>0$, let $d_{\alpha,l}$ be the cost function defined over $(\mathbb{X}\times\mathbb{Y})^2$ as*

$$d_{\alpha,l}((\mathbf{x},y),(\mathbf{x}',y')) := \alpha d(\mathbf{x},\mathbf{x}') + l(y,y'),$$

and $W_{\alpha,l}$ its induced Wasserstein distance between $\mathcal{S}$ and $\mathcal{T}^h$. Let $\mathcal{P}^*\in\Pi(\mathcal{S},\mathcal{T}^h)$ be a transport plan defining $W_{\alpha,l}(\mathcal{S},\mathcal{T}^h)$. Assume there exists a Lipschitz continuous function $f^*\in\mathbb{H}$ such that

$$f^*\in\mathop{\arg\min}_{f\in\mathbb{H}\cap\mathrm{PTL}_\phi(\mathcal{P}^*)}\mathfrak{E}^l_{\mathcal{S}}(f)+\mathfrak{E}^l_{\mathcal{T}}(f), \tag{2.43}$$

for some $\phi:\mathbb{R}\to[0,1]$. Also, assume that $|f^*(\mathbf{x})-f^*(\mathbf{x}')|\leq M,\ \forall\mathbf{x},\mathbf{x}'\in\mathbb{X}$ for some $M>0$. Then,

$$\mathfrak{E}^l_{\mathcal{T}}(h)\leq W_{\alpha,l}(\mathcal{S},\mathcal{T}^h)+\mathfrak{E}^l_{\mathcal{S}}(f^*)+\mathfrak{E}^l_{\mathcal{T}}(f^*)+LM\phi\left(\frac{\alpha}{L}\right). \tag{2.44}$$

Although this bound does not have the form of Equation (2.27), it includes the joint error associated to the ideal joint hypothesis f^*. However, this latter is restricted to hypotheses that verify the probabilistic transfer Lipschitzness (Equation (2.8)) w.r.t. the optimal transport plan $\mathcal{P}^*$.

Based on the MDD In Zhang et al. (2019), the Maximum Disparity Discrepancy is used to bound the misclassification rate on the target domain as stated by the following theorem.

Theorem 2.1.14. *(Zhang et al., 2019, Proposition 3.3) Given label space* $\mathbb{Y} = \{1, \cdots, K\}$ *and hypothesis space of scoring functions* $\mathbb{H}$*, we have for any* $\beta > 0$ *and* $h \in \mathbb{H}$*:*

$$\mathfrak{E}^{01}_{\mathcal{T}}(h) \leq \mathfrak{E}^{(\beta)}_{\mathcal{S}}(h) + d^{(\beta)}_{h,\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda^{(\beta)}, \tag{2.45}$$

where

$$\lambda^{(\beta)} := \min_{f \in \mathbb{H}} \mathfrak{E}^{(\beta)}_{\mathcal{S}}(f) + \mathfrak{E}^{(\beta)}_{\mathcal{T}}(f). \tag{2.46}$$

This bound has the general form presented in Equation (2.27), with the particularity of the dependence of the divergence term on the considered hypothesis h (Definition 2.1.12). Aside from the remarks we made about the MDD $d^{(\beta)}_{h,\mathbb{H}}$ in Section 2.1.2.5, we note the dependence of the non-estimable term on the classification margin, highlighted by parameter $\beta > 0$.

2.1.4 Necessity: Difficulty of Adaptation

From the generic form of DA bounds (Equation (2.27)), we saw that ensuring good performance on the source domain and reducing the distance between the marginals, while assuming that the non-estimable term is small, are sufficient conditions for a successful adaptation. However, what about their necessity? This question is at the heart of Ben-David et al. (2010) that proves the necessity to simultaneously have a small $\mathbb{H}\Delta\mathbb{H}-$divergence between domains and a low joint error. To formalize learning in a DA setting, the authors define a domain adaptation learner as a function that takes a labeled source sample, an unlabeled target one, and outputs a classifier from a hypothesis class $\mathbb{H}$. Then, they define learnability for a domain adaptation task as follows.

Definition 2.1.14 ($(\varepsilon, \delta, m, n)$-learnability)**.** *Let* $\mathbb{Y} = \{0, 1\}$*, let* $\mathfrak{A}$ *be a domain adaptation learner. For* $\epsilon, \delta > 0$*, and* $m_s, m_t \in \mathbb{N}^*$*, we say that* $\mathfrak{A}$ (ϵ, δ, m, n)*-learns* $\mathcal{T}$ *from* $\mathcal{S}$ *relative to* $\mathbb{H}$*, if when given access* $S \sim \mathcal{S}^{m_s}$*,* $T_u \sim {\mathcal{T}_{\mathbb{X}}}^{m_t}$*, with a probability at least* $1 - \delta$ *(over the choice of the samples* S *and* T_u*) we have:*

$$\mathfrak{E}^{01}_{\mathcal{T}}(h_{S,T_u}) \leq \min_{h \in \mathbb{H}} \mathfrak{E}^{01}_{\mathcal{T}}(h) + \epsilon,$$

where h_{S,T_u} *is the hypothesis learned by* $\mathfrak{A}$ *from* S *and* T_u*.*

This definition is analogous to $PAC-$learnability (Definition 1.2.1), but without uniformity over the probability distributions of the domains. It is then proven in (Ben-David et al., 2010, Theorems 1 and 2) that under mild assumptions over the feature space $\mathbb{X}$ and the hypothesis space $\mathbb{H}$, it is possible to construct a source domain $\mathcal{S}$ over $\mathbb{X} \times \mathbb{Y}$ and a marginal target distribution $\mathcal{T}_{\mathbb{X}}$ verifying the following property: for any $\epsilon > 0$ and any domain adaptation learner $\mathfrak{A}$, one can label the target domain by a binary function f such that the covariate shift assumption holds, and either the joint error (Ben-David et al., 2010, Theorem1) or the $\mathbb{H}\Delta\mathbb{H}-$divergence (Ben-David et al., 2010, Theorem 2) is at most ϵ. However, in both cases, with a high probability over the draw of samples $S \sim \mathcal{S}^{m_s}$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^{m_t}$, $\mathfrak{A}$ has a high misclassification rate over the target domain. Put differently, without both the low joint error and low $\mathbb{H}\Delta\mathbb{H}-$divergence assumptions holding simultaneously, a domain adaptation learner $\mathfrak{A}$ may fail.

2.2 Algorithmic Advances

Historically, proposing DA algorithms preceded its theoretical study: in fact, one of the earliest surveys on transfer learning, containing a section about several DA algorithms, can be found in Pan and Yang (2010) and includes the approaches of Roark and Bacchiani

(2003); Blitzer et al. (2006); Daumé and Marcu (2006); Sugiyama et al. (2007); Daumé III (2009) to name a few.

In general, the vast majority of existing DA approaches are based on the common idea of aligning the data from two domains to reduce the discrepancy between their distributions. In fact, after such an alignment, the task at hand reduces to a supervised learning problem as there's no more shift between the two considered distributions. Other approaches rely on different hypotheses, such as the cluster assumption (Section 2.1.1). Qualifying the number of DA algorithms that have been proposed in the literature as gargantuan does not do it justice. In fact, such a proliferation of methods is explained by the number of different possibilities of answering the following questions:

How to measure the dissimilarity between domains? In addition to the ones that we addressed in Section 2.1.2, several other criteria have been proposed.

How and on what variables is the alignment performed? Some algorithms employ feature transformations to align the domains, others act at the level of the instances' empirical distributions. These transformations can be linear or non-linear, and the criterion that defines them depends on the dissimilarity measure between the domains.

Are alignment and learning separated? Whether it is better to move the two domains closer and learn the task at hand separately or jointly remains an open question.

Is the algorithm shallow or deep? Although not exclusive to DA, the question of whether one has to rely on deep neural networks for successful adaptation in various tasks (Wilson and Cook, 2019) or whether shallow methods with deep features are equally good remains open.

In the next sections, we provide a non exhaustive overview of different DA algorithms, where the answers to the previous questions are given for each family of approaches covered.

2.2.1 Instance Re-weighting Approaches

This family of approaches relies on two assumptions: the covariate shift and the absolute continuity assumption of the target's marginal w.r.t. the source one, *i.e.* $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$. Defining the *importance*[8] (Tsuboi et al., 2009) as $w(\mathbf{x}) := \frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x})$, this approach can be deduced as follows:[9]

$$
\begin{aligned}
\mathfrak{E}^{l}_{\mathcal{T}}(h) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{T}_{\mathbb{Y}|\mathbf{x}}}\left[l(h(\mathbf{x}),y)\right]\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}\left[\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x})\mathop{\mathbb{E}}_{y\sim\mathcal{T}_{\mathbb{Y}|\mathbf{x}}}\left[l(h(\mathbf{x}),y)\right]\right] \qquad \text{(due to the absolute continuity assumption)} \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}\left[w(\mathbf{x})\mathop{\mathbb{E}}_{y\sim\mathcal{S}_{\mathbb{Y}|\mathbf{x}}}\left[l(h(\mathbf{x}),y)\right]\right] \qquad \text{(due to the covariate shift assumption)} \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{S}_{\mathbb{Y}|\mathbf{x}}}\left[w(\mathbf{x})l(h(\mathbf{x}),y)\right]\right] = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\left[w(\mathbf{x})l(h(\mathbf{x}),y)\right]. \qquad (2.47)
\end{aligned}
$$

Hence, minimizing the $l-$risk on the target domain boils down to minimizing a weighted risk on the source domain, with weights $w(\mathbf{x})$ that are independent of the labels. For a

[8] In practice, this is the ratio of densities or the ratio of probability masses for discrete probabilities.

[9] For the sake of generality, we avoid introducing densities and considering the quantity $\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mathcal{S}}$ requiring that $\mathcal{T} \ll \mathcal{S}$.

sample $S_u \sim \mathcal{S}_{\mathbb{X}}{}^m$, it can be schematized as follows:

$$\hat{S}_u := \frac{1}{m}\sum_{i=1}^{m}\delta_{\mathbf{x}_i} \xrightarrow{\text{Re-weighting}} \begin{cases} \hat{S}_u^{\mathbf{w}} := \frac{1}{m}\sum_{i=1}^{m} w_i \delta_{\mathbf{x}_i}. \\ \hat{S}_u^{\mathbf{p}} := \sum_{i=1}^{m} p_i \delta_{\mathbf{x}_i} \quad \text{where } \mathbf{p} \in \Delta_{m_s}, \end{cases} \tag{2.48}$$

where Δ_{m_s} is the m_s-dimensional probability simplex and the two presented re-weighting schemes are equivalent and linked by the equality $w_i = m \cdot p_i$. The effect of instance re-weighting on the output classifier is illustrated in Figure 2.4. Instance re-weighting is

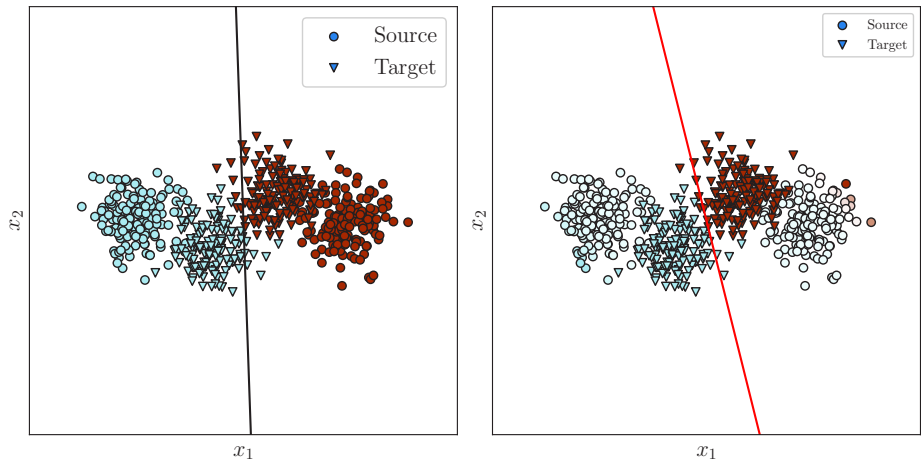


Figure 2.4: Illustration of instance-based approaches: (left) without re-weighting, the classification boundary of the source domain fails to perfectly classify the target data; (right) After re-weighting, the boundary (in red) fits better the target data, where the weights are expressed by the color shades.

one of the earliest approaches (Shimodaira, 2000; Sugiyama et al., 2007) in DA, and the estimation of weights $w(\mathbf{x})$ is the key challenge in these approaches. Several methods, with some of them being briefly described below, have been proposed in the literature to do it indirectly or directly.

2.2.1.1 Indirect Instance Re-weighting Approaches

For this approach, the probability weights of the instances from both domains are estimated, and then the ratio is used to compute w_i. The estimation of every domain's probability weights can be parametric (Shimodaira, 2000) or non parametric, *e.g.* by *kernel density estimation* (Sugiyama et al., 2005; Baktashmotlagh et al., 2014).

2.2.1.2 Direct Instance Re-weighting Approaches

Rather than computing the densities for each domain to deduce the importance ratio, direct re-weighting approaches (Tsuboi et al., 2009; Kanamori et al., 2009) aim at computing the ratio by seeking a new probability distribution on the source domain that matches the target one, *i.e.* directly estimating $\mathbf{p}$ from Equation (2.48). This is generally carried out via minimizing a divergence between the empirical marginal distributions of covariates with one of the most popular choices for this divergence being the MMD distance (Section 2.1.2.3) for which the alignment is known as *Kernel Mean Matching* (KMM) (Huang et al., 2007; Gretton et al., 2009b; Chu et al., 2013; Gong et al., 2013; Yan et al., 2017). Other choices include the $l-$discrepancy for which a re-weighting approach was proposed for a regression task in Mansour et al. (2009b). The solved optimization problem has one of the two following equivalent forms (due the equivalence presented in Equation (2.48)) for two unlabeled samples $S_u \sim \mathcal{S}_{\mathbb{X}}{}^{m_s}$ and $T_u \sim \mathcal{T}_{\mathbb{X}}{}^{m_t}$:

$$\min_{\mathbf{w}\in\mathbb{R}^{m_s}} \text{divergence}\left(\hat{S}_u^{\mathbf{w}}, \hat{T}_u\right) \quad \text{or} \quad \min_{\mathbf{p}\in\Delta_{m_s}} \text{divergence}\left(\hat{S}_u^{\mathbf{p}}, \hat{T}_u\right), \tag{2.49}$$

and varies depending on the used divergence measure. Other constraints can be added to this formulation in order enforce class-related criteria on the source domain. Nevertheless, these approaches are limited by the requirement that the source distribution dominates the target one, an assumption that is violated in several tasks. This assumption is no longer required in the next family of methods.

2.2.2 Feature Transformation Approaches

Instead of considering the instances' weights, feature transformation algorithms focus on modifying the features to move the two domains closer. This can be achieved either by letting one domain fixed and pushing the other towards it (where in most of the cases, it is the source domain that is transformed to match the target one) or by mapping both of the domains to a new latent space $\mathbb{U}$ where they are similar. As in Weiss et al. (2016), we call these two approaches *asymmetric* and *symmetric*, respectively. In what follows, we denote by ϕ a feature transformation map picked from a predetermined set Φ reflecting a prior knowledge about the problem at hand (similar to picking a hypothesis space $\mathbb{H}$ as an inductive bias for supervised classification). The symmetric and asymmetric approaches are summarized in Table 2.2, where F denotes a domain representation other than its probability distribution.

Approach	**Asymmetric**		**Symmetric**	
Mapping	$\phi : \mathbb{X} \to \mathbb{X}$		$\phi : \mathbb{X} \to \mathbb{U}$	
Domain representation	distribution	other	distribution	other
Goal	$\phi\#\mathcal{S} \approx \mathcal{T}$	$F(\phi\#\mathcal{S}) \approx F(\mathcal{T})$	$\phi\#\mathcal{S} \approx \phi\#\mathcal{T}$	$F(\phi\#\mathcal{S}) \approx F(\phi\#\mathcal{T})$
Restriction	Respect class information			

Table 2.2: The two main feature transformation-based DA approaches.

2.2.2.1 Asymmetric Feature Transformation Approaches

The underlying assumption in this case is that the covariate shift assumption holds between the target and the source domains after an unknown transformation ϕ is applied to this latter, *i.e.* $(\phi\#\mathcal{S})_{y|\mathbf{x}} = \mathcal{T}_{y|\mathbf{x}}$, where $\#$ denotes the pushforward operator (Definition A.2.5). Concretely, ϕ is approximated so that $\phi\#\mathcal{S}_{\mathbb{X}} \approx \mathcal{T}_{\mathbb{X}}$ when the goal is to align distributions, or that $F(\phi\#\mathcal{S}_{\mathbb{X}}) \approx F(\mathcal{T}_{\mathbb{X}})$ where F denotes some representation of the domains other than their distributions (Table 2.2).

Representation as distributions In this case, a measure of dissimilarity between the empirical domain distributions is used. For instance, the Wasserstein distance (Section 2.1.2.4) was employed in Courty et al. (2016) to align the domains. The idea of this contribution is to consider the primal form of the Wasserstein distance (Theorem 2.1.5), to calculate it with the computationally attractive entropy regularization (Cuturi, 2013) using the Sinkhorn iterations (Sinkhorn, 1967). Moreover, the authors introduce several transport map regularization strategies that prevent a target instance from receiving high probability masses from two source instances with different classes, hence taking the classification task into account. The problem they minimize has the following form:

$$\min_{\mathbf{P}\in\Pi(\hat{S}_u,\hat{T}_u)} \underbrace{\sum_{i,j}(\mathbf{P})_{ij}c(\mathbf{x}_{s,i},\mathbf{x}_{t,j})}_{\text{Transport cost}} + \lambda \underbrace{\sum_{i,j}(\mathbf{P})_{ij}(-\log((\mathbf{P})_{ij}))}_{\text{Entropic regularization}} + \eta \underbrace{\Omega_c(\mathbf{P})}_{\substack{\text{class-related}\\ \text{regularization}}}, \quad (2.50)$$

where $c : \mathbb{X}^2 \to \mathbb{R}_+$ is a cost function and $\lambda, \eta > 0$ are hyperparameters. Once the optimal coupling $\mathbf{P}^*$ is computed, the authors suggest mapping every source domain point $\mathbf{x}_{s,i}$ to a new point $\hat{\mathbf{x}}_{s,i}$ using the following rule

$$\hat{\mathbf{x}}_{s,i} := \arg\min_{\mathbf{x} \in \mathbb{R}^n} \sum_{j=1}^{m_t} (\mathbf{P}^*)_{ij} c(\mathbf{x}_{s,i}, \mathbf{x}_{t,j}). \tag{2.51}$$

The previous problem has a closed form solution when c is the squared Euclidean distance, and is called the *barycentric mapping* in this case (Ferradans et al., 2014). As indicated in Courty et al. (2016, Section 3.3), the barycentric mapping corresponds to building a new distribution that interpolates between the source and target's empirical distributions using the transport matrix's coefficients.

Other representations These approaches associate each domain with a representation other than its empirical distribution. For example, some lines of work associate each domain to a subspace defined by the matrix of its basis vectors, and thus aligning the domains boils down to aligning these two matrices. This is the case for the method proposed in Harel and Mannor (2011) and the Subspace Alignement (SA) algorithm (Fernando et al., 2013), where the authors represent each domain by $d-$base vectors of its principal components analysis (PCA) (Pearson, 1901) decomposition and align the two via a linear transformation parameterized by $\mathbf{A} \in \mathbb{R}^{n \times n}$. They further take the classes information into account by introducing a trade-off between $\mathbf{A}$'s role in domain alignment and in separating the classes on the source domain using the LMNN's algorithm cost function (Weinberger et al., 2006). Instead of subspaces, some approaches consider non-linear manifolds, such as in Aljundi et al. (2015), where a kernelized version of SA is used. Other domain representations are possible, such as considering the covariance matrices (Sun et al., 2016) used in the popular correlation alignment (CORAL) algorithm. We note that for these approaches, the dissimilarity measure between domains is reduced to comparing their representations and thus does not represent a metric: for example, the same covariance matrix may correspond to two different probability distributions.

2.2.2.2 Symmetric Feature Transformation Approaches

The earliest feature transformation based DA approaches fall in this category. They were motivated by the assumption that some of the features are domain-specific, whereas others are common for both domains. It is then appealing to identify these latter features and to use them to align the two domains. This idea was pushed further into "creating" these features via a transformation $\phi : \mathbb{X} \to \mathbb{U}$ applied to both the source and target instances, with the goal of reducing the distribution shift. In Figure 2.5, we illustrate the idea of these approaches in the case of a linear mapping ϕ.

In such approaches, it is crucial to ensure that Φ is constrained not to contain the trivial transformation $\phi : \mathbf{x} \to 0$, as this latter does align the two domains via collapsing them into one point at the expense of ignoring the label information. Transformation ϕ can be computed either via heuristic approaches, or via solving a minimization problem, and as with asymmetric transformations, different domain representations were considered in the literature.

Heuristic In this subfamily of approaches, the transformation applied to both domains is determined by a heuristic, rather than solving an optimization problem. This is the case of Structural Correspondence Learning (SCL) (Blitzer et al., 2006, 2007), where the authors consider a sentiment analysis task and define *pivot features* that appear frequently in both domains (since they use binary features). They use them to project the instances into a low-dimensional representation, and they concatenate

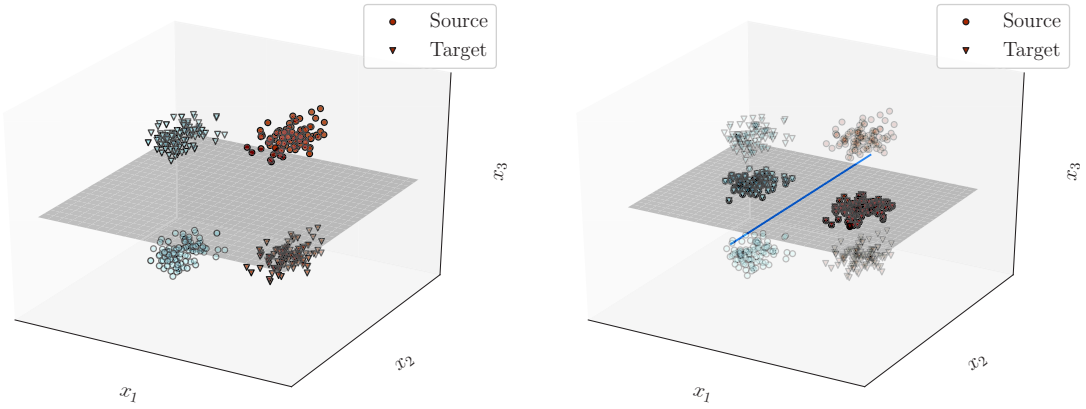


Figure 2.5: Illustration of a symmetric feature transformation approach: the 3D data are projected into a plane where the source and target domains are aligned. The classification boundary in the new space is represented by the blue line.

this latter with the original features as a new common space for both domains. Another heuristic approach was used in Aljundi et al. (2015) to select landmarks onto which the domains are projected with a kernel function used to align them.

Distribution divergence-based Such approaches align the empirical distributions of the two domains by minimizing a certain distance between them. The MMD distance (Section 2.1.2.3) is without a doubt one of the most leveraged metrics for these methods (Pan et al., 2008, 2011; Baktashmotlagh et al., 2016). For instance, in Pan et al. (2011), the authors project the RKHS representation of the data onto components determined by a minimization of the MMD distance between the two domains. These components are rows of a matrix satisfying an orthogonality constraint in the RKHS space, where the introduction of the latter constraint prevents the trivial alignment that collapses both domains to a single point.

Other representations As in the asymmetric case, a domain is associated to a certain representation. For instance, a subspace was used in several lines of work to perform adaptation. More precisely, in Gopalan et al. (2011, 2013); Chopra et al. (2013), the intuition was to mimic the gradual transfer occurring during the human learning by modeling it as points on a path on the Grassmannian manifold[10] that links the two domains. Both the source and target data are then projected into the resulting intermediate subspaces in the hope of bridging the gap between them. This idea was further developed in Gong et al. (2012, 2014) where the data points are projected into a geodesic path of intermediate domains using kernels and in Caseiro et al. (2015) where the projection path is constructed by smooth polynomial functions.

2.2.3 Simultaneously Aligning while Classifying

In most of the previously presented approaches, we gave examples of algorithms that are performed in two stages: aligning, then classifying. Now we consider methods in which these two steps are conducted jointly via solving an optimization problem, yielding directly a classifier that is good on the target domain. These approaches were used in the shallow setting, but they are dominantly adopted in the deep learning setting. Usually, an optimization problem of the following form is solved:

$$\min_{h \in \mathbb{H}} \mathfrak{C}^l_{\hat{S}}(h) + \eta \, \text{divergence}_h(\hat{S}_u, \hat{T}_u), \tag{2.52}$$

[10]The set of all subspaces of a given dimension.

where $\eta > 0$ is hyperparameter controlling the trade-off between minimizing the risk on the source domain (first term) and minimizing the divergence between the two domains (second term), where the latter may depend on hypothesis h. The cost function in this case bares a striking resemblance to general DA bounds without the non-estimable term (Section 2.1.3), and actually some lines of work learn h by minimizing the empirical counterpart of these bounds (Germain et al., 2013, 2016a; Courty et al., 2017).

Unless the weight placed on the alignment term dominates the rest, these methods have the virtue of being naturally immune to the trivial domain alignment problem (*i.e.* reducing both domains to a single point), as the supervised task information is preserved by the first term $\mathfrak{E}^l_{\hat{S}}(h)$.

2.2.3.1 Shallow Learning Approaches

In this case, one uses only the original features or an *a priori* fixed transformation of them (*e.g.* mapping the instances to an RKHS) before learning a classifier directly to have a good performance on the target domain. One of the earliest approaches in this regard was proposed in Morvant et al. (2012), where $(\epsilon, \gamma, \tau)-$good similarity functions (Section 1.4) are leveraged to map the data via ϕ^L from Theorem 1.4.1. Then, the following optimization problem is solved:

$$\min_{\mathbf{w}\in\mathbb{R}^{|L|}} \frac{1}{m_s}\sum_{i=1}^{m_s} l_+(\mathbf{w}^T\phi^L(\mathbf{x}_{i,s}), y_i) + \beta \sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \left\|(\phi^L(\mathbf{x}_s) - \phi^L(\mathbf{x}_t)) \odot \mathbf{w}\right\|_1 + \lambda\|\mathbf{w}\|_1, \tag{2.53}$$

where $\odot$ denotes the Hadamard product[11], $\mathcal{C}_{ST}$ is a pre-selected set of pairs defined by solving a bipartite graph matching problem and β, λ are positive hyperparameters.

In the same spirit, Germain et al. (2013), consider the following objective for linear classification:

$$\min_{\mathbf{w}\in\mathbb{R}^n} \frac{C}{m_s}\sum_{i=1}^{m_s} \ell_c\left(y_{s,i}.\frac{\mathbf{w}^T\mathbf{x}_{s,i}}{\|\mathbf{x}_{s,i}\|_2}\right) + A\left|\frac{1}{m_s}\sum_{i=1}^{m_s}\ell_d\left(\frac{\mathbf{w}^T\mathbf{x}_{s,i}}{\|\mathbf{x}_{s,i}\|_2}\right) - \frac{1}{m_t}\sum_{i=1}^{m_t}\ell_d\left(\frac{\mathbf{w}^T\mathbf{x}_{t,i}}{\|\mathbf{x}_{t,i}\|_2}\right)\right| + \|\mathbf{w}\|_2^2, \tag{2.54}$$

where $\ell_e := \frac{1}{2}(1 - \mathrm{erf}\left(\frac{\cdot}{\sqrt{2}}\right))$, ℓ_c is a convexification of ℓ, $\ell_d(u) := 2\ell(u)\ell(-u)$ and $A, C > 0$ are hyperparameters. This objective function is the empirical counterpart of the estimable part of the bound from Theorem 2.1.11 specialized to linear classification.

While approaches of Equations (2.53) and (2.54) have the same form as Equation (2.52), this latter does not cover all shallow approaches. For instance, Germain et al. (2016a) proposes another PAC-Bayesian approach corresponding to the bound from Theorem 2.1.12 when specialized to linear classifiers:

$$\min_{\mathbf{w}\in\mathbb{R}^n} B\sum_{i=1}^{m_s}\ell_c\left(y_{s,i}.\mathbf{w}^T\frac{\mathbf{x}_{s,i}}{\|\mathbf{x}_{s,i}\|_2}\right) + C\sum_{i=1}^{m_t}\ell_d\left(\mathbf{w}^T\frac{\mathbf{x}_{t,i}}{\|\mathbf{x}_{t,i}\|_2}\right) + \|\mathbf{w}\|_2^2, \tag{2.55}$$

where ℓ_e and ℓ_d are the same as Equation (2.54) and $B, C > 0$ are hyperparameters. Also, in Courty et al. (2017), the authors minimize the Wasserstein distance between the domains' joint distributions and replace the target labels by their prediction given by a hypothesis, resulting in the following optimization problem:

$$\min_{\substack{h\in\mathbb{H}\\ \mathbf{P}\in\Pi(\hat{S}_u,\hat{T}_u)}} \sum_{i=1}^{m_s}\sum_{j=1}^{m_t}(\mathbf{P})_{ij}\left(\underbrace{l(h(\mathbf{x}_{t,j}), y_{s,i})}_{\text{aligning the labels}} + \alpha \underbrace{c(\mathbf{x}_{s,i}, \mathbf{x}_{t,j})}_{\text{aligning the marginals}}\right). \tag{2.56}$$

[11] The Hadamard product of two vectors $\mathbf{u}$ and $\mathbf{v}$ is defined as: $(\mathbf{u}\odot\mathbf{v})_i := (\mathbf{u})_i.(\mathbf{v})_i$.

This cost function is the empirical Wasserstein distance in the bound from Theorem 2.1.13, thus it is theoretically justified even though it does not have the form of optimization problem (2.52).

Other approaches do not enjoy such theoretical guarantees but base their optimization problem on intuition, which is the case of (Quanz and Huan, 2009; Long et al., 2014) who use a projected version of the MMD distance (Section 2.1.2.3). Also, boosting algorithms are noteworthy, as they search for a classifier that is non-linear and that performs domain alignment as done in Habrard et al. (2013a).

2.2.3.2 Deep Learning Approaches

Deep artificial neural networks (Goodfellow et al., 2016) are extremely powerful in extracting new features from raw data representations, and have been applied successfully in several fields, such as computer vision (Lecun et al., 1998; Voulodimos et al., 2018; Grigorescu et al., 2020) and natural language processing (Zhang et al., 2018; Young et al., 2018), to name a few. This high expressive power of neural networks is exploited in the DA setting to learn features that ensure the alignment of the source and target domains in addition to guaranteeing a low risk on the source domain. Specifically, many deep learning approaches have the following form

$$\min_{\substack{h\in\mathbb{H}\\ \phi\in\Phi}} \mathfrak{E}^l_{\hat{S}}(h) + \text{divergence}_h(\phi\#\hat{S}_u, \phi\#\hat{T}_u), \tag{2.57}$$

which, when compared to Equation (2.52), introduces the feature map $\phi \in \Phi$, where the set Φ encodes the chosen neural network architecture.

One of the most influential deep DA approaches, with an optimization problem akin to Equation (2.57), was presented in Ganin et al. (2016). The authors based their approach on the theory introduced in Ben-David et al. (2010) and on adversarial training ideas from (Goodfellow et al., 2014). In particular, they train the network to extract features that allow discrimination between classes on the source domain, while hindering the performance of a second classifier that tries to distinguish between the two domains. Concretely, the divergence term from Equation (2.57) has the following form in Ganin et al. (2016):

$$\text{divergence}(\phi\#\hat{S}_u, \phi\#\hat{T}_u) := -\min_{h_{\text{dom}}\in\mathbb{H}} \sum_{\mathbf{x}\in S_u\cup T_u} l_{\text{soft}}(h_{\text{dom}}(\phi(\mathbf{x})), y_{\text{dom}}), \tag{2.58}$$

where h_{dom} is the domain classifier and y_{dom} is the domain label, taking values in $\{-1, 1\}$ and indicating whether instance $\mathbf{x}$ is from the source or the target distribution. This divergence term is the opposite of the minimum risk of the domain classifier h_{dom}, as suggested by Proposition 2.1.1.

By the same token, Zhang et al. (2019) derived an algorithm by minimizing the estimable part of the bound established in Theorem 2.1.14. In order to make the optimization process more tractable, they replace the ramp loss by the cross-entropy loss[12] and add a weight on the source risk term in the expression of the MDD (Definition 2.1.12). As a result of this modification, their domain alignment term also takes the form of the opposite of a domain classifier's risk.

Aside from approaches considering a domain classifier to define the divergence term, the MMD distance and several of its variations (Tzeng et al., 2014; Long et al., 2015; Bousmalis et al., 2016b; Long et al., 2017; Kang et al., 2019), as well as the Wassertein distance (Shen et al., 2018; Bhushan Damodaran et al., 2018; Le et al., 2019), are largely exploited in deep DA. Also, Equation (2.57) does not cover all of deep DA approaches, with notable exceptions given, for instance, by Chopra et al. (2013); Shu et al. (2018). The idea of the former is to create intermediary domains interpolating between the source

[12]The cross-entropy loss is a generalization of the softplus loss to the multi-class case.

and the target as in Gopalan et al. (2011); Gong et al. (2012), for which non-linear feature representations are learned by separate feature extractors and concatenated as a new data representation alignining the two domains. The latter considers a cost function that enriches the one presented in Equation (2.57) by terms enforcing the cluster assumption (Section 2.1.1.2) on the new shared representation of both domains. Finally, some deep DA approaches are the deep learning extensions of existing shallow DA methods, such as Sun and Saenko (2016) and Bhushan Damodaran et al. (2018) corresponding respectively to Sun et al. (2016) and Courty et al. (2017).

Finally, we note that some recent deep DA approaches relax the constraint of using the same feature map for the source and target domains in order to perform alignment (Rozantsev et al., 2018). For a certainly more exhaustive survey covering deep DA approaches, we refer the interested reader to Csurka (2017); Wang and Deng (2018); Wilson and Cook (2019).

2.2.4 Self-Labeling Approaches

While the two previously presented families of DA algorithms consider domain alignment, other lines of work adopted iterative approaches. In these methods, the classifier is initialized with the one that is good on the source domain, and then is gradually modified by adding target instances, while eliminating source ones according to certain criteria. Some of these methods are inspired by semi-supervised learning, a setting in which the labels are scarce, but the training and testing data are still supposed to come from the same probability distribution. One of the most popular approaches in this regard, called DASVM, was presented in Bruzzone and Marconcini (2010) and can be seen as an extension of transductive SVM approaches (Joachims, 1999; Chen et al., 2003) originally designed for the semi-supervised setting mentioned above. The underlying assumption of this approach is that if adaptability is possible between the two domains, then a classifier trained on the source domain cannot be at the same time confident and wrong on the same instance from the target domain. According to this principle, target data are gradually added while source data are removed and a classifier is fitted at each step until convergence. Following a similar idea, but no longer restricted to the SVM algorithm, Habrard et al. (2013b) used the similarity-based learning framework of Balcan et al. (2008a) to solve a DA task for the edit similarity function. Likewise, the CODA algorithm proposed in Chen et al. (2011a) is another iterative approach that considers logistic regression and uses self-training (McClosky et al., 2006), feature selection via l_1 regularization in order to select features behaving similarly on labeled source data and unlabeled target ones, and co-training (Chen et al., 2011b) in order to progressively adapt the classifier at hand to the target data.

2.2.5 Hybrid Approaches

The previously presented families of algorithms are not mutually exclusive, and in fact, some of them combine ideas from different approaches to improve their performance. For instance, in Morvant et al. (2012), an alternation between solving (2.53) and modifying the similarity induced space is carried out to move closer to each other the source and target domains. Also, in Aljundi et al. (2015), a symmetric feature transformation is used, followed by the asymmetric subspace alignment algorithm (Fernando et al., 2013). Another hybrid approach is presented in Sun and Saenko (2015) where the ideas of subspace alignment and instance re-weighting are combined.

2.2.6 Model Selection

The absence of labels for the target domain poses a real challenge for model selection. Several cross-validation procedures were proposed in the literature, among which the *reverse*

cross-validation introduced in Bruzzone and Marconcini (2010); Zhong et al. (2010). The authors of these works consider the following "reversed" DA problem: given the labeled source sample S and the unlabeled target sample T_u, let T^h be the target sample pseudo-labeled by h, and let S_u be the unlabeled source sample. When h is learned by a DA algorithm, the reverse cross-validation relies on the following assumption: if a hypothesis h_r, called *reverse hypothesis*, learned from T^h and S_u (*i.e.* by inverting the roles of the source and target samples) performs well on S_u (one can assess this performance since the source labels are available), then h is good for the initial DA problem. While this approach seems appealing as it does not need any target labels, the assumption it relies on may not hold in practice (Bousmalis et al., 2016b). Consequently, it is not deployed in the vast majority of DA approaches (Wilson and Cook, 2019, Section 8.2). Instead, several lines of work either use a subset of the target sample for hyperparameter tuning (Bousmalis et al., 2016b, Section 4) or present results with fixed hyperparameters for several tasks (Courty et al., 2017, Section 5). Finally, we note that a recent cross-validation approach for deep DA was proposed in You et al. (2019).

Conclusion

In this chapter, we covered the necessary background on DA required for the rest of this manuscript, both on the theoretical and algorithmic levels.

From the theory standpoint, we presented assumptions that are commonly considered to have a direct impact on the success of adaptation and addressed both their sufficiency and necessity. We are aware that our review of the domain adaptation theory is far from exhaustive, and excludes, for instance, algorithmic domain adaptation bounds of Mansour and Schain (2014) and hardness results linked to sample complexity of Ben-David and Urner (2012). For more exhaustive reviews, we refer the interested reader to Kouw and Loog (2019) and Redko et al. (2020).

At the algorithmic level, we presented several main families of DA approaches in a non-exhaustive manner and adopted a categorization that may be subject to discussion. For example, one can group algorithms according to whether they consider intermediate steps between domains or not. This is the case of self-labeling approaches that solve several intermediate learning problems corresponding to a source domain with a growing number of labeled instances coming from the target, or in Gopalan et al. (2011, 2013) and Gong et al. (2012), where intermediate domains are represented as points on a path on the Grassmannian manifold. For recent surveys on domain adaptation algorithms, we refer the interested reader to Zhang (2019) for DA as a subfield of transfer learning, Kouw and Loog (2019) for single-source DA, Zhuang et al. (2019) for several categorization of approaches and for Wilson and Cook (2019) for deep DA.

Part II

Contributions

Chapter 3

Revisiting (ϵ, γ, τ)-good similarities for domain adaptation

This chapter is based on contributions presented in two peer-reviewed conferences (Dhouib and Redko, 2018a,b).

Abstract This chapter encloses our early contributions and observations related to the $(\epsilon, \gamma, \tau)-$good similarities framework, as well as the reason we abandoned it in the rest of the manuscript. To this end, we first extend the theoretical analysis of similarity learning to the DA setting (Chapter 2), by introducing a new definition of an $(\epsilon, \gamma)-$good similarity function for DA, and by proving several results quantifying the performance of a similarity function on a target domain after it has been trained on a source domain. In particular, we show that if the source distribution dominates the target one, then principally new DA learning bounds can be proved. Finally, we provide a retrospective study, where we theoretically prove that despite the theoretical and algorithmic appeal, learning a bilinear similarity function with a Frobenius norm regularization preceding learning of a linear classifier is redundant, as the two are essentially equivalent.

Introduction

As explained previously (Section 1.4.4), the birth of the similarity learning field is mainly built upon the intuition that instances having the same label should be somehow similar when the classification task is considered. And while the earliest contributions to this field (Bellet et al., 2013) considered a supervised setting, similarity learning has been also widely used in the DA setting (Chapter 2) as confirmed by numerous algorithmic contributions (Geng et al., 2011; Perrot and Habrard, 2015b; Pinheiro, 2018). However, no general theoretical analysis of DA for similarity learning has been proposed before and thus it remained an open question.

In this chapter, we theoretically study the general $(\epsilon, \gamma, \tau)-$good similarities framework in the DA context, where we assume that the domains verify the covariate shift assumption (Section 2.1.1.2). Contrary to the previous works on the analysis of metric learning algorithms in DA (Geng et al., 2011; Morvant et al., 2012; Perrot and Habrard, 2015b), we aim to consider a more general setting without being attached to a particular algorithm in order to investigate to which extent a similarity that is good for a source domain remains good for the target one. This allows us to obtain several results that are novel in two different ways. First, they provide a complete theoretical study of similarity learning in DA that, due to the generality of the considered learning framework, naturally covers many possible learning scenarios. Second, we show that under certain assumptions on the richness of the source domain w.r.t. the target one, the target error can be bounded by terms that all explicitly depend on the source domain error.

The rest of the chapter is organized as follows. Section 3.1 introduces a generalization of the (ϵ, γ, τ)-goodness definition used to provide a theoretical result relating the source and target hinge risks (Equation (1.54)) and presents a brief comparison of the obtained bound with the related work. Apart from the source goodness, the established inequality contains a term reflecting the distance between the distributions of two domains and a worst margin term measuring the worst error obtainable by the similarity function for some instance from the learning sample. We further analyze the obtained worst margin term in Section 3.2, give the intuition for it, and measure the confidence of its empirical estimation. This part is then followed by section Section 3.1.2.1, where we examine more closely the possibility of learning a bilinear $(\epsilon, \gamma, \tau)-$good similarity, as previously done in the supervised learning setting by Bellet et al. (2012), as a potential DA algorithm minimizing the source error will have to rely on it too. Unfortunately, we obtain an interesting, yet negative, result that theoretically shows the equivalence between learning a bilinear similarity function with quadratic regularization and learning a linear classifier directly in the original input space whether it is for the purpose of supervised learning or, more generally, for transferring the learned similarity across two domains in the DA context. We conclude this chapter in Section 3.4 and give several possible future perspectives of this work.

3.1 $(\epsilon, \gamma)-$good Similarity Functions for DA

In this section, we introduce the main contributions of this chapter. First, we give a definition of $(\epsilon, \gamma)-$goodness with an arbitrary distribution of landmarks different from that introduced in Section 1.4.1. We make this particular choice in order to have a flexibility of considering landmarks from the target domain, as opposed to the supervised case, in which the landmarks and the training data are both drawn from the same distribution that, in the context of DA, would necessarily have to be the source distribution. Then, we quantify how well a similarity function that is good on the source domain will perform on the target one. This is done by bounding the target risk by the risk on the source and two other terms relating the two domains.

3.1.1 Problem Setup

In order to proceed, we first introduce the basic elements related to the $(\epsilon, \gamma, \tau)-$good similarity framework. We recall that $\mathbb{X} \subset \mathbb{R}^p$ is the feature space and we consider binary classification with labels encoded as $\mathbb{Y} = \{-1, 1\}$. Also, we recall that as in Balcan et al. (2008a), a similarity function is any function $K : \mathbb{X}^2 \to [-1, 1]$. For the considered problem, we assume having access to samples $S \sim \mathcal{S}^m$ and $T_u \sim {\mathcal{T}_{\mathbb{X}}}^n$, as defined in Section 2.1. Contrary to contribution Dhouib and Redko (2018b) on which this chapter is based, we no longer suppose that labels are deterministic, and consequently we make a clear distinction between any distribution $\mathcal{D}$ over $\mathbb{X} \times \mathbb{Y}$ and its marginal $\mathcal{D}_{\mathbb{X}}$ over $\mathbb{X}$. Also, we assume that the covariate shift assumption holds, *i.e.* $\mathcal{S}_{\mathbb{Y}|\mathbf{x}} = \mathcal{T}_{\mathbb{Y}|\mathbf{x}}$ for $\mathbf{x} \in \mathbb{X}$ (Section 2.1.1.2), as this assumption was considered in several DA contributions (Section 2.2.1) and can be leveraged via a re-weighting the source instances (Section 2.2.1.1).

As hinted in (Balcan et al., 2008a, Note 2, Theorem 14), the instances and landmarks can be potentially drawn from different distributions. Hence, we propose a modification of Definition 1.4.2 given as follows.

Definition 3.1.1. *A similarity function K is $(\epsilon, \gamma)-$good in hinge loss for problem $(\mathcal{D}, \mathcal{L})$ (where $\mathcal{D}$ and $\mathcal{L}$ are the respective data and landmarks distributions) if:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[l_+ \left(\frac{g_{\mathcal{L}}(\mathbf{x})}{\gamma}, y \right) \right] \leq \epsilon,$$

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

This is a generalization of Definition 1.4.2, and the two coincide when we consider the distribution $\mathcal{L}$ defined by $\mathbb{P}_{\mathbf{x}\sim\mathcal{L}_\mathbb{X}}[\mathbf{x} \in A] := \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathbb{X}}[\mathbf{x} \in A | R(\mathbf{x}) = 1]$ for all measurable sets A, where R is an indicator function that is potentially probabilistic, as in Definition 1.4.1. As for parameter τ, it can be seen as an upper bound for $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathbb{X}}[\mathbf{x} \in \operatorname{supp}\mathcal{L}_\mathbb{X}]$ since in this case, we have $\operatorname{supp}\mathcal{L}_\mathbb{X} \subseteq \{R(\mathbf{x}) = 1\}$. This definition captures the intuition often used to design DA algorithms as $\mathcal{L}$ can be thought of as a "universal landmarks domain" consisting in practice of instances coming from both the source or target domains. This is the case of several DA papers based on comparing instances with landmarks, including Morvant et al. (2012); Gong et al. (2013); Aljundi et al. (2015), where the common goal is to reduce the domain shift in the induced similarity space. Such landmarks are also reminiscent of the shared domains' features, as in the case of sentiment classification where they may correspond to negative or positive words used to express one's opinion independently of the type of the concerned product (Blitzer et al., 2007).

In the rest of the chapter, we use the following notations for any data distribution $\mathcal{D}$ and landmark distribution $\mathcal{L}$. We denote the $l_\gamma-$risk (or $\gamma-$scaled hinge risk) of K for problem $(\mathcal{D}, \mathcal{L})$ by

$$\mathfrak{E}_{\mathcal{D},\mathcal{L}}(K) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[l_\gamma \left(g_\mathcal{L}(\mathbf{x}), y \right)_+ \right],$$

where l_γ the γ-scaled hinge loss function defined by:

$$l_\gamma(y, y') = \ell_\gamma(y \cdot y') \quad \text{with} \quad \ell_\gamma : t \mapsto \left(1 - \frac{t}{\gamma}\right)_+ . \tag{3.1}$$

We let $\mathcal{U}$ be a probability over $\mathbb{X} \times \mathbb{Y}$, such that its marginal $\mathcal{U}_\mathbb{X}$ dominates all the other probability distributions over $\mathbb{X}$, and having a conditional distribution $\mathcal{U}_{\mathbb{Y}|\mathbf{x}} = \mathcal{S}_{\mathbb{Y}|\mathbf{x}} = \mathcal{T}_{\mathbb{Y}|\mathbf{x}}$. In addition, $\mathfrak{M}_{\mathcal{D},\mathcal{L}}(K)$ stands for the worst margin achieved by an element $\mathbf{x} \in \operatorname{supp}\mathcal{D}$ associated with landmark distribution $\mathcal{L}$, i.e:

$$\mathfrak{M}_{\mathcal{D},\mathcal{L}}(K) := \sup_{\mathbf{x}\in\operatorname{supp}\mathcal{D}} l_\gamma(g_\mathcal{L}(\mathbf{x}), y).$$

Note that since K takes values in $[-1, 1]$, $l_\gamma(g_\mathcal{L}(\mathbf{x}), y)$ is bounded thanks to the continuity of ℓ_γ. This ensures that $\mathfrak{M}_{\mathcal{D},\mathcal{L}}(K)$ is finite.

3.1.2 Relating the Source and Target $l_\gamma-$Risks

After having introduced the new $(\epsilon, \gamma)-$goodness definition above, we can now consider the general case of similarity learning with two landmark distributions $\mathcal{L}^s$ and $\mathcal{L}^t$, related to the source and target domains, respectively. Given a similarity function that is $(\epsilon, \gamma)-$good in hinge loss for problem $(\mathcal{S}, \mathcal{L}^s)$, our goal is to bound its $l_\gamma-$risk on the target domain for problem $(\mathcal{T}, \mathcal{L}^t)$. This generality makes it possible to investigate the impact of the difference between landmarks in addition to the shift between domains.

3.1.2.1 Shared Landmarks Distribution

In order to prepare for a more general result that relates the goodness of a similarity K for problems $(\mathcal{S}, \mathcal{L}^s)$ and $(\mathcal{T}, \mathcal{L}^t)$, we first provide a preparatory result that considers the same landmark distribution $\mathcal{L} = \mathcal{L}^s = \mathcal{L}^t$. This result is given by the following lemma.

Lemma 3.1.1 (same landmarks)**.** *Let K be an $(\epsilon, \gamma)-$good similarity for problem $(\mathcal{S}, \mathcal{L})$. Then K is $(\epsilon + \epsilon', \gamma)$-good for problem $(\mathcal{T}, \mathcal{L})$, where*

$$\epsilon' = d_{1+,\gamma}(\mathcal{T}, \mathcal{S}) \cdot \mathfrak{M}_{\mathcal{U},\mathcal{L}}(K)$$

and

$$d_{1+,\gamma}(\mathcal{T}, \mathcal{S}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}} \left[\left(\frac{\mathrm{d}\mathcal{T}_\mathbb{X}}{\mathrm{d}\mathcal{U}_\mathbb{X}}(\mathbf{x}) - \frac{\mathrm{d}\mathcal{S}_\mathbb{X}}{\mathrm{d}\mathcal{U}_\mathbb{X}}(\mathbf{x}) \right)_+ [y \cdot g_\mathcal{L}(\mathbf{x}) < \gamma] \right].$$

Moreover, if $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$ then the obtained results holds with

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

where $d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) = \underset{(\mathbf{x},y)\sim\mathcal{S}}{\mathbb{E}}\left[\left(\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) - 1\right)_+^2 [y \cdot g_{\mathcal{L}}(\mathbf{x}) < \gamma]\right]$.

Proof idea. We start by bounding $\mathfrak{E}_{\mathcal{T},\mathcal{L}}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}}(K)$ using an arbitrary dominating measure $\mathcal{U}$. Then, we apply Hölder's inequality for the L_1 and L_∞ norms of random variables to prove the bound with the $d_{1+,\gamma}(\mathcal{T},\mathcal{S})$ term. For the case $\mathcal{S}_{\mathbb{X}} \gg \mathcal{T}_{\mathbb{X}}$, we use the Cauchy-Schwartz inequality. □

Several observations can be made based on these results. First, we note that the expectation in both divergence terms is taken only on the support of the hinge loss, i.e for instances having a signed margin smaller than γ, making these terms problem-dependent. This dependence is quite important as it allows to claim that the presented result can be informative in practice. Second, the obtained bounds both contain the term $\mathfrak{M}_{\mathcal{U},\mathcal{L}}(K)$ which stands for the worst margin achieved by K on some instance of $\operatorname{supp}\mathcal{U}$. In the case of the SVM, this term is analogous to the largest slack variable associated to an instance drawn from the dominating measure $\mathcal{U}$. For several choices of $\mathcal{U}$, this term can be difficult to control, as we can estimate it only by observing data drawn from $\mathcal{S}$. This limitation is tackled by assuming that $\mathcal{S}_{\mathbb{X}}$ dominates $\mathcal{T}_{\mathbb{X}}$ thus motivating the bounds with χ^2 distance used as a divergence measure. These latter clearly show the benefit of assuming $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$: the distance term in the bound is multiplied by $\sqrt{\epsilon}$ meaning that having a similarity function achieving a low error on the source domain can leverage the difference between the domains' distributions. Note that the assumption $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$ is quite common in the DA literature and has already been used in Zhang et al. (2013). As mentioned by the authors, it roughly means that the source domain is richer than the target one, an assumption that is quite reasonable in practice.

Remark The result provided in Lemma 3.1.1 can be strengthened further in the case where $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$ by introducing a less strict supremum term. We provide this result below.

Lemma 3.1.2. *Let K be an $(\epsilon,\gamma)-$good similarity for problem $(\mathcal{S},\mathcal{L})$ and assume that $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$. Then K is $(\epsilon+\epsilon',\gamma)$-good for problem $(\mathcal{T},\mathcal{L})$, with*

$$\epsilon' = \sqrt{2 \sup_{(\mathbf{x},y)\in\operatorname{supp}\mathcal{S}} \left(\left(\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) - 1\right)_+ l_\gamma(g_{\mathcal{L}}(\mathbf{x}), y)\right) d_1(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}})\epsilon}$$

and

$$d_1(\mathcal{T}_{\mathbb{X}},\mathcal{S}_{\mathbb{X}}) = \underset{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}{\mathbb{E}}\left[\left|\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) - 1\right|\right].$$

Proof idea. Letting X and Y denote the non-negative random variables $\left(\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) - 1\right)_+$ and $l_\gamma(yg_{\mathcal{L}}(\mathbf{x}))$, we use the Hölder inequality for Hölder conjugates $p, q \in [1,\infty]$ and the Markov inequality to obtain a bound prametrized by a real number $t > 0$. Then we minimize the bound over the choice of t (as in establishing Chernoff bounds). The resulting bound is proportional to $\epsilon^{\frac{1}{q+1}}$, and the latter is minimized for $q = 1$ that we choose to obtain the expression of ϵ'. □

Compared to the result using the χ^2 distance in Lemma 3.1.1, this bound offers several interesting insights. First, the L^1 distance $d_1(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}})$ can be bounded in terms of other probability metrics, including χ^2 (by the Jensen inequality) and the Kullback-Leibler

divergence (via Pinsker's inequality), however it does not depend on the margin γ, contrary to the $d_{\chi^2_+,\gamma}(\mathcal{T}, \mathcal{S})$ from Lemma 3.1.1. Second, the supremum term concerns product of the l_γ loss and the difference between the Radon-Nikodym derivatives of $\mathcal{T}_\mathbb{X}$ and $\mathcal{S}_\mathbb{X}$ with respect to $\mathcal{S}_\mathbb{X}$. It means that even if there are regions on which the performance of $g_\mathcal{L}$ is poor (i.e high loss l_γ), it can be compensated by the low density of $\mathcal{T}_\mathbb{X}$ with respect to $\mathcal{S}_\mathbb{X}$ in that region, which is not the case for the $\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K)$ term in Lemma 3.1.1. As this result was obtained late during the thesis, we present below only the results from the original contribution (Dhouib and Redko, 2018b) and leave the analysis of the less strict supremum term for future works.

3.1.2.2 Different Landmarks Case

We now turn our attention to a more general case where the landmarks distributions vary across two domains. To this end, we assume that a similarity function K is $(\epsilon, \gamma)-$good for $(\mathcal{S}, \mathcal{L}^s)$. Given these assumptions, our goal now is to provide a learning guarantee for the goodness of K for the $(\mathcal{T}, \mathcal{L}^t)$ learning problem. To proceed, we first rewrite the difference between $\mathfrak{E}_{\mathcal{T},\mathcal{L}^t}(K)$ and $\mathfrak{E}_{\mathcal{S},\mathcal{L}^s}(K)$ as follows:

$$\mathfrak{E}_{\mathcal{T},\mathcal{L}^t}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}^s}(K) = \mathfrak{E}_{\mathcal{T},\mathcal{L}^s}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}^s}(K) + \mathfrak{E}_{\mathcal{T},\mathcal{L}^t}(K) - \mathfrak{E}_{\mathcal{T},\mathcal{L}^s}(K).$$

By analyzing the obtained expression, we note that the difference between the first two terms can be bounded using Lemma 3.1.1 as $\mathfrak{E}_{\mathcal{T},\mathcal{L}^s}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}^s}(K) = \epsilon + \epsilon' - \epsilon = \epsilon'$, where $\epsilon' = \sqrt{d_{\chi^2_+,\gamma}(\mathcal{T}, \mathcal{S}) \cdot \mathfrak{M}_{\mathcal{U},\mathcal{L}}(K)}\sqrt{\epsilon}$ when $\mathcal{T} \ll \mathcal{S}$ and $d_{1+,\gamma}(\mathcal{T}, \mathcal{S})\mathfrak{M}_{\mathcal{U},\mathcal{L}^t}(K)$ otherwise. Consequently, we further focus solely on the last two terms and, similar to the previous case, provide a result based on both the L_1 and χ^2 distances. We prove the following theorem.

Theorem 3.1.1. *Let K be an $(\epsilon, \gamma)-$good similarity for problem $(\mathcal{S}, \mathcal{L}^s)$. Assume that there exists probability $\mathcal{U}'$ over $\mathbb{X} \times \mathbb{Y}$ such that $\mathcal{U}'_\mathbb{X}$ dominates both $\mathcal{L}^s_\mathbb{X}$ and $\mathcal{L}^t_\mathbb{X}$, and that $\mathcal{U}'_{\mathbb{Y}|\mathbf{x}} = \mathcal{L}_{\mathbb{Y}|\mathbf{x}}{}^s = \mathcal{L}_{\mathbb{Y}|\mathbf{x}}{}^t$. Then K is $(\epsilon + \epsilon' + \epsilon'', \gamma)$-good for problem $(\mathcal{T}, \mathcal{L}^t)$, with*

$$\epsilon'' = \frac{1}{\gamma} d_K(\mathcal{L}^s, \mathcal{L}^t) \text{ and } \epsilon' = d_{1+,\gamma}(\mathcal{T}, \mathcal{S})\mathfrak{M}_{\mathcal{U},\mathcal{L}^s}(K),$$

where

$$d_K(\mathcal{L}^s, \mathcal{L}^t) := \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{T}_\mathbb{X} \\ \mathbf{x}'\sim\mathcal{U}'_\mathbb{X}}} \left[\left| \frac{\mathrm{d}\mathcal{L}^s_\mathbb{X}}{\mathrm{d}\mathcal{U}'_\mathbb{X}}(\mathbf{x}') - \frac{\mathrm{d}\mathcal{L}^t_\mathbb{X}}{\mathrm{d}\mathcal{U}'_\mathbb{X}}(\mathbf{x}') \right| \left| K(\mathbf{x}, \mathbf{x}') \right| \right].$$

Moreover, if $\mathcal{T}_\mathbb{X} \ll \mathcal{S}_\mathbb{X}$, then the obtained result holds with

$$\epsilon' = \sqrt{d_{\chi^2_+,\gamma}(\mathcal{T}, \mathcal{S}) \cdot \mathfrak{M}_{\mathcal{U},\mathcal{L}^s}(K)}\sqrt{\epsilon}.$$

Proof. Denoting $\frac{\mathrm{d}\mathcal{L}^s_\mathbb{X}}{\mathrm{d}\mathcal{U}_\mathbb{X}}$ and $\frac{\mathrm{d}\mathcal{L}^t_\mathbb{X}}{\mathrm{d}\mathcal{U}_\mathbb{X}}$ respectively by $\mathfrak{l}_s$ and $\mathfrak{l}_t$, we have:

$$\begin{aligned}
\mathfrak{E}_{\mathcal{T},\mathcal{L}^t}(K) - \mathfrak{E}_{\mathcal{T},\mathcal{L}^s}(K) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{T}} [l_\gamma(y \cdot g_{\mathcal{L}^t}(\mathbf{x})) - l_\gamma(y \cdot g_{\mathcal{L}^s}(\mathbf{x}))] \\
&\leq \frac{1}{\gamma} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{T}} [|y \cdot g_{\mathcal{L}^s}(\mathbf{x}) - y \cdot g_{\mathcal{L}^t}(\mathbf{x})|] \qquad (3.2) \\
&= \frac{1}{\gamma} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \left[\left| \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{U}} \left[\left(\mathfrak{l}_s(\mathbf{x}') - \mathfrak{l}_t(\mathbf{x}') \right) yy' K(\mathbf{x}, \mathbf{x}') \right] \right| \right] \\
&\leq \frac{1}{\gamma} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \left[\mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{U}} \left[\left| \left(\mathfrak{l}_s(\mathbf{x}') - \mathfrak{l}_t(\mathbf{x}') \right) yy' K(\mathbf{x}, \mathbf{x}') \right| \right] \right] \qquad (3.3) \\
&= \frac{1}{\gamma} \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{T}_\mathbb{X} \\ \mathbf{x}'\sim\mathcal{U}_\mathbb{X}}} \left[\left| \mathfrak{l}_s(\mathbf{x}') - \mathfrak{l}_t(\mathbf{x}') \right| \left| K(\mathbf{x}, \mathbf{x}') \right| \right]. \qquad (3.4)
\end{aligned}$$

Here (3.2) holds because l_γ is $\frac{1}{\gamma}-$Lipschitz. (3.3) is obtained applying Jensen inequality with the convexity of the $|\cdot|$ function. Line (3.4) comes from the fact that $|yy'K(\mathbf{x}, \mathbf{x}')| = |K(\mathbf{x}, \mathbf{x}')|$. As for ϵ', it is directly obtained by Lemma 3.1.1 depending on the assumption made about the absolute continuity of the target distribution w.r.t. the source distribution. □

We note that compared to Dhouib and Redko (2018b), we introduced the measure $\mathcal{U}'$ related to the landmark distributions in the same way as $\mathcal{U}$ is related to $\mathcal{S}$ and $\mathcal{T}$. Also, we changed the definition of the divergence between landmark distributions to $d_K(\mathcal{L}^s, \mathcal{L}^t)$, which depends on the considered similarity function and decreases the bound. The obtained result suggests that it is better to consider the same landmark distribution $\mathcal{L} = \mathcal{L}^s = \mathcal{L}^t$ for the two domains, as this assumption minimizes the bound by implying $\epsilon'' = \frac{1}{\gamma} d_K(\mathcal{L}^s, \mathcal{L}^t) = 0$. This conclusion is rather intuitive: in many DA algorithms, the source and target domains are aligned using a shared set of invariant components, and landmarks can be seen as invariant points allowing to adapt the similarity function efficiently across domains. For this reason, we focus on the case of a shared landmark distribution in the rest of the chapter.

3.1.3 Comparison with other Existing Results

We now briefly compare the obtained results with some previous related works. To this end, we note that the vast majority of DA results (Ben-David et al., 2010; Mansour et al., 2009b; Cortes and Mohri, 2011; Morvant et al., 2012) have the following form that we already presented in Equation (2.27):

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{S}}(h) + \text{divergence}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + t(\mathcal{S}, \mathcal{T}). \tag{3.5}$$

From Equation (3.5), we note that our result with χ^2 distance drastically differs from the traditional DA bounds as, contrary to them, it suggests that the source risk directly impacts all the terms in the bound. Indeed, the inequality in Equation (3.5) prompts us to minimize both the source error $\mathfrak{E}^l_{\mathcal{S}}(h)$ and the divergence term "divergence$(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$", assuming that $t(\mathcal{S}, \mathcal{T})$ is small, while our result shows that source error given by the goodness of the similarity function can partially leverage the divergence between the two domains as it multiplies the latter. To the best of our knowledge, the only two other results that have this multiplicative dependence between the source error and the divergence term are Mansour et al. (2009d) and Germain et al. (2016a), where variations of Rényi divergence were considered. Contrary to their contributions, our bound involves a divergence term that is restricted to the $[y \cdot g_{\mathcal{L}}(\mathbf{x}) < \gamma]$ set making it intrinsically linked to the considered hypothesis class. Also, while the results of Mansour et al. (2009a, Theorem 2,) relies on Renyi divergences, in our case we use the χ^2 divergence. Furthermore, we note that the bounds proposed in Germain et al. (2016a) involve a non-estimable term that, similar to λ in Equation (3.5) is assumed to be small while the worst margin term presented in our result is subject to the analysis provided in the next section.

3.2 Analysis of the Worst Margin Term

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

3.2.1 A Simple Bound for the Worst Margin

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

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

The last inequality comes from the fact that $K:\mathbb{X}\times\mathbb{X}\to[-1,1]$. Based on the obtained expression, we note from Lemma 3.1.1 that the target goodness can now be bounded in terms of both values that characterize the similarity function in the source domain, *i.e.* γ and ϵ as follows:

$$\mathfrak{E}_{\mathcal{T},\mathcal{L}}(K) \leq \epsilon + \left(1+\frac{1}{\gamma}\right) d(\mathcal{S},\mathcal{T}), \tag{3.6}$$

where $d(\mathcal{S},\mathcal{T})$'s expression depends on whether $\mathcal{S}_{\mathbb{X}}$ dominates $\mathcal{T}_{\mathbb{X}}$. On the other hand, replacing the worst margin term in the bound by constant γ prevents us from taking it into account when attempting to design a new adaptation algorithm based on the obtained bounds. Such an algorithm may take the form of a minimization of the bound over the choice of the similarity function K from a predefined set, akin to hypothesis spaces used for classification. In this case, the supremum term becomes a maximum over a finite set, which can be taken into account by a finite number of constraints. Hence, it can be useful to estimate this term empirically from the observed data sample by taking the empirical maximum for the source instances and the empirical mean for the landmarks.

3.2.2 An Empirical Estimation of the Worst Margin

We intend to measure our confidence in the empirical estimation of the worst margin term by bounding the deviation between the real worst margin term and its empirical counterpart. To this end, we suppose having access to a labeled data sample $S=\{(\mathbf{x}_1,y_1),...,(\mathbf{x}_m,y_m)\}\sim\mathcal{S}^m$, inducing an empirical distribution $\hat{S}$. Similarly, we define a sample $L=\{(\mathbf{x}'_1,y'_1),\ldots,(\mathbf{x}'_r,y'_r)\}$ and the corresponding empirical distribution $\hat{L}$. We can now prove the following result.

Theorem 3.2.1. *Let K be a similarity function defined on a feature space $\mathbb{X}$. Let $\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K)$ denote its worst performance associated to loss function l_γ and achieved by an example drawn from $\mathcal{S}$, where $\mathcal{L}$ is the landmarks distribution. Assume that $\mathcal{T}_{\mathbb{X}}\ll\mathcal{S}_{\mathbb{X}}$ and that the cumulative distribution function F_{l_γ} of the loss function associated with $\mathcal{S}$ and $\hat{L}$ is k times differentiable at $\mathfrak{M}_{\mathcal{S},\hat{L}}(K)$, and that $k>0$ is the minimum integer such that $F^{(k)}_{l_\gamma}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K))\neq 0$. Then, for all $\alpha>1, r\geq 1$, there exists $m_0\geq 1$ such that for all $m\geq m_0$, we have with probability at least $1-\delta$:*

$$\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K) \leq \mathfrak{M}_{\hat{S},\hat{L}}(K) + \frac{2}{\gamma}\operatorname{Rad}_r\left(\mathbb{H}_1(K)\right) + \frac{1}{\gamma}\sqrt{2\frac{\log\left(\frac{4}{\delta}\right)}{r}} + \left(\frac{(-1)^{k+1}\log\left(\frac{2\alpha}{\delta}\right)k!}{F^{(k)}_{l_\gamma}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K))m}\right)^{\frac{1}{k}},$$

where $\mathbb{H}_1(K)$ is the hypothesis class defined by $\mathbb{H}_1(K) := \{h_{\mathbf{x}} : \mathbf{x}'\mapsto K(\mathbf{x},\mathbf{x}'), \mathbf{x}\in\operatorname{supp}\mathcal{S}_{\mathbb{X}}\}$.

Proof idea. We begin by writing $\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K)-\mathfrak{M}_{\hat{S},\hat{L}}(K)=M_1+M_2$ where

$$M_1=\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K)-\mathfrak{M}_{\mathcal{S},\hat{L}}(K),\quad M_2=\mathfrak{M}_{\mathcal{S},\hat{L}}(K)-\mathfrak{M}_{\hat{S},\hat{L}}(K).$$

To bound M_1, we use classical learning theory techniques to establish a concentration inequality w.r.t. hypothesis class $\mathbb{H}_1(K)$. As for M_2, we use a Taylor expansion of the probability that it exceeds a given threshold, which allows to establish an asymptotic concentration inequality for the supremum. □

This theorem shows that under certain conditions, the empirical maximum is guaranteed to converge in probability to the real supremum of the distribution's support. The convergence rate depends on the complexity of the similarity function search space represented by the Rademacher complexity (Definition 1.2.5) term and on the regularity of the loss distribution function reflected by the $m^{-\frac{1}{k}}$ term. This last term dominates the convergence rate when $k > 2$, and we have in general a convergence rate that is $\mathcal{O}(m^{-\frac{1}{\max\{2,k\}}})$.

Due to the bound's dependence on the regularity of F_{l_γ}, knowing this cumulative distribution function is necessary for an explicit computation of the bound. Furthermore, in the case when k increases, it implies that we may need more data in order to have a truthful estimation of the function's regularity. Thus, this quantity may become non estimable, which goes in line with several other theoretical contributions (Morvant et al., 2012; Ben-David et al., 2010; Mansour et al., 2009b; Cortes and Mohri, 2011) where the learning bound includes an a priori non estimable term.

3.3 Limits of Learning with $(\epsilon, \gamma)-$good Similarity Functions

Originally, the aim behind the current chapter was to provide the theoretical foundation for a potential DA algorithm based on $(\epsilon, \gamma, \tau)-$good similarity functions. In this regard, we note that the established analysis suggests using the same landmarks for both the source and target distributions as it leads to tighter bounds on the target hinge risk. On the other hand, it crucially depends on learning a good similarity function which can be done following the approach proposed in Bellet et al. (2012). We chose to concentrate on this particular approach as we were attracted by some of its appealing properties, notably the sparsity of the resulting classifier in the similarity space and the non linear extension using the KPCA (Schölkopf et al., 1997). In this section, we conduct a retrospective study on the potential interest of learning a bilinear $(\epsilon, \gamma)-$good similarity function from data belonging to the set

$$\{K_{\mathbf{A}} : (\mathbf{x}, \mathbf{x}') \mapsto \mathbf{x}^T \mathbf{A} \mathbf{x}'; \mathbf{A} \in \mathbb{R}^{n \times n}; \|\mathbf{A}\|_2 \leq 1\},$$

i.e. , the set considered in Bellet et al. (2012).

3.3.1 Learning a Bilinear Similarity Function is Equivalent to Learning a Linear Classifier

To proceed, we fix the same landmark distribution for both domains and note that any considered cost function will depend on $g_{\mathcal{L}}$ as follows from Definition 3.1.1. For bilinear similarity functions, $g_{\mathcal{L}}(\mathbf{x})$ has the following form for all $\mathbf{x} \in \mathbb{X}$:

$$\begin{aligned} g_{\mathcal{L}}(\mathbf{x}) &= \mathop{\mathbb{E}}_{(\mathbf{x}', y') \sim \mathcal{L}} \left[\mathbf{x}^T \mathbf{A} \mathbf{x}' y'\right] \\ &= \mathbf{x}^T \mathbf{A} \mathop{\mathbb{E}}_{(\mathbf{x}', y') \sim \mathcal{L}} \left[\mathbf{x}' y'\right] = \mathbf{x}^T \mathbf{A} \boldsymbol{\mu}', \end{aligned}$$

where $\boldsymbol{\mu}' := \mathop{\mathbb{E}}_{(\mathbf{x}', y') \sim \mathcal{L}} [\mathbf{x}' y']$. Consequently, as long as one considers a shared landmark distribution, any cost function that minimized over the choice of $K_{\mathbf{A}}$ is a function of $\mathbf{A}\boldsymbol{\mu}'$, *i.e.* the problem to solve reduces to

$$\min_{\substack{\mathbf{A} \in \mathbb{R}^{n \times n} \\ \|\mathbf{A}\|_2 \leq 1}} \phi\left(\mathcal{S}, \mathcal{T}_{\mathbb{X}}, \mathbf{A}\boldsymbol{\mu}'\right), \tag{3.7}$$

where ϕ can be any function. With the previous general formulation, let us consider the set

$$\mathfrak{B} = \{\mathbf{A}\boldsymbol{\mu}'; \mathbf{A} \in \mathbb{R}^{n \times n} \text{ and } \|\mathbf{A}\|_2 \leq 1\}. \tag{3.8}$$

Problem (3.7) can be re-written as follows:

$$\min_{\mathbf{w}\in\mathfrak{B}} \phi(\mathcal{S}, \mathcal{T}_{\mathbb{X}}, \mathbf{w}). \tag{3.9}$$

In other words, minimizing the cost function of Equation (3.7) boils down to minimizing the latter cost function over $\mathbf{w}$, as long as the $\mathbf{w}$ has the form $\mathbf{A}\boldsymbol{\mu}'$ for some matrix $\mathbf{A}$ with $\|\mathbf{A}\|_2 \leq 1$. To understand whether one can relax the restriction on $\mathbf{w}$, we notice the following inclusion

$$\mathfrak{B} \subseteq \{\mathbf{w} \in \mathbb{R}^n; \|\mathbf{w}\| \leq \|\boldsymbol{\mu}'\|\} \tag{3.10}$$

that is due to the constraint $\|\mathbf{A}\|_2 \leq 1$. An interesting question is whether this inclusion is strict. To this end, let us study the converse: for any $\mathbf{w} \in \mathbb{R}^n$ with $\|\mathbf{w}\| \leq \|\boldsymbol{\mu}'\|$, we have:

$$\mathbf{w} = \mathbf{A}_{\mathbf{w}}\boldsymbol{\mu}' \text{ for } \mathbf{A}_{\mathbf{w}} := \frac{\mathbf{w}\boldsymbol{\mu}'^T}{\|\boldsymbol{\mu}'\|^2} \in \mathbb{R}^{n\times n}. \tag{3.11}$$

The norm of $\mathbf{A}_{\mathbf{w}}$ verifies the following:

$$\|\mathbf{A}_{\mathbf{w}}\| = \left\|\frac{\mathbf{w}\boldsymbol{\mu}'^T}{\|\boldsymbol{\mu}'\|^2}\right\| = \frac{\|\mathbf{w}\|\|\boldsymbol{\mu}'\|}{\|\boldsymbol{\mu}'\|^2} = \frac{\|\mathbf{w}\|}{\|\boldsymbol{\mu}'\|} \leq 1 \tag{3.12}$$

Hence, the above inclusion is an equality:

$$\mathfrak{B} = \{\mathbf{w} \in \mathbb{R}^n; \|\mathbf{w}\| \leq \|\boldsymbol{\mu}'\|\}. \tag{3.13}$$

où Consequently, Problem (3.7) becomes equivalent to:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^n \\ \|\mathbf{w}\|\leq\|\boldsymbol{\mu}'\|}} \phi(\mathcal{S}, \mathcal{T}_{\mathbb{X}}, \mathbf{w}), \tag{3.14}$$

and a solution to Problem (3.7) is $\mathbf{A}_{\mathbf{w}_*}$, where $\mathbf{w}_*$ is a solution of Problem (3.14). This leads to a significant decrease of the computational burden as the number of optimization variables is reduced from n^2 to n.

To summarize, we proved that finding an (ϵ, γ)-good bilinear similarity function parametrized by matrix $\mathbf{A}$ with $\|\mathbf{A}\|_2 \leq 1$ is exactly equivalent to finding a linear classifier $\mathbf{w}$ with $\|\mathbf{w}\| \leq \|\boldsymbol{\mu}'\|$, independently of the considered cost function and of the landmark distribution $\mathcal{L}$. This finding includes the work of Bellet et al. (2012) in which the authors consider learning a bilinear similarity function for supervised classification. Interestingly, the authors observe that solving the problem for different choices of landmark points does not change the performance of the obtained similarity, which is in line with the fact that our analysis is independent of the choice of the landmark distribution $\mathcal{L}$.

3.3.2 $\mathbf{A}_{\mathbf{w}_*}$ is of Rank 1

We saw that for every linear classifier $\mathbf{w}$, matrix $\mathbf{A}_{\mathbf{w}}$ (Equation (3.11)) satisfies $\mathbf{A}\mu' = \mathbf{w}$. Hence, we provided a constructive proof for the fact that the linear mapping

$$\begin{aligned} \Psi :&\mathbb{R}^{n\times n} \to \mathbb{R}^n \\ &\mathbf{A} \mapsto \mathbf{A}\boldsymbol{\mu}' \end{aligned}$$

is surjective, and it is not injective because $\dim(\mathbb{R}^{n\times n}) > \dim(\mathbb{R}^n)$, implying that for every $\mathbf{w} \in \mathbb{R}^n$, an infinity of matrices $\mathbf{A}$ verify $\mathbf{A}\boldsymbol{\mu}' = \mathbf{w}$. This motivates an investigation on what characterizes the particular solution $\mathbf{A}_{\mathbf{w}}$. For a fixed $\mathbf{w} \in \mathbb{R}^n$, let $\mathbf{A} \neq \mathbf{A}_{\mathbf{w}}$ be a matrix verifying $\mathbf{A}\boldsymbol{\mu}' = \mathbf{w}$. Taking into account the expression of $\mathbf{A}_{\mathbf{w}}$ from Equation (3.7), we obtain:

$$\langle \mathbf{A} - \mathbf{A}_{\mathbf{w}}, \mathbf{A}_{\mathbf{w}} \rangle \propto \left\langle \mathbf{A} - \mathbf{A}_{\mathbf{w}}, \mathbf{w}\boldsymbol{\mu}'^T \right\rangle = \mathrm{Tr}\left\{(\mathbf{A} - \mathbf{A}_{\mathbf{w}})\boldsymbol{\mu}'\mathbf{w}^T\right\} = 0, \tag{3.15}$$

where the last equality comes from the fact that $\mathbf{A}_{\mathbf{w}}\boldsymbol{\mu}' = \mathbf{A}\boldsymbol{\mu}' = \mathbf{w}$. Hence, for any matrix $\mathbf{A}$ verifying $\mathbf{A}\boldsymbol{\mu}' = \mathbf{w}$, we have $\mathbf{A} - \mathbf{A}_{\mathbf{w}} \perp \mathbf{A}_{\mathbf{w}}$. This shows that $\mathbf{A}_{\mathbf{w}}$ is the matrix with the minimum Frobenius norm and verifying $\mathbf{A}\boldsymbol{\mu}' = \mathbf{w}$. Now, if $\mathbf{w}_*$ is a vector solving Problem (3.14), then $\mathbf{A}_* := \mathbf{A}_{\mathbf{w}_*}$ is a solution of Problem (3.7). This latter is equivalent to the following regularized cost function:

$$\min_{\mathbf{A} \in \mathbb{R}^{n \times n}} \phi\left(\mathcal{S}, \mathcal{T}_{\mathbb{X}}, \mathbf{A}\boldsymbol{\mu}'\right) + \beta \|\mathbf{A}\|_2^2. \tag{3.16}$$

As a result, any matrix verifying $\mathbf{A} \neq \mathbf{A}_{\mathbf{w}_*}$ and verifying $\mathbf{A}\boldsymbol{\mu}' = \mathbf{w}_*$ will not modify the first term in Equation (3.16), while violating the minimality of the $2-$norm of the solution, proved in Equation (3.15). Hence, $\mathbf{A}_{\mathbf{w}_*}$ is the solution of Problem (3.16).

We note that the previous analysis can be made more explicit by considering the particular case of learning a bilinear similarity function using the quadratic loss in a supervised setting, which is the topic of Appendix E. In this latter, the solution $\mathbf{A}_*$ of the considered minimization problem has a closed form and is of rank 1.

3.3.3 Consequences for Learning a Linear Classifier in the Bilinear Similarity Space

The practical interest of the similarity functions studies in (Balcan et al., 2008b,a) is to learn a linear classifier in the induced similarity space, as suggested by Theorem 1.4.1. In the case of a bilinear similarity function, the mapping ϕ^L from Theorem 1.4.1 is linear. In fact, one has

$$\begin{aligned} \phi^L(\mathbf{x})^T &= (K_{\mathbf{A}_*}(\mathbf{x}, \mathbf{x}'_1), \ldots, K_{\mathbf{A}_*}(\mathbf{x}, \mathbf{x}'_{n'})) && (3.17) \\ &= (\mathbf{x}^T\mathbf{A}_*\mathbf{x}'_1, \ldots, \mathbf{x}^T\mathbf{A}_*\mathbf{x}'_{n'}) = \mathbf{x}^T\mathbf{A}_*\mathbf{X}'^T, && (3.18) \end{aligned}$$

where $\mathbf{X}'$ is the matrix with the drawn landmarks stacked as rows. Moreover, for our learnt bilinear similarity, $\mathbf{A}_*$ is of rank 1, hence ϕ^L has rank 1 too:

$$\phi^L(\mathbf{x}) = \mathbf{X}'\mathbf{A}_*^T\mathbf{x} = \mathbf{X}'\mathbf{A}_{\mathbf{w}_*}^T\mathbf{x} = \mathbf{X}'\boldsymbol{\mu}'(\mathbf{w}_*^T\mathbf{x}),$$

meaning that ϕ^L projects all the points onto the vector $\mathbf{X}'\boldsymbol{\mu}'$. Hence, all of the features in the ϕ^L space will be fully correlated, resulting in a badly conditioned optimization problem when searching for a classifier in the ϕ^L space. Furthermore, any linear classifier $\alpha \in \mathbb{R}^{n'}$ in the ϕ^L space would verify:

$$\boldsymbol{\alpha}^T\phi^L(\mathbf{x}) = (\boldsymbol{\alpha}^T\mathbf{X}'\boldsymbol{\mu}')\mathbf{w}^T\mathbf{x},$$

meaning that $\phi^L(\mathbf{x})$ is equal to $\mathbf{w}^T\mathbf{x}$ up to a multiplicative constant where the latter is a classifier in the original space. Hence, not only learning a quadratically regularized bilinear similarity function with shared landmarks is theoretically redundant with directly learning a classifier, but it also leads to a badly conditioned optimization problem when searching for a classifier in the similarity space.

3.4 Conclusions and Future Perspectives

In this chapter, we provided general theoretical guarantees for the similarity learning framework in the DA context. The obtained results contain a divergence term between the two domain distributions that naturally appears when bounding the deviation between the same similarity's performance on them, and a worst margin term measuring the worst error obtainable by the similarity function for some instance from the learning sample. Contrary to the previous generalization bounds established for DA problem, we showed that when the source distribution dominates the target one, the bound can be improved via

a multiplication by the $\sqrt{\epsilon}$ factor. We further analyzed the worst margin term and showed that its convergence to the true value depends on the complexity of the search space of the similarity function, as well as on the regularity of the hinge loss's cumulative distribution function in the neighborhood of its maximum (worst) value. Finally, we rigorously proved that in spite of the theoretical appeal of applying Theorem 1.4.1 (Balcan et al., 2008a) to bilinear similarity functions with quadratic regularization, and regardless of the considered loss function, the learned bilinear similarity does not improve classification performance w.r.t. directly learning a linear classifier in the original feature space. This last point was the reason why we preferred concentrating on the DA problem without involving the $(\epsilon, \gamma, \tau)-$good formalism in the rest of the thesis.

In the future, it is possible to extend the contributions of this chapter in multiple directions. First, in our new definition of the $(\epsilon, \gamma)-$goodness, the landmark distribution is assumed to be different from that used to generate source and target data samples, and thus a question about the existence of a landmark distribution that leads to tighter bounds naturally arises. Second, it would be interesting to explore the semi-supervised scenario, where the landmarks used to learn a similarity function emanate from the source and target distributions at the same time. In this case, one can expect to obtain a result showing that the goodness of a similarity function learned with source landmarks only is worse than when considering a mixture distribution. Finally, an extension of the established bounds beyond the covariate shift scenario and after relaxing the assumption $\mathcal{S}_{\mathbb{X}} \gg \mathcal{T}_{\mathbb{X}}$ seems necessary to tackle more challenging DA scenarios. In the light of the last section of this work, one may ask what regularization term to use in order to obtain a bilinear similarity function that is not equivalent to directly learning a linear classifier, and that has a rank greater than 1. More generally, determining under what conditions it becomes interesting to learn a good similarity function (in the sense defined in (Balcan et al., 2008a)), *i.e.* when it performs better than directly searching for a classifier, or using a predefined similarity function (*e.g.* a kernel), is an interesting future direction.

Chapter 4

Margin-aware Adversarial Domain Adaptation

This chapter is based on contributions Dhouib et al. (2019, 2020b).

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

Introduction

Since the inception of the DA field (Chapter 2), several theoretical contributions were proposed to analyze this problem in the statistical learning framework (Section 2.1). The general idea behind any such analysis usually consists in bounding the target domain error rate by a source error rate plus an estimable term reflecting a certain distance between domains, called the alignment or divergence term, plus a non-estimable term that is assumed to be small for adaptation to be possible (Section 2.1.3). To this end, the seminal work of Ben-David et al. (2007) considered the bounds for 0-1 loss in binary classification setting by introducing a divergence term that takes into account the complexity of the hypothesis space. Their results were further generalized for any loss function verifying the triangle inequality in Mansour et al. (2009b) and to a case when the hypothesis space is an RKHS in Cortes and Mohri (2014); Cortes et al. (2019). A somewhat different result was recently proposed in Zhang et al. (2019) where the authors provided generalization bounds for DA in the case of multi-class classification with source domain error defined by the margin violation rate. Finally, several DA bounds were proposed in Redko et al. (2017); Courty et al. (2017); Shen et al. (2018) for the specific case when the considered alignment term is given by the Wasserstein distance (Santambrogio, 2015).

At the algorithmic level, there have been a plethora of approaches that deal with the unsupervised domain adaptation problem (Section 2.2), and they can be roughly divided to shallow and deep methods. Most of shallow methods try to solve the problem in a two-step fashion by first aligning the source and target domains to make them indistinguishable, which then allows to apply classical supervised algorithms on the transformed data. A

notable recent approach to perform alignment is the use of optimal transport (Courty et al., 2016, 2017), which provides a well-founded way of finding a mapping aligning the source and target domains that minimizes the cost of transforming the source distribution into the target one (Sections 2.2.2.1 and 2.2.2.2). Deep domain adaptation methods have also known an impressive surge in their number, with the basic idea being the exploitation of their feature extraction capacity to learn representations that align the two domains, while distinguishing between the different classes of the source domain (Tzeng et al., 2014). One of the main reasons of this surge is the adversarial training procedure (Goodfellow et al., 2014) for the first time used for domain adaptation in Ganin and Lempitsky (2015), where the main idea is built upon the theoretical contribution in Ben-David et al. (2007).

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

1. We establish a first provisional bound on the hinge risk on the target domain with an arbitrary margin parameter. It is written as the sum of a hinge risk on the source domain, an alignment term, and a non-estimable one. The bound offers a better assessment of the quality of separation between classes in the target domain due to the margin parameter, and has an estimable part that is convex in the considered classifier.

2. Using the previous results as a starting point, we further concentrate on bounding the margin violation rate in the target domain by its counterpart from the source domain, a novel symmetric alignment term and a non-estimable term that shows the benefit of large margin separation on source domain for the success of adaptation. This novel result addresses most of the drawbacks of the previous bound, includes the work of Ben-David et al. (2007) (Theorem 2.1.6) as a special case, does not require the loss function to satisfy the triangle inequality as in Mansour et al. (2009b) (Theorem 2.1.7) and strengthens the result of Zhang et al. (2019) (Theorem 2.1.14) by replacing the misclassification target error with a stricter target margin violation risk.

3. We upper bound our alignment term by a distance defining an adversarial variation of the classic Monge-Kantorovich problem (Theorem 2.1.5). This latter is further shown to be upper-bounded by the original Wasserstein distance considered in Redko et al. (2017); Courty et al. (2017); Shen et al. (2018), thus leading to tighter bounds.

4. We derive a first OT-driven adversarial DA algorithm that outputs a classifier minimizing the estimable part of the obtained bound. This classifier is shown to outperform other OT-based DA methods on both synthetic and real-world datasets.

The rest of the chapter is organized as follows. Section 4.1 introduces the required preliminary knowledge and notations. Section 4.2.2 is dedicated to our theoretical contributions presenting a novel bound on the margin violation error on the target domain, its thorough analysis, and relation to other existing bounds. Then, in Section 4.3, we use it to derive an algorithm that is further specialized to linear classifiers, resulting in a convex programming formulation. Finally, in the last section, we evaluate our algorithm on a toy dataset and on a benchmark real-world problem.

4.1 Preliminary Knowledge

In this section, we present the problem setup of our study with the notations used throughout the chapter. We also recall background knowledge on learning bounds in DA (presented in Section 2.1) to allow a further comparison to them in the rest of the chapter.

4.1.1 Problem Setup and Notations

We consider binary classification in an unsupervised domain adaptation setting, in which source and target data are respectively drawn from $\mathcal{S}$ and $\mathcal{T}$, the joint distributions over the product space of instances and labels $\mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subseteq \mathbb{R}^p$ and $\mathbb{Y} = \{-1, 1\}$. We recall that although both domains are assumed to be labeled, only the labels of the source instances are observable during the learning stage.

To proceed, let $\mathbb{H}$ and $\mathbb{H}'$ denote two hypothesis spaces acting on $\mathbb{X}$ and taking values in $[-1, 1]$. For further developments, we define several quantities to assess classifiers' performances on different domains. Let $l_{\rho,\beta}$ be the loss function defined for any $1 > \rho, \beta \geq 0$ by

$$l_{\rho,\beta}(y, y') := \ell_{\rho,\beta}(y \cdot y'),$$

i.e. as in Equation (1.18), where $\ell_{\rho,\beta}$ is defined for any ρ, β as follows:

$$\ell_{\rho,\beta}(t) := \begin{cases} 1 - \frac{(t-\rho)}{\beta}, & \text{if } \beta > 0 \text{ and } \rho \leq t \leq \beta + \rho \\ [t < \rho], & \text{otherwise.} \end{cases} \tag{4.1}$$

From its definition, we note that $\ell_{\rho,0}(t) = [t < \rho]$, and that it verifies the following inequality for all $\rho, \beta > 0$ and $t \in \mathbb{R}$:

$$\ell_{\rho,0}(t) = [t < \rho] < \ell_{\rho,\beta}(t) < \ell_{\rho+\beta,0}(t) = [t < \rho + \beta], \tag{4.2}$$

illustrated in Figure 4.1. Depending on the values of ρ and β, loss function $l_{\rho,\beta}$ includes

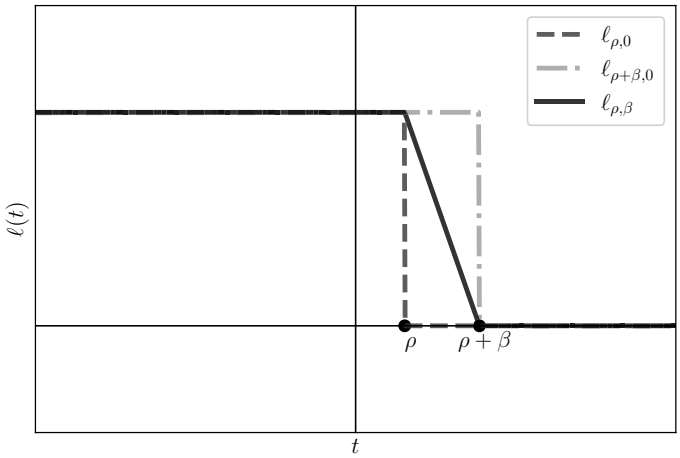


Figure 4.1: Loss function $\ell_{\rho,\beta}$ with its characteristic points and an illustration of the property from Equation (4.2).

cases that we already presented in Table 1.2: $l^{\rho,0}$ is the $\rho-$margin violation loss, $l^{0,\beta}$ is the ramp loss, whereas $l^{0,0}$ is the classic $0-1$ loss.

4.1.2 Background on DA Theory

All previous analyses of the DA problem aimed at bounding the target domain error by its source counterpart, an estimable divergence term and a non-estimable term representing the a priori adaptability of the problem (Section 2.1.3). Consequently, the major differences between the available results lie in the considered definition of the error, the divergence measure they introduce, and the form of the non-estimable term.

To this end, the first rigorous theoretical analysis of domain adaptation, presented in Ben-David et al. (2007) (and later in Ben-David et al. (2010)), introduces the alignment term given by the $\mathbb{H}\Delta\mathbb{H}-$divergence and the ideal joint error λ (Theorem 2.1.6). This

result was further generalized in Mansour et al. (2009b), where the authors considered an arbitrary symmetric loss function l verifying the triangle inequality (Theorem 2.1.7). This assumption makes their result more general than the one in Ben-David et al. (2010), holding only for 0-1 loss. Such a generalization to arbitrary loss function was later also provided for the bounds with the Wasserstein distance between joint (Courty et al., 2017) and marginal domains' distributions (Redko et al., 2017). While being more general than the bound of Ben-David et al. (2010)[1], these bounds, however, do not cover the margin violation loss as this latter does not verify the triangle inequality[2].

The MDD introduced recently in Zhang et al. (2019) (Definition 2.1.12) offers new insights on DA by introducing the ramp loss and scoring functions that give the confidence level of belonging to a class of interest rather than functions with binary outputs. However, as they bound the 0-1 risk on the target domain, *i.e.* $\mathfrak{E}^{01}_{\mathcal{T}}(h)$, their bound does not indicate the behavior of the margin violation rate on this latter.

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

4.2 Margin-aware Bounds on the Target Risk

This section concerns our theoretical contributions. We begin by bounding the scaled hinge risk on the target domain (as defined in Equation (3.1)) for a classifier h picked from a given hypothesis class $\mathbb{H}$. The bound establishes an intermediate result for which we outline several significant downsides related to the introduced divergence term. To strengthen the obtained result, we further bound the margin violation rate on the target domain and introduce a more meaningful divergence term, while maintaining the bound's dependence on the classification margin. Finally, we introduce a convex proxy for the obtained divergence term and use it further to derive an efficient optimization procedure allowing to minimize it.

4.2.1 A First Bound on the Scaled Hinge Risk on the Target Domain

In this section, we present our first attempt to propose a domain adaptation learning guarantee via directly bounding the scaled hinge risk, *i.e.* the loss function defined in Equation (3.1), of a given classifier on the target domain. Contrary to the previous chapter, we do this in a strictly more general setting without the covariate shift assumption and without assuming the dominance of the source marginal distribution of its target counterpart. The above-mentioned analysis was the main topic of our publication (Dhouib et al., 2019), where the bound concerned a classifier k defined by the means of a similarity function K, with a common landmark distribution for both domains, *i.e.* as for classifier $g_{\mathcal{L}}$ (from Definition 3.1.1). As $(\epsilon, \gamma, \tau)-$good similarities framework was shown to have some fundamental flaws when dealing with its algorithmic implementation, we extend the results from Dhouib et al. (2019) to a setting where no particular form is imposed for the considered classifier, apart from it taking values in $[-1, 1]$. At the same time, we keep considering the classification margin that lies at the heart of the theory of Balcan et al. (2008a) as it provides more information on the separation quality between classes in the target domain. We state our bound on the $l_\rho-$risk on the target domain in the following proposition.

Proposition 4.2.1. *Let l_ρ be a loss function defined by $l(y, y') := \ell_\rho(y \cdot y')$, where $\ell : t \mapsto$*

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

[2] We provide a proof for this claim in Proposition C.1.1.

$\left(1-\frac{t}{\rho}\right)_+$. *Then, for any* $h \in \mathbb{H}$, *and any* $0 < \rho \leq 1$, *we have:*

$$\mathfrak{E}^{\rho}_{\mathcal{T}}(h) \leq \mathfrak{E}^{\rho}_{\mathcal{S}}(h) + \frac{1}{\rho}\tilde{\Delta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \frac{1}{\rho}\tilde{\lambda}(\mathcal{S}, \mathcal{T}), \tag{4.3}$$

where

$$\tilde{\Delta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) := \sup_{h' \in \mathbb{H}'} \mathop{\mathbb{E}}_{\mathbf{x}_t \sim \mathcal{T}_{\mathbb{X}}} \left[\left| h(\mathbf{x}_t)h'(\mathbf{x}_t) - \mathop{\mathbb{E}}_{\mathbf{x}_s \sim \mathcal{S}_{\mathbb{X}}} \left[h(\mathbf{x}_s)h'(\mathbf{x}_s) \right] \right| \right], \tag{4.4}$$

$$\tilde{\lambda}(\mathcal{S}, \mathcal{T}) := \inf_{f \in \mathbb{H}'} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left[|y - f(\mathbf{x})| \right] + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \left[|y - f(\mathbf{x})| \right]. \tag{4.5}$$

Proof. We start by writing:

$$\mathfrak{E}^{\rho}_{\mathcal{T}}(h) = \underbrace{\mathfrak{E}^{\rho}_{\mathcal{T}}(h) - \mathfrak{E}^{\rho}_{\mathcal{T}}(h,f)}_{t_1} + \underbrace{\mathfrak{E}^{\rho}_{\mathcal{T}}(h,f) - \mathfrak{E}^{\rho}_{\mathcal{S}}(h,f)}_{t_2} + \underbrace{\mathfrak{E}^{\rho}_{\mathcal{S}}(h,f) - \mathfrak{E}^{\rho}_{\mathcal{S}}(h)}_{t_3} + \mathfrak{E}^{\rho}_{\mathcal{S}}(h). \tag{4.6}$$

For term t_3, we have:

$$\begin{aligned}
\mathfrak{E}^{\rho}_{\mathcal{S}}(h,f) - \mathfrak{E}^{\rho}_{\mathcal{S}}(h) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}} \left[\ell_\rho(h(\mathbf{x})f(\mathbf{x})) \right] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}} \left[\ell_\rho(h(\mathbf{x})y) \right] \\
&\leq \frac{1}{\rho} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left[h(\mathbf{x})(y - f(\mathbf{x}))_+ \right] \leq \frac{1}{\rho} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left[|y - f(\mathbf{x})| \right],
\end{aligned}$$

where we used the $\frac{1}{\rho}$–Lipschitzness of function ℓ_ρ and the fact h takes its values in $[-1, 1]$. Term t_1 can be bounded in the same manner. Concerning term t_2, we have:

$$\begin{aligned}
\mathfrak{E}^{\rho}_{\mathcal{T}}(h,f) - \mathfrak{E}^{\rho}_{\mathcal{S}}(h,f) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}} \left[\ell_\rho(h(\mathbf{x})f(\mathbf{x})) \right] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}} \left[\ell_\rho(h(\mathbf{x})f(\mathbf{x})) \right] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}} \left[\ell_\rho(h(\mathbf{x})f(\mathbf{x})) \right] - \ell_\rho \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}} \left[h(\mathbf{x})f(\mathbf{x}) \right] \right) \qquad (4.7) \\
&\leq \frac{1}{\rho} \mathop{\mathbb{E}}_{\mathbf{x}_t\sim\mathcal{T}_{\mathbb{X}}} \left[\left| h(\mathbf{x}_t)f(\mathbf{x}_t) - \mathop{\mathbb{E}}_{\mathbf{x}_s\sim\mathcal{S}_{\mathbb{X}}} \left[h(\mathbf{x}_s)f(\mathbf{x}_s) \right] \right| \right] \qquad (4.8) \\
&\leq \frac{1}{\rho} \sup_{h\in\mathbb{H}} \mathop{\mathbb{E}}_{\mathbf{x}_t\sim\mathcal{T}_{\mathbb{X}}} \left[\left| h(\mathbf{x}_t)h'(\mathbf{x}_t) - \mathop{\mathbb{E}}_{\mathbf{x}_s\sim\mathcal{S}_{\mathbb{X}}} \left[h(\mathbf{x}_s)h'(\mathbf{x}_s) \right] \right| \right],
\end{aligned}$$

where (4.7) is obtained by applying Jensen's inequality to the convex function ℓ_ρ. The $\frac{1}{\rho}$–Lipschitzness of the latter allows to obtain the last two lines. □

While being similar in shape to the previous DA bounds (Section 2.1.3), the result presented above concerns the scaled hinge risk, thus allowing to offer a better assessment on the quality of separation between classes. This is different from both Ben-David et al. (2010) and Zhang et al. (2019) that bounded the misclassification rate in the target domain. Also, the hinge risk is not included by the general study of Mansour et al. (2009b), as it does not verify the triangle inequality[3]. Another advantage of our bound is the convexity of its estimable part as a function of classifier h, making it convenient for optimization. Indeed, deriving an algorithm consisting in minimizing the estimable part of this bound was the topic of our contribution Dhouib et al. (2019).

Despite all the positive aspects discussed above, the previous result suffers from several limitations. First, the alignment term is not tight because in the case where the domains are identical, implying that $\mathcal{S}_{\mathbb{X}} = \mathcal{T}_{\mathbb{X}}$, it becomes

$$\tilde{\Delta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}}, \mathcal{S}_{\mathbb{X}}) := \sup_{h' \in \mathbb{H}'} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}} \left[\left| h(\mathbf{x})h'(\mathbf{x}) - \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}\sim\mathcal{S}_{\mathbb{X}}} \left[h(\tilde{\mathbf{x}})h(\tilde{\mathbf{x}}) \right] \right| \right].$$

[3] We prove this point in Corollary C.1.1.

This is the mean absolute deviation of random variable $hh'\#\mathcal{S}_{\mathbb{X}}$, which may well be different from zero, making the bound loose for identical domains. Second, by taking an expectation over distribution $\mathcal{S}_{\mathbb{X}}$ in the alignment term, we associate all of the target points to one "centroid"-alike source point potentially leading to another source of the bound's looseness. Finally, the non-estimable term is based on the absolute value loss, and this latter is not convenient for comparing between a classifier (more precisely, a scoring function) from $\mathbb{H}'$ and labels in the finite set $\{-1, 1\}$. Below, we aim at proposing a new analysis that addresses these drawbacks, while maintaining the virtues of Proposition 4.2.1 discussed above.

4.2.2 Bounding the Target Margin Violation Risk

Below, we present the main contributions from Dhouib et al. (2020b), where we achieve the goal stated above by proposing a bound on the margin violation rate in the target domain with a more meaningful divergence term admitting an interesting adversarial interpretation.

4.2.2.1 A Bound with a Non-convex Divergence between Distributions

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

Theorem 4.2.1. *Assume that for any $h' \in \mathbb{H}'$, we have $\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[h'(\mathbf{x}) = 0] = \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}}[h'(\mathbf{x}) = 0] = 0$. Let $\rho, \beta, \alpha > 0$ be such that $\rho + \beta < \alpha < 1$ and let $\rho' := \frac{\rho+\beta}{\alpha}$. Then, for any $h \in \mathbb{H}$, the following bound holds:*

$$\mathfrak{E}^{\rho,0}_{\mathcal{T}}(h) \leq \mathfrak{E}^{\rho',0}_{\mathcal{S}}(h) + d^{\rho,\beta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda_{\alpha},$$

where

$$d^{\rho,\beta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) := \sup_{h'\in\mathbb{H}'} \left| \mathfrak{E}^{\rho,\beta}_{\mathcal{S}}(h, h') - \mathfrak{E}^{\rho,\beta}_{\mathcal{T}}(h, h') \right|$$

and

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

Proof. In this proof, we use Lemma C.1.1 from Appendix C. By the first point of lemma Lemma C.1.1, we have:

$$\mathfrak{E}^{\rho,0}_{\mathcal{T}}(h) \leq \mathfrak{E}^{\rho,0}_{\mathcal{T}}(h, f) + \mathfrak{E}^{0,0}_{\mathcal{T}}(f). \tag{4.9}$$

Now, let us concentrate on bounding $\mathfrak{E}^{\rho,0}_{\mathcal{T}}(h, f)$. We have

$$\begin{aligned}
\mathfrak{E}^{\rho,0}_{\mathcal{T}}(h, f) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}}[[h(\mathbf{x})f(\mathbf{x}) < \rho]] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}}[l_{\rho,\beta}(h(\mathbf{x}), f(\mathbf{x}))] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}}[l_{\rho,\beta}(h(\mathbf{x}), f(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[l_{\rho,\beta}(h(\mathbf{x}), f(\mathbf{x}))] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[l_{\rho,\beta}(h(\mathbf{x}), f(\mathbf{x}))] \\
&\leq \sup_{h'\in\mathbb{H}'} \left| \mathfrak{E}^{\rho,\beta}_{\mathcal{T}_{\mathbb{X}}}(h, h') - \mathfrak{E}^{\rho,\beta}_{\mathcal{S}_{\mathbb{X}}}(h, h') \right| + \mathfrak{E}^{\rho+\beta,0}_{\mathcal{S}}(h, f),
\end{aligned} \tag{4.10}$$

where we used the lower bound on the function $\ell_{\rho,\beta}$ from equation (4.2). Finally, from the fact that $f \in \mathbb{H}'$, we take the supremum over $\mathbb{H}'$ and we use the upper bound on $\ell_{\rho,\beta}$,

again from equation (4.2), to obtain (4.10).
What is left to bound is $\mathbb{P}_{\mathbf{x}\sim\mathcal{S}_\mathbb{X}}[h(\mathbf{x})f(\mathbf{x}) < \rho + \beta]$, for which we use the second point of Lemma C.1.1 to obtain:

$$\mathfrak{E}_{\mathcal{S}}^{\rho+\beta,0}(h,f) \leq \mathfrak{E}_{\mathcal{S}}^{\frac{\rho+\beta}{\alpha},0}(h,f) + \mathfrak{E}_{\mathcal{S}}^{\alpha,0}(f)\,. \tag{4.11}$$

To sum up the different developments established up to now, *i.e.* , (4.9), (4.10) and (4.11), we have:

$$\mathfrak{E}_{\mathcal{T}}^{\rho,0}(h) \leq \mathfrak{E}_{\mathcal{T}}^{0,0}(f) + \mathfrak{E}_{\mathcal{S}}^{0,0}(f) + \mathbb{P}_{\mathbf{x}\sim\mathcal{S}_\mathbb{X}}[|f(\mathbf{x})| < \alpha] + d_{h,\mathbb{H}}^{\rho,\beta}(\mathcal{S}_\mathbb{X},\mathcal{T}_\mathbb{X}) + \mathfrak{E}_{\mathcal{S}}^{\frac{\rho+\beta}{\alpha},0}(h)\,.$$

Given that previous inequality holds for any choice of $f \in \mathbb{H}'$, minimizing it over this choice yields the result and introduces the ideal joint error. □

Similar to our previous result of Proposition 4.2.1, the bound from Theorem 4.2.1 offers a better estimation of the quality of separation between the classes in the target domain due to the margin violation loss instead of the misclassification rate. However, its estimable part given by the alignment term $d_{h,\mathbb{H}'}^{\rho,\beta}(\mathcal{S}_\mathbb{X},\mathcal{T}_\mathbb{X})$ is equal to 0 for $\mathcal{S} = \mathcal{T}$, making it tighter for similar domains. Also, compared to the bound from Equation (2.26) (from Zhang et al. (2019)), it does not introduce the decision function associated to a hypothesis (*i.e.* its sign in case of binary classification with labels encoded as -1 and 1), and thus avoids discontinuities in this term making it more suitable for optimization algorithms. One can further show that it is an integral probability metric (Zolotarev, 1984) as for a fixed $h \in \mathbb{H}$, the set $\{\mathbf{x} \mapsto l_{\rho,\beta}(h(\mathbf{x})h'(\mathbf{x})); h' \in \mathbb{H}'\}$ is constituted of bounded measurable functions, implying that $d_{h,\mathbb{H}'}^{\rho,\beta}(\mathcal{S}_\mathbb{X},\mathcal{T}_\mathbb{X})$ is a pseudometric on the space of probability distributions over $\mathbb{X}$. Finally, the non-estimable term λ_α, contrary to its counterpart from Proposition 4.2.1, quantifies the loss of the ideal joint hypothesis using the $0-1$ loss instead of the absolute value. Besides, it has the particularity of being non symmetric w.r.t. the the source and target domains' roles as, in addition to having low errors in both domains, it requires *only* a large absolute margin on the source domain, reflected by $\mathbb{P}_{\mathbf{x}\sim\mathcal{S}_\mathbb{X}}[|f(\mathbf{x})| < \alpha]$ where f_α is the function achieving the minimum in the expression of λ_α. This latter hence reflects an intuitively understandable behaviour: if one has a large margin of separation on $\mathcal{S}$, *i.e.* there exists α large enough with $\mathbb{P}_{\mathbf{x}\sim\mathcal{S}_\mathbb{X}}[|f_\alpha(\mathbf{x})| < \alpha]$ small enough, the space of classifiers that are good for the source becomes bigger. Consequently, it becomes more likely to find among them a classifier that is not only good on the source, but on the target domain too. This claim is also supported by the fact that for fixed $\rho, \beta \geq 0$, the source error term is the violation rate of margin $\rho' = \dfrac{\rho+\beta}{\alpha}$: as α increases, that rate decreases, implying that less concentration on the performance in the source domain is needed. Of course, no gain can be obtained from augmenting α if it considerably increases λ_α.

Additionally, compared to non-estimable terms from previous works, λ_α is at least equal to its counterpart λ from Ben-David et al. (2010), and is not directly comparable to $\lambda^{(\beta)}$ from Theorem 2.1.14 (Zhang et al., 2019). Finally, for the moment, β appears as the cost of the Lipschitz property of the loss function used in defining the discrepancy term. Its role will become clear when we bound our alignment term by a convex proxy.

In the following corollary, we formerly link our bound to that of Theorem 2.1.6.

Corollary 4.2.1. *If $\mathbb{H} = \mathbb{H}'$ is a class of binary hypotheses taking values in $\{-1,1\}$, the bound from Theorem 4.2.1 implies the one in Theorem 2.1.6 from Ben-David et al. (2010).*

Proof. Let $h \in \mathbb{H}$, $h' \in \mathbb{H}'$ and $y \in \{-1,1\}$. For $0 < \rho, \beta < 1$ and $1 > \alpha > \rho + \beta$, we have:

$$l_{\rho,\beta}(h(\mathbf{x})h'(\mathbf{x})) = [h(\mathbf{x})h(\mathbf{x}') < 0], \tag{4.12}$$

$$\left[y \cdot h(\mathbf{x}) < \frac{\rho+\beta}{\alpha}\right] = l_{\frac{\rho+\beta}{\alpha},0}(h(\mathbf{x}), y) = [y \cdot h(\mathbf{x}) < 0]. \tag{4.13}$$

This holds because for any $1 > \rho, \beta, \alpha > 0$ verifying $\rho + \beta < \alpha < 1$, functions $\ell_{\rho,\beta}$ and $[\cdot < 0]$ take the same values when restricted to the set $\{-1, 1\}$. Consequently, all the margins in the estimable part of our bound can be omitted, and it becomes equal to $\mathfrak{E}^{0,0}_{\mathcal{S}}(h) + d^{0,0}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$. We obtain:

$$\begin{aligned}
\mathfrak{E}^{0,0}_{\mathcal{T}}(h) &\leq \mathfrak{E}^{0,0}_{\mathcal{S}}(h) + d^{0,0}_{h,\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda_\alpha && (4.14)\\
&= \mathfrak{E}^{0,0}_{\mathcal{S}}(h) + \sup_{h' \in \mathbb{H}} \left| \mathfrak{E}^{0,0}_{\mathcal{S}}(h, h') - \mathfrak{E}^{0,0}_{\mathcal{T}}(h, h') \right| + \lambda_\alpha && (4.15)\\
&\leq \mathfrak{E}^{0,0}_{\mathcal{S}}(h) + \sup_{h,h' \in \mathbb{H}} \left| \mathfrak{E}^{0,0}_{\mathcal{S}_{\mathbb{X}}}(h, h') - \mathfrak{E}^{0,0}_{\mathcal{T}_{\mathbb{X}}}(h, h') \right| + \lambda_\alpha \\
&= \mathfrak{E}^{0,0}_{\mathcal{S}}(h) + d_{\mathbb{H}\Delta\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda_\alpha, && (4.16)
\end{aligned}$$

where (4.14) is from Theorem 4.2.1 and (4.15) comes from properties (4.12) and (4.13).

Then, taking the supremum over $h \in \mathbb{H}$ and the definition of $d_{\mathbb{H}\Delta\mathbb{H}}$ yield the rest of the developments. Finally, for any $f \in \mathbb{H}$, $\mathbb{P}_{\mathbf{x} \sim \mathcal{S}_{\mathbb{X}}}[|f(\mathbf{x})| < \alpha] = 0$ since $|f(\mathbf{x})| = 1$ for all $\mathbf{x} \in \mathbb{X}$ as f is a binary hypothesis. Hence,

$$\lambda_\alpha = \inf_{f \in \mathbb{H}} \mathfrak{E}^{0,0}_{\mathcal{T}}(f) + \mathfrak{E}^{0,0}_{\mathcal{S}}(f) + \mathbb{P}_{\mathbf{x} \sim \mathcal{S}_{\mathbb{X}}}[|f(\mathbf{x})| < \alpha] = \inf_{f \in \mathbb{H}} \mathfrak{E}^{0,0}_{\mathcal{T}}(f) + \mathfrak{E}^{0,0}_{\mathcal{S}}(f) = \lambda.$$

Combining this result with equation (4.16) yields the final result. □

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

4.2.2.2 A Convex Domain Divergence based on Optimal Transport

Although the bound from Theorem 4.2.1 offers several novel insights on the behavior of the target margin violation risk w.r.t. different components of the bound, its estimable part is non-convex as a function of hypothesis $h \in \mathbb{H}$. In particular, it involves the margin violation loss $[\cdot < \rho]$ for which the minimization is an NP-hard problem (Arora et al., 1997). In order to convexify its two components, namely the margin violation risk on the source domain and the divergence term, it is sufficient to bound the first using a convex surrogate loss l, while for the second we propose to leverage optimal transport (OT) theory (Section 2.1.2.4).

With the previous notations, the convex bound for $d^{\rho,\beta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$ is given in the following proposition involving the set of transport plans between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$ (Definition 2.1.10).

Proposition 4.2.2 (Convex bound for alignment term)**.** *For any $\rho, \beta > 0$, we have*

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

where

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

and $\Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$ is the set of transport plans between $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$.

Proof.

$$d^{\rho,\beta}_{h,\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) = \sup_{h' \in \mathbb{H}'} \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{T}_{\mathbb{X}}} \left[\ell_{\rho,\beta}(h(\mathbf{x}) h'(\mathbf{x})) \right] - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{S}_{\mathbb{X}}} \left[\ell_{\rho,\beta}(h(\mathbf{x}) h'(\mathbf{x})) \right] \right|$$

$$\leq \sup_{h' \in \mathbb{H}'} \sup_{|\varphi|_{\text{Lip}} \leq \frac{1}{\beta}} \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{T}_{\mathbb{X}}} \left[\varphi(h(\mathbf{x})h'(\mathbf{x}))\right] - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{S}_{\mathbb{X}}} \left[\varphi(h(\mathbf{x})h'(\mathbf{x}))\right] \right| \tag{4.17}$$

$$= \frac{1}{\beta} \sup_{h' \in \mathbb{H}'} W_1(hh'\#\mathcal{S}_{\mathbb{X}}, hh'\#\mathcal{T}_{\mathbb{X}}) \tag{4.18}$$

$$= \frac{1}{\beta} \sup_{h' \in \mathbb{H}'} \inf_{\mathcal{P} \in \Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})} \mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} \left[\left| h(\mathbf{x}_s)h'(\mathbf{x}_s) - h(\mathbf{x}_t)h'(\mathbf{x}_t) \right| \right] \tag{4.19}$$

$$\leq \frac{1}{\beta} \inf_{\mathcal{P} \in \Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})} \sup_{h' \in \mathbb{H}'} \mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} \left[\left| h(\mathbf{x}_s)h'(\mathbf{x}_s) - h(\mathbf{x}_t)h'(\mathbf{x}_t) \right| \right], \tag{4.20}$$

where we used the $\frac{1}{\beta}-$Lipchtizness of $\ell_{\rho,\beta}$ to obtain (4.17), in which the term $\sup_{|\varphi|_{\text{Lip}} \leq \frac{1}{\beta}}$ denotes a supremum over all $\frac{1}{\beta}-$Lipschitz functions. Then using the dual form of the Wasserstein distance W_1 between 1-dimensional distributions $hh'\#\mathcal{S}_{\mathbb{X}}$ and $hh'\#\mathcal{T}_{\mathbb{X}}$, we obtain (4.18). In the next line (4.19), we express the Wasserstein distance in its primal form (Theorem 2.1.5). Finally, we use the inf-sup inequality to obtain (4.20). □

For less cumbersome notations, we denote the $\Pi(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})$ by Π in what follows. To see the convexity of the introduced alignment term, we note that $d^{\rho,\beta}_{h,\mathbb{H}'}$ can be bounded by $\Delta_{\mathbb{H}'}(h, \mathcal{P})$ for any transport plan $\mathcal{P}$. Furthermore, the sets

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

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

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

The bound in Proposition 4.2.2 also has the form of a robust version of the Wasserstein distance between 1D distributions $hh'\#\mathcal{S}_{\mathbb{X}}$ and $hh'\#\mathcal{T}_{\mathbb{X}}$ and admits the following adversarial interpretation: for a fixed joint distribution $\mathcal{P} \in \Pi$, taking the supremum over $h' \in \mathbb{H}'$ is trying to separate the two domains, while taking $\mathcal{P}$ that achieves the infimum resists to this separation.

Combined with Theorem 4.2.1, Proposition 4.2.2 allows to immediately deduce a domain adaptation bound involving the introduced OT-based divergence.

Proposition 4.2.3 (Convex bound on the target risk)**.** *With the assumptions and notations of Theorem 4.2.1 and Proposition 4.2.2, assume further that l is a loss function defined by $l(h(\mathbf{x}), y) := \ell(y \cdot h(\mathbf{x}))$, where ℓ is non increasing and verifying $\ell(\rho') \neq 0$. Then, for any $h \in \mathbb{H}$:*

$$\mathfrak{E}^{\rho,0}_{\mathcal{T}}(h) \leq \frac{1}{\ell(\rho')} \mathfrak{E}^{l}_{\mathcal{S}}(h) + \frac{1}{\beta} \inf_{\mathcal{P} \in \Pi} \Delta_{\mathbb{H}'}(h, \mathcal{P}) + \lambda_{\alpha}. \tag{4.21}$$

Proof. We have:

$$\mathfrak{E}^{\rho',0}_{\mathcal{S}}(h) = \mathop{\mathbb{P}}_{(\mathbf{x},y) \sim \mathcal{S}} \left[y \cdot h(\mathbf{x}) < \rho' \right]$$

$$\leq \mathop{\mathbb{P}}_{(\mathbf{x},y) \sim \mathcal{S}} \left[\ell(y \cdot h(\mathbf{x})) \geq \ell(\rho') \right] \tag{4.22}$$

$$\leq \frac{1}{\ell(\rho')} \mathop{\mathbb{E}}_{(\mathbf{x},y) \sim \mathcal{S}} \left[\ell(y \cdot h(\mathbf{x})) \right] = \frac{1}{\ell(\rho')} \mathfrak{E}^{l}_{\mathcal{S}}(h), \tag{4.23}$$

where (4.22) is obtained due to the implication $y \cdot h(\mathbf{x}) < \rho' \Rightarrow \ell(y \cdot h(\mathbf{x})) \geq \ell(\rho')$ which is an immediate consequence of ℓ being non increasing. Then (4.23) is due to the Markov inequality. Combining the latter result with the bound from Proposition 4.2.2 yields the result. □

We emphasize that this bound is different from the one in our contribution Dhouib et al. (2020b), as in this latter we do not rigorously bound the $\rho'-$margin violation risk on the source domain, but we only state that we replace it using a convex surrogate. Moreover, in contrast with the bound from Proposition 4.2.1, the association between the source and target instances is more refined due to the infimum term that allows us to choose the optimal transport plan.

Compared to other DA bounds involving the Wasserstein distance (Redko et al., 2017; Courty et al., 2017; Shen et al., 2018), our divergence term (second term in the r.h.s. of Equation (4.21)) takes into account the considered hypothesis classes, making it a pseudo-metric that is less strict than the Wasserstein distance between marginal distributions of the domains. To support this claim, we bound our optimal transport based alignment term by the Wasserstein distance between the two domains.

Proposition 4.2.4 (Bounding by the Wasserstein distance)**.** *Let $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ be a metric, and assume that all of the hypotheses from $\mathbb{H}$ and $\mathbb{H}'$ verify the $L-$Lipschitz continuity w.r.t. metric d for some $L > 0$. Then, the following holds*

$$\sup_{h\in\mathbb{H}}\left(\inf_{\mathcal{P}\in\Pi}\Delta_{\mathbb{H}'}(h,\mathcal{P})\right) \leq \inf_{\mathcal{P}\in\Pi}\sup_{\substack{h\in\mathbb{H}\\ h'\in\mathbb{H}'}}\mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}}\left[\left|hh'(\mathbf{x}_s)-hh'(\mathbf{x}_t)\right|\right] \leq 2LW_1(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}}),$$

where

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

is the Wasserstein distance associated to metric d.

Proof. We have for any $h \in \mathbb{H}$, $h' \in \mathbb{H}'$ and $\mathbf{x}_s, \mathbf{x}_t \in \mathbb{X}$:

$$\begin{aligned}
\left|h(\mathbf{x}_s)h'(\mathbf{x}_s)-h(\mathbf{x}_t)h'(\mathbf{x}_t)\right| &= \left|h(\mathbf{x}_s)(h'(\mathbf{x}_s)-h'(\mathbf{x}_t))+h'(\mathbf{x}_t)(h(\mathbf{x}_s)-h(\mathbf{x}_t))\right| \\
&\leq |h(\mathbf{x}_s)|\left|h'(\mathbf{x}_s)-h'(\mathbf{x}_t)\right| + \left|h'(\mathbf{x}_t)\right||h(\mathbf{x}_s)-h(\mathbf{x}_t)| \quad (4.24)\\
&\leq \left|h'(\mathbf{x}_s)-h'(\mathbf{x}_t)\right| + |h(\mathbf{x}_s)-h(\mathbf{x}_t)| \quad (4.25)\\
&\leq 2L\cdot d(\mathbf{x}_s,\mathbf{x}_t), \quad (4.26)
\end{aligned}$$

where we apply the triangle inequality to obtain (4.24) and we use the fact that the hypotheses from $\mathbb{H}$ and $\mathbb{H}'$ have values in $[-1,1]$ to obtain (4.25). Then we use the $L-$Lipschitz property of h and h'.

Since inequality (4.26) holds for all choices of h et h', we can take the supremum over these two hypotheses, then the infimum over $\mathcal{P} \in \Pi$ to obtain:

$$\inf_{\mathcal{P}\in\Pi}\sup_{\substack{h\in\mathbb{H}\\ h'\in\mathbb{H}'}}\mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}}\left[\left|hh'(\mathbf{x}_s)-hh'(\mathbf{x}_t)\right|\right] \leq 2LW_1(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}}).$$

Then, bearing in mind the expression of $\Delta_{\mathbb{H}'}(h,\mathcal{P})$ (Proposition 4.2.2), the inf-sup inequality allows to write:

$$\sup_{h\in\mathbb{H}}\inf_{\mathcal{P}\in\Pi}\Delta_{\mathbb{H}'}(h,\mathcal{P}) \leq \inf_{\mathcal{P}\in\Pi}\sup_{\substack{h\in\mathbb{H}\\ h'\in\mathbb{H}'}}\mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}}\left[\left|hh'(\mathbf{x}_s)-hh'(\mathbf{x}_t)\right|\right]$$

which concludes the proof. □

This inequality comes essentially from the fact that the space of $L-$Lipchitz functions is richer than the considered hypothesis spaces $\mathbb{H}$ and $\mathbb{H}'$, and the supremum over $h \in \mathbb{H}$ is due to the fact that the Wasserstein distance between distributions is independent of classifier h. It formally shows that the attachment of our alignment term to the task at hand, via $h \in \mathbb{H}$ and the supremum over $h' \in \mathbb{H}$, makes it far less strict than the Wasserstein distance between marginal distributions of the domains.

4.3 Domain Adaptation Algorithm

With our considered setting and notations, the goal of our domain adaptation task is to find $h \in \mathbb{H}$ such that $\mathfrak{E}_{\mathcal{T}}^{\rho,0}(h)$, the margin violation rate on the target domain, is as small as possible. As we assume that we have no access to the labels of $\mathcal{T}$, we look for a hypothesis that minimizes the estimable part of our bound of Proposition 4.2.3.

4.3.1 Minimizing the Estimable Part of the Bound

As in most of DA approaches, we assume that our non-estimable term λ_α is small for adaptation to be possible. Our DA algorithm then consists in learning a classifier h that minimizes the remaining estimable part of the bound. Taking into account that the domain alignment term in Equation (4.21) is an infimum over $\mathcal{P} \in \Pi$ of convex functions of h (as we mentioned in Section 4.2.2.2), our bound results in the following optimization problem:

$$\min_{\substack{h\in\mathbb{H}\\ \mathcal{P}\in\Pi}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} [l(h(\mathbf{x}),y)] + \frac{\ell(\rho')}{\beta}\Delta_{\mathbb{H}'}(h,\mathcal{P}). \tag{4.27}$$

We underscore that the cost function in this case contains a supremum over the potentially infinite hypothesis space $\mathbb{H}'$. Hence, it might be difficult to solve Problem (4.27), even though convexity is verified. However, we show in the next section that a particular choice of $\mathbb{H}$ and $\mathbb{H}'$ allows one to deal efficiently with this term.

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

4.3.2 Application to Linear Classification

We consider our algorithm's formulation in the linear classification case, where $\mathbb{H}$ and $\mathbb{H}'$ are hypothesis spaces of bounded linear classifiers.

Proposition 4.3.1. *Let $\mathbb{H}$ and $\mathbb{H}'$ be the spaces of linear classifiers respectively with bounded Euclidean norm and $1-$norm. Then, Problem* (4.27) *can be equivalently expressed as the following convex program:*

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^n\\ \mathcal{P}\in\Pi}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \big[l(\mathbf{w}^T\mathbf{x},y)\big] + \delta\left\| \mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}} \big[\big|(\mathbf{x}_s\mathbf{x}_s^T - \mathbf{x}_t\mathbf{x}_t^T)\mathbf{w}\big|\big]\right\|_\infty + \zeta\|\mathbf{w}\|_2^2, \tag{4.28}$$

where l is a loss function defined as in Proposition 4.2.3 and $\delta, \zeta > 0$ are two hyper-parameters related to the bounds on $\mathbb{H}$ and $\mathbb{H}'$.

Proof. Let $\nu > 0$ and $\eta > 0$ be the respective radii of $\mathbb{H}$ and $\mathbb{H}'$, *i.e.*

$$\mathbb{H} \simeq \{\mathbf{w} \in \mathbb{R}^n; \|\mathbf{w}\|_2 \leq \nu\} \quad \text{and} \quad \mathbb{H}' \simeq \{\mathbf{v} \in \mathbb{R}^n; \|\mathbf{v}\|_1 \leq \eta\}, \tag{4.29}$$

where $\simeq$ denotes an equality up to an isomorphism of vector spaces. Also, let $\{\mathbf{e}_1, ..., \mathbf{e}_d\}$ be the canonical basis of $\mathbb{R}^n$. For $h \in \mathbb{H}$, *i.e.* , for $\mathbf{w} \in \mathbb{R}^n$ with $\|\mathbf{w}\|_2 \leq \nu$, the alignment term becomes:

$$\begin{aligned}
\Delta_{\mathbb{H}'}(h,\mathcal{P}) &= \sup_{h'\in\mathbb{H}'} \mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}} \big[\big|h(\mathbf{x}_s)h'(\mathbf{x}_s) - h(\mathbf{x}_t)h'(\mathbf{x}_t)\big|\big] \\
&= \sup_{\|\mathbf{v}\|_1\leq\eta} \mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}} \big[\big|\mathbf{v}^T(\mathbf{x}_s\mathbf{x}_s^T - \mathbf{x}_t\mathbf{x}_t^T)\mathbf{w}\big|\big] \\
&= \eta \sup_{1\leq k\leq d} \mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}} \big[\big|\mathbf{e}_k^T(\mathbf{x}_s\mathbf{x}_s^T - \mathbf{x}_t\mathbf{x}_t^T)\mathbf{w}\big|\big] \qquad (4.30) \\
&= \eta \sup_{1\leq k\leq d} \mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}} \big[\mathbf{e}_k^T\big|(\mathbf{x}_s\mathbf{x}_s^T - \mathbf{x}_t\mathbf{x}_t^T)\mathbf{w}\big|\big] \qquad (4.31)
\end{aligned}$$

$$
\begin{aligned}
&= \eta \sup_{1 \leq k \leq d} \mathbf{e}_k^T \mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} \left[\left| (\mathbf{x}_s \mathbf{x}_s^T - \mathbf{x}_t \mathbf{x}_t^T) \mathbf{w} \right| \right] \\
&= \eta \left\| \mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} \left[\left| (\mathbf{x}_s \mathbf{x}_s^T - \mathbf{x}_t \mathbf{x}_t^T) \mathbf{w} \right| \right] \right\|_\infty ,
\end{aligned}
$$

where after replacing h and h' by linear classifiers $\mathbf{w}$ and $\mathbf{v}$, we use the convexity of the function $\mathbf{v} \mapsto \mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} \left[\left| \mathbf{v}^T (\mathbf{x}_s \mathbf{x}_s^T - \mathbf{x}_t \mathbf{x}_t^T) \mathbf{w} \right| \right]$ for a fixed $\mathbf{w}$, and the fact that the $1-$norm unit ball is a polytope with vertices $\{\pm \mathbf{e}_k; 1 \leq k \leq d\}$ to obtain (4.30). Then, we use the identity $\left| \mathbf{e}_k^T \mathbf{u} \right| = \mathbf{e}_k^T |\mathbf{u}|, \ \forall \mathbf{u} \in \mathbb{R}^n$ to obtain (4.31). The last steps are deduced by the linearity of the expectation and the definition of the $\infty-$norm.

Satisfying constraint $\|\mathbf{w}\|_2 \leq \nu$ is equivalent to adding $\zeta \|\mathbf{w}\|^2$, where ζ is a Lagrange multiplier which has a one-to-one correspondence with ν. Finally, as $\frac{\ell(\rho')}{\beta}$ multiplies $\Delta_{\mathbb{H}'}(h, \mathcal{P})$ in Problem (4.27), setting $\delta = \frac{\eta \ell(\rho')}{\beta}$ yields the result. □

This proposition introduces two hyper-parameters linked to regularization of the classifier and to the alignment term. The further the domains are from each other, the more concentration we need on the alignment term, which is achieved by increasing δ. Also, we note that this is a strongly convex optimization problem, due to the strong convexity of the regularization $\|\mathbf{w}\|_2^2$, which is an important feature for numerical optimization with gradient descent, thus justifying our choice of the space $\mathbb{H}$.

4.3.3 Optimization Procedure for the Discrete Problem

In the empirical case, one has access to finite datasets $S = \{(\mathbf{x}_{s,i}, y_{s,i})\}_{i=1}^{m_s} \sim \mathcal{S}^{m_s}$ and $T_u = \{\mathbf{x}_{t,j}\}_{j=1}^{m_t} \sim \mathcal{T}_{\mathbb{X}}{}^{m_t}$, with S_u denoting the unlabeled part of S. The empirical cost function of Problem (4.28) becomes:

$$
\min_{\substack{\mathbf{w} \in \mathbb{R}^n \\ \mathbf{P} \in \Pi}} \quad \frac{1}{m_s} \sum_{1 \leq i \leq m_s} l(\mathbf{w}^T \mathbf{x}_{s,i}, y_{s,i}) + \delta \left\| \sum_{\substack{1 \leq i \leq m_s \\ 1 \leq j \leq m_t}} (\mathbf{P})_{ij} \left| (\mathbf{x}_{s,i} \mathbf{x}_{s,i}^T - \mathbf{x}_{t,j} \mathbf{x}_{t,j}^T) \mathbf{w} \right| \right\|_\infty + \zeta \|\mathbf{w}\|_2^2 , \tag{4.32}
$$

where $\Pi = \Pi(\hat{S}_u, \hat{T}_u)$ the set of transport matrices (Definition 2.1.10) between empirical distributions $\hat{S}_u$ and $\hat{T}_u$.

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

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

2. For a fixed linear classifier $\mathbf{w}$, the minimization over $\mathbf{P}$ only involves the term multiplied by δ in (4.32), and due to the positivity of all coordinates of vector $\sum_{ij} (\mathbf{P})_{ij} \left| (\mathbf{x}_{s,i} \mathbf{x}_{s,i}^T - \mathbf{x}_{t,j} \mathbf{x}_{t,j}^T) \mathbf{w} \right|$, this minimization is equivalent to[4]:

$$
\min_{\mathbf{q} \in \Delta_n} \max_{\mathbf{P} \in \Pi} \left(- \sum_{ij} (\mathbf{P})_{ij} \mathbf{q}^T \left| (\mathbf{x}_{s,i} \mathbf{x}_{s,i}^T - \mathbf{x}_{t,j} \mathbf{x}_{t,j}^T) \mathbf{w} \right| \right) , \tag{4.33}
$$

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

[4] A detailed derivation of Equation (4.33) is available in Section C.2.

We use smooth proxies of the positive part $(\cdot)_+$, the absolute value $|\cdot|$ and the infinite norm $\|\cdot\|_\infty$ (detailed in Section C.3.1).

4.3.4 Learning in Similarity Induced Spaces

Although the previous section concerns linear classification, this latter can be considered for non-linearly separable datasets after a certain transformation, such as the kernel trick for SVM's presented in Section 1.2.6.1. Different from supervised classification where one hopes that data becomes linearly separable after such transformation, its purpose in the DA setting is rather to guarantee the existence of a low-error ideal joint hypothesis. It turns out that for our choice of classifier class $\mathbb{H}'$, the analysis of the $(\epsilon, \gamma, \tau)-$good similarities framework (Section 1.4) implies that the ideal joint error term may be small in this case. Indeed, if we fix a similarity function K that happens to be good for the DA task at hand, then it is theoretically guaranteed that there exists a linear classifier with bounded $1-$norm, *i.e.* a classifier in $\mathbb{H}'$, that separates the data with a low error in the similarity space (Theorem 1.4.1). Note that we fix the similarity function K and do not learn it, in accordance with our retrospective study in Section 3.3. Mapping an instance to the similarity space is done as follows:

$$\phi^L(\mathbf{x}) = (K(\mathbf{x}, \tilde{\mathbf{x}}_1), ..., K(\mathbf{x}, \tilde{\mathbf{x}}_{n'})) \tag{4.34}$$

where $L = \{\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_{n'}\}$ is a finite set of landmarks usually defined as a subset of the original dataset. In the next experimental section, we solve Problem (4.28) after applying the mapping ϕ^L, but with a regularization term $\|\mathbf{w}\|_2^2$. Note that this latter does not depend on the similarity matrix $(\mathbf{K})_{ij} := K(\mathbf{x}_i, \mathbf{x}_j)$ for $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{X}$, unlike in the case of kernelized approaches where it is equal to $\mathbf{w}^T\mathbf{K}\mathbf{w}$, with $\mathbf{K}$ being the kernel matrix. We note further that in Balcan et al. (2008a), the authors recommend bounding the $1-$norm of $\mathbf{w}$ as a constraint. However, this is equivalent to bounding its ℓ_2 norm due to norm equivalence in finite dimension, which in turn is equivalent to adding a quadratic regularization. This latter is more suitable for optimization using gradient descent.

4.4 Empirical Evaluation

In this section, we evaluate our method on two domain adaptation problems: a toy set with controllable adaptation difficulty and a real-world sentiment analysis problem. For all experiments, we use the version of our algorithm specialized to linear classifiers as described in problem (4.28). We further use a similarity function K to be specified for each dataset considered as in Equation (4.34) to calculate the features from the raw data. Finally, we denote our method by **MADAOT** following the abbreviation of paper Dhouib et al. (2020b)'s title. The code for the different experiments is available on this link[5].

4.4.1 Hyper-parameter Tuning

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

[5] `https://github.com/sofiendhouib/MADAOT`.

Hence, we choose to present our algorithm's performance for two cases. In the first one, we do not use target labels during training phase, but we use them as a validation set to select the best hyper-parameters (defined in Proposition 4.3.1) via a 5-fold cross-validation procedure for 10 values of δ ranging from 10^{-2} to 10^{2}, and 10 values for ζ from 10^{-6} to 10^{-2}, both on a logarithm scale. This is a rather standard procedure in unsupervised domain adaptation used in several other papers on the subject (Courty et al., 2016; Bousmalis et al., 2016a). We use this procedure for the first dataset. As for the real-world dataset, we run all experiments by setting $\delta = 1$ and $\zeta = 10^{-5}$.

4.4.2 Intertwining Moons Data Set

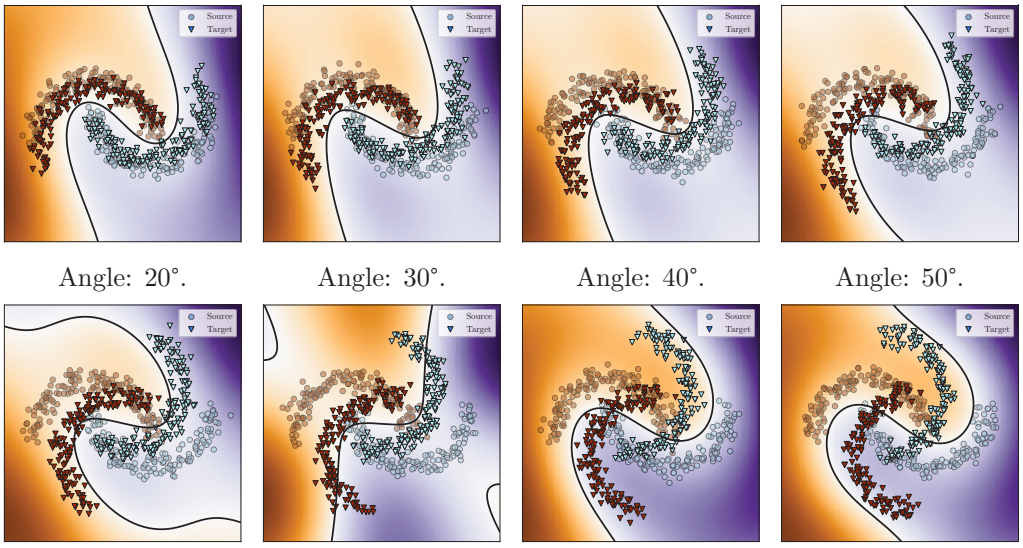
We carry on our experiments on the moons dataset used in Courty et al. (2016). For this dataset, the source domain's data sample is represented by two intertwining moons centered at the origin $(0, 0)$ and composed of 300 instances. The source domain's data are then rotated around their center by a certain angle to get the target domain data. Obviously, the greater is the angle, the further from each other the two domains are and the harder is the adaptation. Similar to Courty et al. (2016), we cope with the non-linearity of this dataset by using a Gaussian kernel as similarity function K, where the width parameter is chosen as the mean Euclidean distance between the source instances, as suggested in Kar and Jain (2011). As for the baselines, our algorithm is compared to an SVM classifier with Gaussian kernel trained on the source domain (without adaptation) and two optimal transport based domain adaptation algorithms: OT-GL (Courty et al., 2016) and JDOT (Courty et al., 2017). Note that we report only the variation of the method proposed in Courty et al. (2016) with the group-Lasso regularization as this latter was shown to be the most efficient for this dataset. Finally, as the results for JDOT on moons were not presented in the original paper, we run it with the hyper-parameter ranges suggested by the authors. The final results averaged over 10 tests on independent datasets of 1000 instances are presented in Table 4.1. From it, we can make several conclusions. First, all considered DA baselines manage to achieve an almost perfect score on the angles from 10° to 40°, while SVM without adaptation has a 30% drop in accuracy for these angles. This shows that the moons dataset presents a challenging adaptation task that goes beyond the generalization capacities of a standard supervised learning algorithm. As for the DA baselines, their performance is rather not surprising as for these angles the adaptation problem remains fairly easy. Starting from 50° (Table 4.1(left)) and up to 90° (Table 4.1(left)), our method provides a better performance that those obtained with both JDOT and OT-GL with the most significant improvement obtained for the angle of 50°. One should note that OT-GL method relies on the information about the source labels encoded in the group-lasso term but even this does not help to maintain its performance for larger angles. We conclude by saying that the theoretical features of the introduced OT-based distance used by our algorithm are highlighted by its efficiency in this experiment compared to strong OT baselines. To visually illustrate the classifier h output by our algorithm, we plotted the decision boundary for some rotation angles on Figure 4.2. Besides, the influence of hyperparameter δ is highlighted in Figure 4.3 and shows how it is crucial for the success of adaptation. More figures are available in Section C.3.2.

4.4.3 Sentiment Analysis Data Set

Below, we consider the famous Amazon product reviews dataset (Blitzer et al., 2007) related to the sentiment analysis task. For this dataset, we choose 4 of its subsets corresponding to different product categories, namely: books, DVD, electronics, and kitchen (denoted by B, D, E, K, respectively). This leads to 12 domain adaptation tasks of varying difficulty as the proximity and the number of semantic relationships between the different

Angle (°)	10	20	30	40	50	70	90
SVM (Courty et al., 2016)	**100**	89.6	76	68.8	60	26.6	17.2
OT-GL (Courty et al., 2016)	**100**	**100**	**100**	**98.7**	80.4	62.2	49.2
JDOT (Courty et al., 2017)	98.9	95.5	90.6	86.5	81.5	**70.5**	**60**
MADAOT	**99.5**	**99.3**	**99.6**	**99.6**	**98.9**	**77**	**64.1**

Table 4.1: Average accuracy over 10 realizations for the moons toy set.



Angle: 20°. Angle: 30°. Angle: 40°. Angle: 50°.

Angle: 60°. Angle: 70°. Angle: 80°. Angle: 90°.

Figure 4.2: Decision boundary for the inter-twinning moons dataset with different rotation angles to obtain the target domain. In this figure, the hyperparameters have the fixed values $\delta = 1$ and $\zeta = 10^{-5}$.

domains vary a lot. As the original data is represented by over 100 000 features given by uni- and bigrams, we follow the pre-processing of Chen et al. (2011a) (resulting in between 20000 and 40000 features) and consider a linear kernel as a similarity function K. For each task, we use predefined sets of 2000 instances of source and target data samples for training and keep 4000 instances of the target domain for testing. We compare our method to SVM with cross-validated hyper-parameters as a baseline, to the state-of-the-art adversarial DA approach DANN (Ganin et al., 2016) and JDOT (NN) with a neural network used as a classifier, as done in Courty et al. (2017). As our method uses only linear classifiers, we also run two baseline shallow algorithms for comparison, JDOT with a linear SVM and OT-GL with 1-Nearest Neighbor classifier (Courty et al., 2016). The results of our experiments are reported in Tables 4.2 and 4.3. From these tables, we see that MADAOT outperforms other methods on 8 out of 12 tasks, and has the second-best performance on 2 others. This is rather surprising considering that both DANN and JDOT (NN) rely on neural networks. Indeed, these latter are expected to have higher discriminative power than the class of linear classifiers. Consequently, we attribute this performance gain to the efficiency of our task-dependent OT-based alignment term that manages to better align the two distributions compared to the minimization of the original 2-Wasserstein distance considered in OT-GL and JDOT. Furthermore, the minimax formulation of our alignment term addresses the curse of dimensionality problem related to OT as the sample complexity of this latter is known to scale exponentially in dimension. Several approaches were proposed to address this problem recently in order for the calculation of the Wasserstein distance to make sense for high-dimensional data and our findings show that applying it

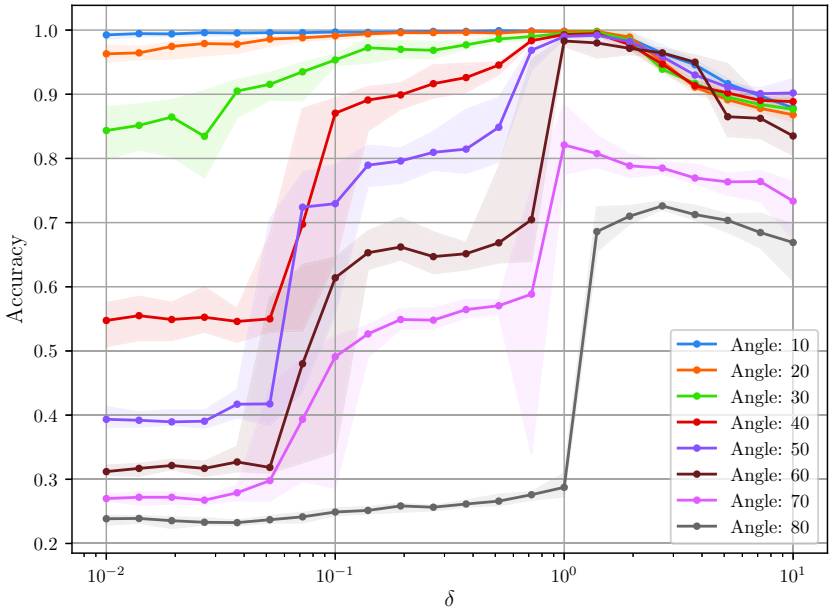


Figure 4.3: Influence of hyperparameter δ (with a logarithmic scale) on the accuracy, with $\zeta = 10^{-5}$ for all of the experiments. The experiments were repeated 30 times and the median is reported along with the interquartile range.

in the DA context can lead to improved performance.

Task	B→D	B→E	B→K	D→B	D→E	D→K
SVM (CV)	79.5	69.2	71.4	**78.5**	71.7	76.6
DANN	**80.6**	74.7	76.7	74.7	73.8	76.5
OT-GL	75.7	75.4	77.9	75.9	70	78
JDOT_{SVM}	75	**77.1**	77.9	70.9	**78.3**	78
JDOT_{NN}	79.5	**78.1**	**79.4**	76.3	**78.8**	**82.1**
MADAOT	**82.4**	75	**80.4**	**80.9**	73.5	**81.5**

Table 4.2: Accuracy on the Amazon Reviews dataset (part 1).

Task	E→B	E→D	E→K	K→B	K→D	K→E
SVM (CV)	71	73.8	85.3	72.3	73.4	84.4
DANN	71.8	72.6	85	71.8	73	**84.7**
OT-GL	71.2	69.6	81.4	**73.1**	74.1	81.7
JDOT_{SVM}	65.9	71.7	79.4	66.8	66.1	77.5
JDOT_{NN}	**74.9**	**73.7**	**87.2**	72.8	**76.5**	84.5
MADAOT	**77.2**	**78.1**	**88.1**	**75.6**	**75.9**	**87.1**

Table 4.3: Accuracy on the Amazon Reviews dataset.

4.5 Conclusion and Future Perspectives

In this chapter, we presented a novel theoretical analysis of the unsupervised domain adaptation problem for binary classification that considers the margin violation loss on

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

The future research directions of this chapter are many, both in terms of the theoretical results and of the derived algorithm. Primarily, our established bounds would be strengthened when extended to the multi-class case. One way of doing this is by considering the generalized margin used in Equation (2.24) that would require to come up with a new approach for establishing a bound that generalizes ours. Another way is to reason on pairs of instances and to consider a similarity function instead of classifier h. This choice, however, comes with the challenge of sampling the pairs that are not independent, in addition to their number that scales quadratically with the size of the dataset. Aside from the previous line of research, we plan to investigate in more detail the theoretical properties of our data dependent optimal transport term by establishing a concentration inequality for this latter. This would allow to theoretically highlight the success of our algorithm for high-dimensional data.

On the algorithmic level, other choices of $\mathbb{H}$ and $\mathbb{H}'$ are to be investigated. In the linear case, a different choice of $\mathbb{H}$ will boil down to changing the regularization term, while setting $\mathbb{H}'$ to be the space of L_p bounded linear classifiers may be less trivial [6]. More generally, in the non-linear case, we would like to study the direct maximization of our non-convex alignment term introduced in Theorem 4.2.1 using a deep adversarial approach. Even though our method offers a remarkable performance compared to several deep learning baselines, its efficiency can be further improved by this latter extension.

[6] We detail this point in Section C.4.

Chapter 5

Minimax Optimal Transport

This chapter is based on contribution Dhouib et al. (2020a), proposed in collaboration with Tanguy Kerdoncuff[1], Rémi Emonet[2] et Marc Sebban[3] from Hubert Curien Laboratory, University of Jean Monnet, St Etienne.

Abstract The Optimal transport (OT) problem and its associated Wasserstein distance have recently become a topic of great interest in the machine learning community to compare widely used probability distributions. However, calculating OT in practice requires choosing the ground metric that reflects in the best possible way the user's knowledge about the problem at hand. In this chapter, inspired by the adversarial optimal transport problem variation derived in the previous chapter, we propose a general formulation of a minimax OT problem that jointly optimizes the ground metric and the transport plan, allowing us to define a robust distance between distributions. We analyze the proposed formulation theoretically in several broad cases of practical interest, show its tight links and advantages compared to previously proposed works, and derive efficient algorithms for each case considered. Additionally, we use this method to define a notion of stability, allowing us to select a ground metric that is robust to bounded perturbations. Finally, we provide an experimental study highlighting the efficiency of our approach.

Introduction

In many scientific areas, we are often confronted with the necessity of comparing different objects to assess their relatedness. In machine learning, for instance, these objects may be individual data points in similarity-based classification (e.g., k-nearest neighbors (Cover and Hart, 1967), non-linear support vector machines (Boser et al., 1992)) or probability distributions in generative modelling (Goodfellow et al., 2014) and hypothesis testing. For this latter case, the optimal transportation (OT) metric (also called the Wasserstein distance) has recently emerged as a powerful tool used to compare complex objects based on the OT problem (Monge, 1781) that roughly quantifies the minimal amount of effort required to transform one distribution into another. Several key features of this metric lead to its widespread use in many different applications and setups (Gramfort et al., 2015; Kusner et al., 2015; Bonneel et al., 2016; Courty et al., 2017; Laclau et al., 2017). First, it takes into account the geometry of the underlying data distributions by the means of pairwise costs calculated for the points that they are supported on. Second, it allows to compare distributions with disjoint supports thus avoiding the vanishing gradient problem (Arjovsky et al., 2017) when used as a loss function.

[1] `https://hv0nnus.github.io/index.html`
[2] `https://home.heeere.com/`
[3] `https://perso.univ-st-etienne.fr/sebbanma/`

In the previous chapter, we proposed a variation of the Wasserstein distance that depends on the considered hypothesis classes, resulting in an adversarial formulation for the total transport cost. In this chapter, we generalize the last formulation by studying the OT problem with a minimax objective function, where one seeks an OT plan w.r.t. the worst possible ground cost function (which can be a metric), belonging to an arbitrary and possibly infinite convex set. Such a minimax formulation is of particular interest as it has been shown previously (i) to reduce the sample complexity and increase the robustness to noise of the original OT problem for high-dimensional data (Paty and Cuturi, 2019), (ii) to allow to consider submodular cost functions (Alvarez-Melis et al., 2018) and (iii) to use it as a loss in generative models (Genevay et al., 2018). We advance the study of the minimax OT further by providing the following contributions. First, for an infinite set of cost functions defined by a Mahalanobis distance, we reformulate the minimax OT problem as a minimization of the arbitrary dual norm of the matrix of second-order displacements and show how one can use it to smoothly interpolate between the original OT problem and a special case of the minimax formulation of Paty and Cuturi (2019). Second, we provide a generic solver for minimax OT for both regularized and unregularized minimax OT problems, and both finite and infinite families of cost functions, contrary to previous work (Paty and Cuturi, 2019; Alvarez-Melis et al., 2018) that considered optimization algorithms theoretically justified for smooth minimax cost functions only. Finally, we introduce the notion of cost matrix stability and solve its underlying optimization problem. It consists in finding a cost function from a list of possible candidates that leads to a stable transportation cost in its unit ball neighborhood.

The rest of this chapter is structured as follows. In Section 5.1, we provide the necessary introductory definitions related to the OT problem and the notations used throughout the chapter. We then present in Section 5.2 the main contributions of this chapter, including a general minimax formulation for the OT problem with an arbitrary convex compact set of cost matrices, and discuss an optimization procedure used to solve it as well as its theoretical guarantees. We further proceed by considering the important special cases of the previously introduced problem and showing their relationship to other works on the subject. Finally, in Section 5.3, we present an experimental evaluation of our approach for several considered use-cases.

5.1 Preliminary Knowledge

We now briefly recall the general notions related to the optimal transport problem (Section 2.1.2.4) and then introduce the most relevant works concerning its minimax formulation.

Optimal transport Optimal transport (OT) can be seen as the search for a transportation plan that moves (transports) a probability measure $\mathcal{D}$ onto another measure $\mathcal{D}'$ with a minimum cost measured by some function $c : \mathbb{X} \times \mathbb{X}' \to \mathbb{R}_+$, where $\mathbb{X}$ and $\mathbb{X}'$ are some complete metric spaces that, in most applications, are taken to be Euclidean spaces. More formally, the Kantorovich (Kantorovich, 1942) formulation of OT seeks for an optimal coupling $\mathcal{P}$ having marginals $\mathcal{D}$ and $\mathcal{D}'$, which minimizes the following quantity:

$$W_c(\mathcal{D}, \mathcal{D}') = \inf_{\mathcal{P} \in \Pi(\mathcal{D}, \mathcal{D}')} \mathop{\mathbb{E}}_{(\mathbf{x}, \mathbf{x}') \sim \mathcal{P}} \left[c(\mathbf{x}, \mathbf{x}') \right], \tag{5.1}$$

where $c(\mathbf{x}, \mathbf{x}')$ is the cost of moving $\mathbf{x}$ to $\mathbf{x}'$ (drawn from distributions $\mathcal{D}$ and $\mathcal{D}'$, respectively), and $\Pi(\mathcal{D}, \mathcal{D}')$ is defined as in Definition 2.1.10. When c is the squared Euclidean distance, we write W_2^2. In the discrete version of the problem, *i.e.* when $\mathcal{D}$ and $\mathcal{D}'$ are defined as empirical measures supported on vectors $\{\mathbf{x}_i\}_{i=1}^m$, $\{\mathbf{x}'_j\}_{j=1}^{m'}$ in $\mathbb{R}^n$ with probability

vectors $\mathbf{r} \in \Delta_m$ and $\mathbf{c} \in \Delta_{m'}$, the previous problem can be expressed as follows:

$$\mathbf{P}^* \in \underset{\mathbf{P} \in \Pi(\mathbf{r},\mathbf{c})}{\arg\min} \langle \mathbf{P}, \mathbf{C} \rangle \,, \tag{5.2}$$

where $\mathbf{C} \in \mathbb{R}^{m \times m'}$ is a cost matrix defined by $(\mathbf{C})_{ij} := c(\mathbf{x}_i, \mathbf{x}'_j)$, representing the pairwise costs of transporting $\mathbf{x}_i$ to $\mathbf{x}'_j$, and $\mathbf{P}$ is a matrix of size $m \times m'$ belonging to the transportation polytope $\Pi(\mathbf{r}, \mathbf{c})$ (also called Birkhoff polytope for $m = m'$) defined as[4]

$$\Pi(\mathbf{r}, \mathbf{c}) := \{\mathbf{P} \in \mathbb{R}_+^{m \times m'}; \mathbf{P}\mathbf{1}_{m'} = \mathbf{c}; \mathbf{P}^T\mathbf{1}_m = \mathbf{r}\}.$$

Remark In what follows, we denote by Π the sets $\Pi(\mathcal{D}, \mathcal{D}')$ and $\Pi(\mathbf{r}, \mathbf{c})$ respectively in the continuous and discrete cases, where the distinction between the two depends on the context.

Note that Problem (5.2) is a linear program (LP), but its dimensions scale quadratically with the size of the sample. Alternatively, one can consider a regularized version of the problem (Cuturi, 2013), which has the extra benefit of being faster to solve.

Minimax OT Two other studies considered the minimax formulation of the OT problem in a setting similar to ours. In the first one, Paty and Cuturi (2019) showed that one can see the OT problem, with c taken to be the squared Euclidean distance, as a trace minimization problem of the second-order displacement matrix defined for any $\mathcal{P} \in \Pi$. Specifically, one has:

$$\begin{aligned} &W_2^2(\mathcal{D}, \mathcal{D}') = \min_{\mathcal{P} \in \Pi} \mathrm{Tr}(\mathbf{V}_{\mathcal{P}}), \\ &\text{with} \quad \mathbf{V}_{\mathcal{P}} := \mathop{\mathbb{E}}_{(\mathbf{x}, \mathbf{x}') \sim \mathcal{P}} \left[(\mathbf{x} - \mathbf{x}')(\mathbf{x} - \mathbf{x}')^T \right]. \end{aligned}$$

Paty and Cuturi (2019) further used this expression of the 2-Wasserstein distance to introduce the Subspace Robust Wasserstein (SRW) distance as follows:

$$\mathcal{S}_k^2(\mathcal{D}, \mathcal{D}') := \min_{\mathcal{P} \in \Pi} \max_{\substack{\mathbf{0} \preccurlyeq \mathbf{M} \preccurlyeq \mathbf{I_n} \\ \mathrm{Tr}\{\mathbf{M}\} = k}} \langle \mathbf{V}_{\mathcal{P}}, \mathbf{M} \rangle = \min_{\mathcal{P} \in \Pi} \sum_{i=1}^{k} \lambda_i(\mathbf{V}_{\mathcal{P}}), \tag{5.3}$$

where $\preccurlyeq$ denotes the Loewener order for PSD matrices[5] and $\{\lambda_i\}_{i=1}^k$ are the $k \leq n$ largest eigenvalues of $\mathbf{V}_{\mathcal{P}}$. Note that considering only the maximization over the k largest eigenvalues allows learning a cost matrix of a reduced rank. Thus, it tackles the curse of dimensionality issue of calculating the Wasserstein distance for high-dimensional data.

A somehow different way of using the minimax formulation of OT was proposed in Alvarez-Melis et al. (2018) for c taken to be a submodular function $F : 2^V \to \mathbb{R}$ with $\mathbb{V}$ denoting a certain set of available items. In this case, taking the Lovász extension f of F leads to the following optimization problem:

$$\mathrm{StrOT}(\mathcal{D}, \mathcal{D}') := \min_{\mathbf{P} \in \Pi} \max_{\mathbf{C} \in \mathfrak{B}_F} \langle \mathbf{P}, \mathbf{C} \rangle \,,$$

where $\mathfrak{B}_F$ is the base polytope of F defined as $\mathfrak{B}_F = \{y \in \mathbb{R}^{|\mathbb{V}|} | y(\mathbb{V}) = F(\mathbb{V}); y(S) \leq F(S), \forall S \subseteq \mathbb{V}\}$. When F is a modular function, the size of $\mathfrak{B}_F$ is 1 thus recovering the original OT problem.

[4] This definition is more general than Definition 2.1.10, as it takes non uniform weights over the source and target samples into account.

[5] The Loewener order is defined for two PSD matrices $\mathbf{A}$ and $\mathbf{B}$ by: $\mathbf{A} \preccurlyeq \mathbf{B} \Leftrightarrow \mathbf{B} - \mathbf{A}$ is PSD.

Other related work Three other papers presented an OT-based minimax formulation distantly related to ours. In Genevay et al. (2018), the authors studied a generative model that uses Sinkhorn divergence as a fitting criterion and proposed to learn a cost function in this framework. Their problem is intrinsically different from ours as we do not consider the density fitting problem where one optimizes the parameters of the fitted distribution. On the other hand, in Li et al. (2019), the authors reduced the regularized OT formulation with relaxed marginal constraints into a minimax problem. Their formulation, however, is also different from ours as it does not seek to learn a cost matrix. Finally, the line of work on the Wasserstein distributionally robust optimization (Kuhn et al., 2019) is also very dissimilar from our contribution presented in the current chapter as this latter considers finding the best estimator of a density from a Wasserstein ball of a certain radius.

We now proceed to the presentation of our contributions.

5.2 Robust OT with a Convex Set of Cost Functions

Below, we formulate the general robust OT problem and highlight its properties in several cases of interest. We further propose and theoretically analyze a general algorithm that can be used to solve it.

5.2.1 Problem Formulation

Let $\mathfrak{G}$ (as in "Ground") be an arbitrary set of cost functions defined over $\mathbb{X} \times \mathbb{X}'$. This set may represent, for instance, a convex combination of cost function candidates provided by several experts, or it can be described by an infinite set of parameters. We impose no particular constraints on the cost functions belonging to $\mathfrak{G}$ as long as the corresponding Kantorovich problems admit a solution. We now consider the following minimax problem:

$$\mathrm{RKP}(\Pi, \mathfrak{G}) = \min_{\mathcal{P} \in \Pi} \max_{c \in \mathfrak{G}} \mathop{\mathbb{E}}_{(\mathbf{x}, \mathbf{x}') \sim \mathcal{P}} \left[c(\mathbf{x}, \mathbf{x}') \right], \tag{5.4}$$

where we look for a coupling $\mathcal{P}^*$ that is robust to the choice of a cost function $c \in \mathfrak{G}$, by considering the worst achievable transportation cost. A game-theoretic interpretation of this formulation is to consider two players, where Player 1 aims at aligning the two distributions by picking a coupling $\mathcal{P}$, while Player 2 resists to it by choosing the cost function c from the set of admissible costs $\mathfrak{G}$.

We denote the value at the solution of this problem by $\mathrm{RKP}(\Pi, \mathfrak{G})$, where RKP stands for robust Kantorovich problem. We abuse the notation and use $\mathrm{RKP}(\mathfrak{P}, \mathfrak{G})$ for any set $\mathfrak{P} \subseteq \Pi$ (even non convex) to denote $\mathrm{RKP}(\mathrm{conv}(\mathfrak{P}), \mathfrak{G})$, *i.e.* solving for $\mathcal{P} \in \mathrm{Conv}\,(\mathfrak{P})$, where $\mathrm{Conv}\,(\cdot)$ denotes the convex hull. We also extend the notation W_c, presented before by defining $W_{\mathfrak{G}}(\mathcal{D}, \mathcal{D}') := \mathrm{RKP}(\Pi, \mathfrak{G})$.

5.2.2 Choice of $\mathfrak{G}$

Below, we consider two possible choices for the convex set $\mathfrak{G}$. First, we study the infinite family of squared Mahalanobis distance cost matrices widely used in the metric learning literature (Section 1.4.4.2). Second, we consider a convex hull of a finite family of cost functions as in the example given above.

5.2.2.1 Infinite Family of Mahalanobis Distances

For any matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, we recall the definition of its Schatten $p-$norm as the $p-$norm (Definition A.1.5) of the vector constituted of its singular values, *i.e.*

$$\|\mathbf{M}\|_p^p = \sum_{1 \leq i \leq n} \sigma_i^p(\mathbf{M}),$$

where $p \in [1, +\infty]$ and $\{\sigma_i(\mathbf{M})\}$ are $\mathbf{M}$'s singular values. In particular, if $\mathbf{M} \in \mathbb{S}_+^{n\times n}$, then $\|\mathbf{M}\|_p = \mathrm{Tr}(\mathbf{M}^p)^{\frac{1}{p}}$. We also recall that the dual of the Schatten $p-$norm is the $q-$norm with q equal to $\frac{p}{p-1}$ if $p > 1$, to ∞ if $p = 1$ and to 1 if $p = \infty$ (Magnus, 1987) (analogous to the result on $p-$norms from Proposition A.1.2).

We now define $\mathfrak{G}$ as a family of Mahalanobis cost functions, parametrized by bounded matrices $\mathbf{M}$:

$$\mathfrak{G} = \{c^{\mathbf{M}} : (\mathbf{x}, \mathbf{x}') \mapsto (\mathbf{x} - \mathbf{x}')^T \mathbf{M}(\mathbf{x} - \mathbf{x}'); \|\mathbf{M}\|_p \leq 1\}. \tag{5.5}$$

We can now state the following proposition.

Proposition 5.2.1. *Let $\mathfrak{G}$ be defined as in (5.5) for matrices $\mathbf{M} \in \mathbb{S}_+^{n\times n}$. Then, $\mathfrak{G}$ is a convex compact set of cost functions, and for any $p, q \in [1, +\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$ the following holds:*

1. $\mathrm{RKP}(\Pi, \mathfrak{G}) = \min_{\mathcal{P}\in\Pi} \|\mathbf{V}_\mathcal{P}\|_q$. *In particular, we have:*

$$\mathrm{RKP}(\Pi, \mathfrak{G}) = \begin{cases} W_2^2(\mathcal{D}, \mathcal{D}'), & \text{if } q = 1, \\ \mathcal{S}_1^2(\mathcal{D}, \mathcal{D}'), & \text{if } q = \infty. \end{cases}$$

2. *For any $\mathcal{P} \in \Pi$, $\|\mathbf{M}^*\|_p = 1$ and*

$$\mathbf{M}^* = \underset{\substack{\mathbf{M}\in\mathbb{S}_+^{n\times n} \\ \|\mathbf{M}\|_p\leq 1}}{\arg\max} \langle \mathbf{V}_\mathcal{P}, \mathbf{M}\rangle = \left(\frac{\mathbf{V}_\mathcal{P}}{\|\mathbf{V}_\mathcal{P}\|_q}\right)^{\frac{q}{p}}.$$

In particular, for $p = 2$, one does not need to impose the PSD condition on $\mathbf{M}$, i.e.

$$\mathbf{M}^* = \underset{\|\mathbf{M}\|_2\leq 1}{\arg\max} \langle \mathbf{V}_\mathcal{P}, \mathbf{M}\rangle = \frac{\mathbf{V}_\mathcal{P}}{\|\mathbf{V}_\mathcal{P}\|_2}.$$

Proof idea. $\mathfrak{G}$ is the image of a convex compact set of $\mathbb{R}^{n\times n}$ by a linear mapping with a finite dimensional domain, hence it is convex and compact. Point 1 is a consequence of the equality case of Hölder's inequality, the positive semi-definiteness of matrix $\mathbf{V}_\mathcal{P}$ and the fact that the Schatten $p-$norm is the classic $p-$norm for the vector of a matrix's singular values, which tends to the $\infty-$norm as $q \to \infty$. The second point is a direct consequence of the equality case of Hölder's inequality for Schatten $p-$norms (Magnus, 1987) using the fact that $\mathbf{V}_\mathcal{P}$ is PSD. The last particular case is an application of the Cauchy-Schwarz inequality. □

This theorem highlights several novel insights. First, it provides a different point of view for a general minimax OT problem with the infinite family of Mahalanobis distances. In particular, it shows that the original OT problem can be seen as a minimax problem when one takes the least restrictive infinity norm for the bound on the matrix parameterizing the Mahalanobis distance, while SRW with $k = 1$ corresponds to the case of the $1-$norm[6]. This observation is illustrated in Figure 5.1 where we smoothly interpolate between the two boundary cases by solving Problem (5.4) with intermediate values of q. We note that such an interpolation may have interesting implications in practice when one seeks for an explicit control between the original and the minimax OT problems. Second, the optimal expression for $\mathbf{M}^*$ shows that it is proportional to $\mathbf{V}_\mathcal{P}$ and if this latter

[6]Other values of k for SRW are also covered when using a truncated Schatten $p-$norm.

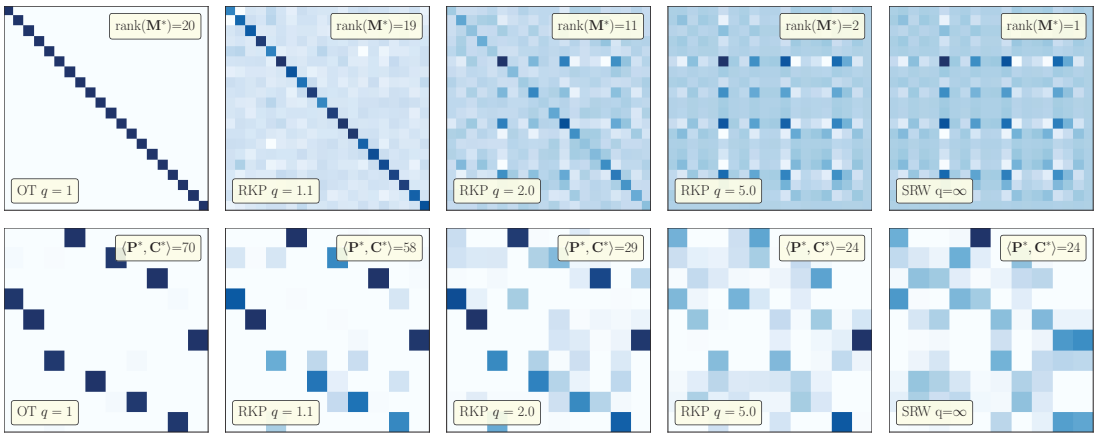


Figure 5.1: Interpolation between OT and $\text{SRW}_{k=1}$ on a binary toy classification problem with each class consisting of 5 points sampled from Gaussians centered on the edge of a 10-dimensional hypercube with $\sigma = 1$ with 10 additional random noise features. The transport is computed between the 2 classes using the setting of Proposition 5.2.1 with $q \in \{1, 1.1, 2, 5, \infty\}$. (top row) Mahalanobis matrices $\mathbf{M}^*$ and their rank; (bottom row)

captures the displacement in lower dimensions, then $\mathbf{M}^*$ is expected to do so too. In fact, in the discrete case we have:

$$\mathbf{V}_{\mathbf{P}^*} = \sum_{i=1}^{m}\sum_{j=1}^{m'}(\mathbf{P}^*)_{ij}(\mathbf{x}_i - \mathbf{x}'_j)(\mathbf{x}_i - \mathbf{x}'_j)^T,$$

where $\mathbf{P}^*$ is the optimal transport matrix. $\mathbf{M}^*$ is then a linear combination of $(\mathbf{x}_i - \mathbf{x}'_j)(\mathbf{x}_i - \mathbf{x}'_j)^T$, where $\{i, j\}$ are indices for which $(\mathbf{P}^*)_{ij} > 0$, making its image included in the span of $\{(\mathbf{x}_i - \mathbf{x}'_j); (\mathbf{P}^*)_{ij} > 0\}$, *i.e.* the span of displacement directions. This intuition is confirmed in our experiments (Figure 5.6) where we show that even without the rank constraint, solving Problem (5.4) results in a matrix of a reduced rank. The last point of the previous proposition shows that the case $p = 2$ (Frobenius norm) can be very convenient in practice as PSD constraints, needed for $p \neq 2$, increase considerably the computational burden of any optimization problem, yet they are necessary for the obtained cost function to be a true pseudo-metric.

To conclude the theoretical analysis of the considered case for the minimax problem, we establish a general bound on $\text{RKP}(\Pi, \mathfrak{G})$ in terms of the original 2-Wasserstein distance.

Corollary 5.2.1. *With the assumptions from Proposition 5.2.1, the following inequality holds for any $p \in [1, +\infty]$:*

$$\frac{1}{d^{\frac{1}{p}}} W_2^2(\mathcal{D}, \mathcal{D}') \leq W_{\mathfrak{G}}(\mathcal{D}, \mathcal{D}') \leq W_2^2(\mathcal{D}, \mathcal{D}').$$

Proof. Let $\mathcal{P} \in \Pi$. We have for any $p \geq 1$, $\|\mathbf{V}_{\mathcal{P}}\|_p \leq \|\mathbf{V}_{\mathcal{P}}\|_1$ by the monotonicity of Schatten $p-$norm, and we have $\min_{\mathcal{P}\in\Pi} \|\mathbf{V}_{\mathcal{P}}\|_1 = W_2^2(\mathcal{D}, \mathcal{D}')$ (Proposition 5.2.1). Taking the infimum over $\mathcal{P} \in \Pi$ yields the r.h.s. inequality in (5.2.1). To obtain the left hand side, notice that $\mathbf{A} := d^{-\frac{1}{p}} I_d$ verifies $\|\mathbf{A}\|_p \leq 1$, and that $\mathbf{A}$ is PSD, so that $c^{\mathbf{A}} \in \mathfrak{G}$. Thus,

$$\min_{\mathcal{P}\in\Pi} \langle \mathbf{V}_{\mathcal{P}}, \mathbf{A} \rangle = \min_{\mathcal{P}\in\Pi} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}} \left[c^{\mathbf{A}}(\mathbf{x}, \mathbf{x}')\right] \leq \min_{\mathcal{P}\in\Pi} \max_{c\in\mathfrak{G}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}} \left[c(\mathbf{x}, \mathbf{x}')\right] = W_{\mathfrak{G}}(\mathcal{D}, \mathcal{D}').$$

Finally, notice that the left hand side in the previous inequality equals $\frac{1}{d^{\frac{1}{p}}} W_2^2(\mathcal{D}, \mathcal{D}')$ by $\mathbf{A}$'s expression, thus concluding the proof. □

Note that compared to a similar bound given in Paty and Cuturi (2019, Proposition 2) for the SRW distance, our result does not involve the k term on the left-hand side as we do not impose any explicit constraint on the rank of $\mathbf{M}$.

5.2.2.2 Finite Set of Cost Functions

Let $\{c_1, ..., c_K\}$ denote a family of candidate cost functions, and let $\mathfrak{G} = \text{Conv}(\{c_1, ..., c_K\})$ implying that $\mathfrak{G}$ is a convex compact space as it is the convex combination of a finite set, or more shortly a polytope. As mentioned in Section 2, the optimization of the OT problem with a submodular function F taken as a cost function can be equivalently seen as a minimax OT problem of the following form:

$$\min_{\mathbf{P}\in\Pi} \max_{\mathbf{C}\in\mathfrak{B}_F} \langle \mathbf{P}, \mathbf{C} \rangle ,$$

where $\mathfrak{B}_F$ is the base polytope of F. We note that the number of vertices of $\mathfrak{B}_F$ is finite and thus one can show that the StrOT distance is a particular case of Problem (5.4) when $\mathfrak{G}$ is a finite set of cost functions, *i.e.*

$$\text{RKP}(\Pi, \mathfrak{B}_F) = \text{StrOT}(\mathcal{D}, \mathcal{D}').$$

This result establishes the link between our general formulation and that considered in Alvarez-Melis et al. (2018).

A second case where $\mathfrak{G}$ is finite was already introduced in the domain adaptation context in Chapter 4, more precisely in Equation (4.28). In this latter's r.h.s., the second term is given by the following optimization problem:

$$\min_{\mathcal{P}\in\Pi} \left\| \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}} \left[\left| (\mathbf{x}\mathbf{x}^T - \mathbf{x}'\mathbf{x}'^T)\mathbf{w} \right| \right] \right\|_\infty , \tag{5.6}$$

which is an instance of the RKP problem for the cost set

$$\mathfrak{G} := \text{Conv}(\{c_1, \cdots, c_n\}) \quad \text{with} \quad c_k : (\mathbf{x}, \mathbf{x}') \mapsto \left| \mathbf{e}_k^T (\mathbf{x}\mathbf{x}^T - \mathbf{x}'\mathbf{x}'^T)\mathbf{w} \right|, \tag{5.7}$$

where $\mathbf{e}_k$ denotes the k^{th} vector of $\mathbb{R}^n$'s canonical basis.

5.2.3 Proposed Optimization Strategy

We now propose a general solution for optimizing Problem (5.4) in the discrete case where $\mathbb{X}$ and $\mathbb{X}'$ are identified respectively with finite sets $\{\mathbf{x}_i\}_{i=1}^m$ and $\{\mathbf{x}'_j\}_{j=1}^{m'}$, while $\mathfrak{G}$ is identified with an arbitrary convex set of cost matrices with entries $(\mathbf{C})_{ij} = c(\mathbf{x}_i, \mathbf{x}'_j)$. Since $\mathbb{X}$ and $\mathbb{X}'$ are finite, hence bounded, all results from Section 5.2.2 hold in the discrete case.

We propose to adapt the cutting set method presented in Mutapcic and Boyd (2009) for robust optimization to Problem (5.4) that allows us to cover both unregularized and regularized minimax OT problems. In a nutshell, this method consists in alternating between solving a worst-case problem and the corresponding sampled robust minimization problem w.r.t. a set of constraints that grows linearly with iterations and requires for optimized functions to be convex only. In application to Problem (5.4), the high level idea of the proposed algorithm thus would be to solve the maximization problem over $\mathfrak{G}$ w.r.t. a set $\mathfrak{P} \subset \Pi$ and add one transportation matrix to $\mathfrak{P}$ at each iteration. The implementation of this idea, however, is not straightforward and requires two obstacles to be addressed. First, the original algorithm presented by the authors allows to solve a minimax problem of the form

$$\min_{\mathbf{C}\in\mathfrak{G}} \max_{\mathbf{P}\in\Pi} = - \max_{\mathbf{C}\in\mathfrak{G}} \min_{\mathbf{P}\in\Pi}$$

and thus requires from us to prove

$$\min_{\mathbf{P}\in\Pi} \max_{\mathbf{C}\in\mathfrak{G}} = \max_{\mathbf{C}\in\mathfrak{G}} \min_{\mathbf{P}\in\Pi}$$

in order to apply it. Second, and similar to the projected supergradient algorithm proposed for SRW, the authors of Mutapcic and Boyd (2009) disregard the optimal solution for the variable over which the minimization is performed, *i.e.* $\mathbf{P}^*$ in our case, and provide a solution for $\mathbf{C}^*$ only. To address these issues, we now present the following result.

Proposition 5.2.2. *Let $\mathfrak{P}$ be a finite subset of Π. Then, the following holds:*

1. $\mathrm{RKP}(\mathfrak{P}, \mathfrak{G}) := \mathrm{RKP}(\mathrm{Conv}(\mathfrak{P}), \mathfrak{G})$ *has a saddle point* $(\mathbf{P}^*, \mathbf{C}^*)$ *verifying:*

$$\langle \mathbf{P}^*, \mathbf{C}^* \rangle = \min_{\mathbf{P} \in \mathrm{Conv}(\mathfrak{P})} \max_{\mathbf{C} \in \mathfrak{G}} \langle \mathbf{P}, \mathbf{C} \rangle = \max_{\mathbf{C} \in \mathfrak{G}} \min_{\mathbf{P} \in \mathfrak{P}} \langle \mathbf{P}, \mathbf{C} \rangle . \tag{5.8}$$

2. $\mathrm{RKP}(\mathfrak{P}, \mathfrak{G})$ *is equivalent to*

$$\begin{aligned} \mathbf{C}^* \in \underset{\mathbf{C} \in \mathfrak{G}, \mu \geq 0}{\arg\max} \ & \mu \\ s.t. \quad \langle \mathbf{P}, \mathbf{C} \rangle \geq \ & \mu, \ \forall \mathbf{P} \in \mathfrak{P}. \end{aligned} \tag{5.9}$$

3. $\mathbf{P}^* = \sum_{l=1}^{|\mathfrak{P}|} q_l \mathbf{P}_l$, *where* $\mathfrak{Q} = \{q_l\}_{l=1}^{|\mathfrak{P}|}, \sum_i q_i = 1$, *are dual variables of Problem* (5.9).

Proof idea. Point 1 is an application of Sion's minimax theorem (Sion, 1958), which holds due to the bilinearity of the inner product, and to the convexity and compactness of both $\mathfrak{P}$ and $\mathfrak{G}$. Point 2 is a reformulation of the r.h.s. of Equation (5.8). As for the last point, we first express the Lagrange dual of Problem (5.9). Then, using the strong Lagrange duality, we prove $\mathbf{P}^*$'s expression as a function of the dual variables in $\mathfrak{Q}$. □

Several remarks are in order here. First, we note that solving Problem (5.9) directly for $\mathfrak{P} = \Pi$ is intractable in practice for sufficiently large m and m' as the number of constraints (size of $\mathfrak{P}$) grows extremely fast with the number of points (e.g., equal to $m'!$ for $m = m'$). This motivates the use of the cutting plane algorithm that gradually increases the size of the set $\mathfrak{P}$ with iterations and allows to solve intermediate problems with a reduced number of constraints efficiently. Second, the proposition is valid for any finite subset $\mathfrak{P}$ of Π so that 1) solving $\mathrm{RKP}(\Pi, \mathfrak{G})$ can be done by setting $\mathfrak{P}$ to the set of vertices of Π and 2) solving the regularized minimax formulation with added convex regularizer on $\mathbf{P}$ is covered by considering $\mathrm{RKP}(\tilde{\Pi}, \mathfrak{G})$, where $\tilde{\Pi}$ is a convex compact subset of Π.

To see the latter point, we note that a regularized OT problem has the following form:

$$\min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C} \rangle + \lambda R(\mathbf{P}) \tag{5.10}$$

where R is a convex function representing the regularization. The previous problem is equivalent to solving the Kantorovich problem for $\mathbf{P} \in \tilde{\Pi} = \{\mathbf{P} \in \Pi; R(\mathbf{P}) \leq \eta\}$ (the equivalence is justified by the KKT conditions (Karush, 1939; Kuhn and Tucker, 1951)), which is a convex set.

Our final algorithm inspired by Mutapcic and Boyd (2009, Section 5.1) then boils down to alternately performing the following two steps for $t \in \{0, \ldots, T\}$:

Step 1. Find $\mathbf{C}_t$ solving (5.9) over $(\mathfrak{P}_t, \mathfrak{G})$, where $\mathfrak{P}_t$ is a finite subset of Π; let μ_t be the value at the solution.

Step 2. For a fixed matrix $\mathbf{C}_t$ obtained at Step 1, find $\mathbf{P}_t \in \arg\min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C}_t \rangle$.

Step 2 of each iteration can make use of any efficient algorithm for solving the classic unregularized optimal transport. Empirically, we observed that even the approximate solutions obtained by solving the entropy regularized formulation of the optimal transport problem ensure the convergence. We further use the constraint dropping strategy Mutapcic and Boyd (2009, Section 5.3.2) and provide a complete pseudo-code in Algorithm 1, where thresholds τ_1 and τ_2 respectively control the stopping criterion and the constraint elimination. The proposed algorithm is generic and can also be used to solve the problems underlying the SRW and StrOT distances seen previously. Moreover, it acts as a meta-algorithm by implicitly choosing (or learning depending on the construction of the set $\mathfrak{G}$)

Algorithm 1 Cutting set method for $\mathrm{RKP}(\Pi, \mathfrak{G})$ with constraint elimination

Input: T, $\mathfrak{G}$, $\mathfrak{P}_0 \subset \Pi$, τ_1, τ_2
$t, \nu_{-1} \leftarrow 0$
$\varepsilon, \mu_{-1} \leftarrow \infty$
while $t < T$ and $\varepsilon_t > \tau_1$ and $\frac{\mu_{t-1}-\mu_t}{\mu_{t-1}} > \tau_1^2$ **do**
  Solve (5.9) to obtain $(\mu_t, \mathbf{C}_t), \mathfrak{Q}$
  for j in $\{0, ..., |\mathfrak{P}_t| - 1\}$ **do**
    if $q_j \leq \tau_2$ **then**
      $\mathfrak{P}_t \leftarrow \mathfrak{P}_t \setminus \{\mathbf{P}_j\}$
      $\mathfrak{Q} \leftarrow \mathfrak{Q} \setminus \{\mathbf{q}_j\}$
    end if
  end for
  Find $\mathbf{P}_t \in \arg\min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C}_t \rangle$
  $\nu_t \leftarrow \max(\nu_{t-1}, \langle \mathbf{P}_t, \mathbf{C}_t \rangle)$
  $\varepsilon_t \leftarrow \mu_t - \nu_t$
  $\mathfrak{P}_{t+1} = \mathfrak{P}_t \cup \{\mathbf{P}_t\}$
  $t \leftarrow t + 1$
end while
return $\sum_{l=0}^{|\mathfrak{P}_t|-1} q_l \mathbf{P}_l$, $\mathbf{C}_t$

the "right" cost function. This differs from other existing methods on learning the cost matrix in the OT framework (Cuturi and Avis, 2014; Zhao and Zhou, 2018) that usually learn this latter using the *a priori* similarity between the histograms.

Finally, Algorithm 1 is guaranteed to converge in a finite number of iterations T, with the latter being upper-bounded thanks to the following proposition.

Proposition 5.2.3. *Let T be the number of iterations required by Algorithm 1 to reach error $\varepsilon_T \leq \tau_1$. Then*

$$T \leq \left(\frac{\mathrm{diam}_\infty(\mathfrak{G}) + \mathrm{RKP}(\mathfrak{P}_0, \mathfrak{G})}{2\tau_1} + 1 \right)^{\dim(\mathfrak{G})+1}$$

where

$$\mathrm{diam}_\infty(\mathfrak{G}) := \sup_{\substack{\mathbf{C}^1, \mathbf{C}^2 \in \mathfrak{G} \\ i,j}} \left| (\mathbf{C}^1)_{ij} - (\mathbf{C}^2)_{ij} \right|,$$

and $\dim(\mathfrak{G})$ *is the dimension of the affine hull of* $\mathfrak{G}$*. Also, we have*

$$0 \leq \mathrm{RKP}(\mathfrak{P}_t, \mathfrak{G}) - \mathrm{RKP}(\Pi, \mathfrak{G}) \leq \varepsilon_t.$$

Proof idea. We adapt the proof technique presented in (Mutapcic and Boyd, 2009, Section 5.2) to our case, after re-writing the r.h.s. of Equation (5.8) as

$$\min_{\mathbf{C}\in\mathfrak{G}} \max_{\mathbf{P}\in\mathfrak{P}_t} \left(- \langle \mathbf{P}, \mathbf{C} \rangle \right),$$

to make our problem coincide with the authors' formulation. □

This theorem offers interesting insights regarding the convergence speed of the proposed algorithm. First, it introduces the dependence of the latter on $\mathrm{diam}_\infty(\mathfrak{G})$, which can be interpreted as a degree of disagreement between the cost matrices in $\mathfrak{G}$ so that one may need more iterations to reach error ε when they disagree. Second, the presence of the value of the initial nominal problem $\mathrm{RKP}(\mathfrak{P}_0, \mathfrak{G})$ reflects the influence of the initialization $\mathfrak{P}_0$: the smaller is the value of $\mathrm{RKP}(\mathfrak{P}_0, \mathfrak{G})$, the less iterations are needed. Finally, when $\mathfrak{G}$ lies in a subspace of a much smaller dimension than $m \times m'$ (*i.e.* in the case of the

Mahalanobis distance, $\mathfrak{G}$ is the image of $n \times n$ matrices by a linear mapping, while for the finite number of matrices, $\dim(\mathfrak{G})$ is $\dim(\text{span}(\mathbf{C}_1, ..., \mathbf{C}_n)) - 1)$, the algorithm needs much less iterations as highlighted by the presence of $\dim(\mathfrak{G})$ in the exponent.

Comparison to other optimization strategies used in minimax OT

We recall that we would like to compute the saddle point of our problem, not only the transport plan that minimizes the function $\max_{\mathbf{C}\in\mathfrak{G}} \langle ., \mathbf{C}\rangle$, or the cost matrix that maximizes the function $\min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \cdot\rangle$. The latter is concave, but it is non differentiable for the unregularized Kantorovich problem, due to the transport polytope Π having a non smooth boundary. This lack of differentiability is highlighted in Paty and Cuturi (2019, Section 5.2), and is the reason why the authors consider regularized optimal transport in their Algorithm 2, allowing them to obtain the saddle point, *i.e.* both optimal cost matrix and transport plan.

As for StrOT (Alvarez-Melis et al., 2018), the authors find the saddle point in the case of unregularized optimal transport by applying the Saddle Point Mirror Prox algorithm (SP-MP, see for example Bubeck (2015, Section 5.2.3)). This being said, SP-MP requires the cost function of the min-max problem (an inner product in our case and in the case of Alvarez-Melis et al. (2018)) to be smooth, whereas the algorithm of Mutapcic and Boyd (2009) that we use, after adapting it to find the saddle point, works even when the the considered cost is non differentiable. We can imagine imposing some constraints on the cost matrix $\mathbf{C}$ by solving the following problem:

$$\min_{\mathbf{P}\in\Pi} \max_{\mathbf{C}\in\mathfrak{G}} \langle \mathbf{P}, \mathbf{C}\rangle + \Psi(\mathbf{C})$$

where Ψ is a non differentiable concave Lipschitz function. This is the case of $\Psi : \mathbf{C} \to -\|\mathbf{C} - \mathbf{C}_0\|_1$ for example, where $\mathbf{C}_0$ is some reference cost matrix. In this case, the algorithm from Mutapcic and Boyd (2009) is theoretically guaranteed to find the solution, whereas SP-MP is not.

5.2.4 Variations for Different Choices of $\mathfrak{G}$

Below, we express Problem (5.9) over $\mathfrak{P}_t \times \mathfrak{G}$ at step $t \geq 0$ of Algorithm 1, for both choices of $\mathfrak{G}$ considered in Section 5.2.2, in a more convenient way.

Proposition 5.2.4 (Finite set $\mathfrak{G}$)**.** *Let $\mathfrak{G} = \text{Conv}(\{\mathbf{C}_1, ..., \mathbf{C}_K\})$. Then, for $t \geq 0$, solving the problem given in* (5.9) *over $\mathfrak{P}_t \times \mathfrak{G}$ is equivalent to the following linear program*

$$\begin{aligned} \min_{\mathbf{p}\in\mathbb{R}_+^K} \quad & \mathbf{1}_K^T \mathbf{p} \\ s.t. \quad & \mathbf{G}\mathbf{p} \geq \mathbf{1}_{|\mathfrak{P}_t|}, \end{aligned} \tag{5.11}$$

where $\mathbf{G} \in \mathbb{R}^{|\mathfrak{P}_t|\times K}$ with $\mathbf{G}_{kl} = \langle \mathbf{P}_k, \mathbf{C}_l\rangle$. Moreover, the saddle point $(\mathbf{P}^*, \mathbf{C}^*)$ is given by

$$\mathbf{C}^* = \frac{\sum_{k=1}^K p_k^* \mathbf{C}_k}{\sum_{k=1}^K p_k^*}, \qquad \mathbf{P}^* = \frac{\sum_{l=1}^{|\mathfrak{P}_t|} q_l^* \mathbf{P}_l}{\sum_{l=1}^{|\mathfrak{P}_t|} q_l^*},$$

where $\mathbf{p}^* = (p_1, \cdots, p_K)$ and $\mathbf{q}^* = (q_1, \cdots, q_{|\mathfrak{P}_t|})$ are optimal solutions of (5.11) *and its dual.*

Proof idea. We first establish an intermediary result in Lemma D.1.1, giving Problem (5.9)'s dual, with $\mathbf{p}^{**} \in \Delta_K$ and $\mathbf{q}^{**} \in \Delta_{|\mathfrak{P}_t|}$ denoting their respective solutions. Then, by re-writing the probability simplex Δ_K as

$$\Delta_K = \left\{ \frac{\mathbf{p}}{\mathbf{1}_K^T \mathbf{p}}; \mathbf{p} \in \mathbb{R}_+^K \setminus \{0\} \right\}, \tag{5.12}$$

we establish that Problem (5.9) and its dual are respectively equivalent to Problem (5.11) and its dual. These latter have respective solutions $\mathbf{p}^* \in \mathbb{R}_+^K$ and $\mathbf{q}^* \in \mathbb{R}_+^{|\mathfrak{P}_t|}$, and taking into account the re-writing of the probability simplex in Equation (5.12), we obtain that

$$\mathbf{p}^{**} = \frac{\mathbf{p}^*}{\mathbf{1}_K^T \mathbf{p}^*} \quad \text{and} \quad \mathbf{q}^{**} = \frac{\mathbf{q}^*}{\mathbf{1}_{|\mathfrak{P}_t|}^T \mathbf{q}^*},$$

from which the expressions of $\mathbf{P}^*$ and $\mathbf{C}^*$ are deduced. □

For the case of the infinite family of Mahalanobis distances, we propose a more general result that considers the following set Mahalanobis distances translated by an a priori fixed cost matrix $\mathbf{C}$:

$$\mathfrak{G}_{\mathbf{C}} = \{\mathbf{C} + \mathbf{E}^{\mathbf{M}} \in \mathbb{R}^{m \times m'}; (\mathbf{E}^{\mathbf{M}})_{ij} = (\mathbf{x}_i - \mathbf{x}'_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}'_j); \mathbf{M} \in \mathbb{S}_+^{n \times n}; \|\mathbf{M}\|_p \leq r\}. \quad (5.13)$$

for an arbitrary radius $r > 0$.

Proposition 5.2.5 (Non centered family of Mahalanobis distances)**.** *For a fixed matrix* $\mathbf{C}$*, let* $\mathfrak{G}_{\mathbf{C}}$ *be defined as in* (5.13)*. Then, for* $t \geq 0$*, solving* (5.9) *over* $\mathfrak{P}_t \times \mathfrak{G}_{\mathbf{C}}$*, is equivalent to solving the following convex program*

$$\min_{\mathbf{P} \in \mathrm{Conv}(\mathfrak{P}_t)} r \|\mathbf{V}_{\mathbf{P}}\|_q + \langle \mathbf{P}, \mathbf{C} \rangle . \quad (5.14)$$

Moreover, if $\mathbf{P}^*$ *is an optimal solution of Problem* (5.14)*, then* $\mathbf{M}^*$ *is expressed as in Proposition 5.2.1 with probability distribution* $\mathcal{P}$ *replaced by matrix* $\mathbf{P}^*$*.*

Proof. We recall that

$$\mathfrak{G}_{\mathbf{C}} = \{\mathbf{C} + \mathbf{E}^{\mathbf{M}} \in \mathbb{R}^{m \times m'}; (\mathbf{E}^{\mathbf{M}})_{ij} = (\mathbf{x}_i - \mathbf{x}'_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}'_j); \mathbf{M} \in \mathbb{S}_+^{n \times n}; \|\mathbf{M}\|_p \leq r\}.$$

This set of cost matrices is the same as the one considered in Proposition 5.2.1, up to a translation by matrix $\mathbf{C}$. Hence, it is convex and compact.

By Proposition 5.2.2, solving Problem (5.9) is equivalent to solving

$$\min_{\mathbf{P} \in \mathrm{Conv}(\mathfrak{P}_t)} \max_{\mathbf{D} \in \mathfrak{G}_{\mathbf{C}}} \langle \mathbf{P}, \mathbf{D} \rangle$$

However for any matrix $\mathbf{P} \in \mathrm{Conv}(\mathfrak{P}_t)$, and for $r\mathbb{B}_{p+}^n = \{r\mathbf{M}, \mathbf{M} \in \mathbb{B}_{p+}^n\}$ ($\mathbb{B}_{p+}^n$ is defined as in the proof of Proposition 5.2.1), we have:

$$\begin{aligned}
&\max_{\mathbf{D} \in \mathfrak{G}_{\mathbf{C}}} \langle \mathbf{P}, \mathbf{D} \rangle \\
=&\max_{\mathbf{M} \in r\mathbb{B}_{p+}^n} \sum_{ij} (\mathbf{P})_{ij} ((\mathbf{x}_i - \mathbf{x}'_j) \mathbf{M} (\mathbf{x}_i - \mathbf{x}'_j) + (\mathbf{C})_{ij}) \\
=&\max_{\mathbf{M} \in \mathbb{B}_{p+}^n} r \sum_{ij} (\mathbf{P})_{ij} ((\mathbf{x}_i - \mathbf{x}'_j) \mathbf{M} (\mathbf{x}_i - \mathbf{x}'_j)) + \langle \mathbf{P}, \mathbf{C} \rangle \\
=&r \|\mathbf{V}_{\mathbf{P}}\|_q + \langle \mathbf{P}, \mathbf{C} \rangle
\end{aligned}$$

where in the last line, we used the developments done in Proposition 5.2.1, from which we also get the expression of $\mathbf{M}^*$. □

In the rest of the chapter, we consider the case of $p = 2$. We recall that by Proposition 5.2.1, $\mathbf{M}^*$ is PSD even without imposing such a constraint.

5.2.5 Towards a Notion of Stability for Cost Matrices

In this section, we define a new notion of OT stability for a cost matrix $\mathbf{C}$ based on the convex set $\mathfrak{G}_{\mathbf{C}}$ given by (5.13).

Definition 5.2.1. *For a cost matrix* $\mathbf{C}$ *and its associated convex set* $\mathfrak{G}_{\mathbf{C}}$ *introduced in* (5.13)*, and for some* $r > 0$*, we define the instability* $\mathrm{WS}_{\mathbf{C},r}$ *as follows:*

$$\mathrm{WS}_{\mathbf{C},r} := W_{\mathfrak{G}_{\mathbf{C}}}(\mathcal{D}, \mathcal{D}') - W_{\mathbf{C}}(\mathcal{D}, \mathcal{D}') = \min_{\mathbf{P}\in\Pi} \max_{\|\mathbf{M}\|\leq r} \left\langle \mathbf{P}, \mathbf{C} + \mathbf{E}^{\mathbf{M}} \right\rangle - \min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C} \rangle \,.$$

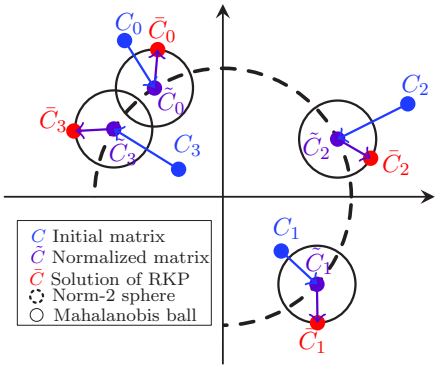


Figure 5.2: Illustration of the notion of matrix cost stability. Every matrix $\mathbf{C}_i$ is normalized so as to get a matrix $\tilde{\mathbf{C}}_i$ which lies on the norm-2 sphere. $\bar{\mathbf{C}}_i$ is the minimizer of Problem (5.8). The stability (Definition 5.2.1) comes from the difference of the cost transports induced by $\bar{\mathbf{C}}_i$ and $\tilde{\mathbf{C}}_i$.

Roughly speaking, Definition 5.2.1 tells us that the Wasserstein distance between $\mathcal{D}$ and $\mathcal{D}'$ associated with a stable cost matrix $\mathbf{C}$ should not differ much from the Wasserstein distance calculated based on the worst cost matrix in the neighborhood of $\mathbf{C}$. Note that the latter is defined as a Mahalanobis ball allowing us to define the stability of $\mathbf{C}$ w.r.t. the finite sets $\mathbb{X}$ and $\mathbb{X}'$. To be able to compare different instability values for a family of cost matrices $\{\mathbf{C}_i\}_{i=1}^K$, we normalize each $\mathbf{C}_i$ either by dividing its elements by its Frobenius norm or by the associated transport cost $W_{\mathbf{C}_i}(\mathcal{D}, \mathcal{D}')$. Figure 5.2 illustrates the intuition behind the notion of cost matrix stability where the Frobenius norm is used for the normalization.

5.3 Experiments

In this section, we first illustrate our algorithm's speed of convergence and compare it to solving the original LP Problem (5.9). Then, we reproduce a simulated problem from Paty and Cuturi (2019) to assess the algorithm's ability to correctly identify the subspace of lower dimensionality in which the transformation between the two samples lies. In what follows, we concentrate on comparing our approach with the authors' implementation of SRW while leaving aside the comparison with StrOT for which the implementation is not publicly available. The second part of our experiments is related to the notion of stability defined in Section 5.2.5. We first bring to light a correlation between the stability and the noise resistance of a cost matrix. Then, we show that selecting the most stable matrix allows to efficiently transport colors between two images in a color transfer task. The code for the different experiments is available on this link[7], and more details on the different experiments are provided in Section D.2.

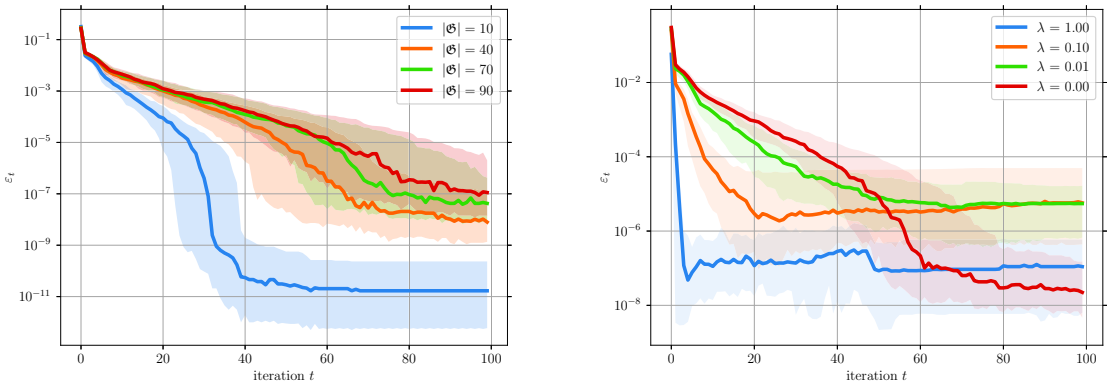
[7] `https://github.com/sofiendhouib/minimax_OT`.

5.3.1 Convergence and Execution Time

We consider the case where $\mathfrak{G}$ is the convex hull of a finite set of cost matrices. The convergence of Algorithm 1 is illustrated in Figure 5.3a (left) by plotting the evolution of the quantity $\varepsilon_t := \mu_t - \nu_t$ along the iterations for $|\mathfrak{G}| \in \{10, 40, 70, 90\}$. From this plot, we see that the convergence becomes slower as $|\mathfrak{G}|$ grows, which is expected because μ_t is the value at the solution of Problem (5.9) over $\text{Conv}(\mathfrak{P}_t) \times \mathfrak{G}$: as $\mathfrak{G}$ grows, minimizing a maximum over it becomes more difficult. Second, for $|\mathfrak{G}| = 10$, Algorithm 1 already achieves an error $\varepsilon_t \leq 10^{-10}$ after $t = 100$ iterations. This confirms that $\mathfrak{P}_t$ does not have to grow until it becomes the set of vertices of Π, as $|\mathfrak{P}_{100}| \leq |\mathfrak{P}_0| + 100 \ll m! = 100!$. We also test our algorithm with the entropic regularization of the transport matrix with $\lambda \in \{1, 0.1, 0.01, 0\}$ as regularization parameter, using Sinkhorn algorithm (Cuturi, 2013) for $|\mathfrak{G}| = 40$. For this setting, we initialize it with $\mathfrak{P}_0 = \{\frac{1}{mm'}\mathbf{1}_m\mathbf{1}_{m'}^T\}$ for any $\lambda > 0$, as this set $\mathfrak{P}_0$ is included in the feasible set of entropy-regularized transport (as suggested in the discussion of Proposition 5.2.2). Interestingly, we have noticed that the algorithm does not converge if $\mathfrak{P}_0$ is a subset of the vertices of transportation polytope Π in the regularized case. The results of this experiment are reported in Figure 5.3b (middle), where we observe the convergence even with the entropy regularization. Additionally, we note that due to the linearity of the mapping $\langle \cdot, \mathbf{P} \rangle$ for all $\mathbf{P} \in \Pi$, Problem (5.4) can be reformulated as the following LP:

$$\begin{aligned} \min_{\substack{\mathbf{P} \in \Pi \\ \eta \geq 0}} \quad & \eta, \\ \text{s.t.} \quad & \langle \mathbf{P}, \mathbf{C}_l \rangle \leq \eta \qquad \forall 1 \leq l \leq n. \end{aligned}$$

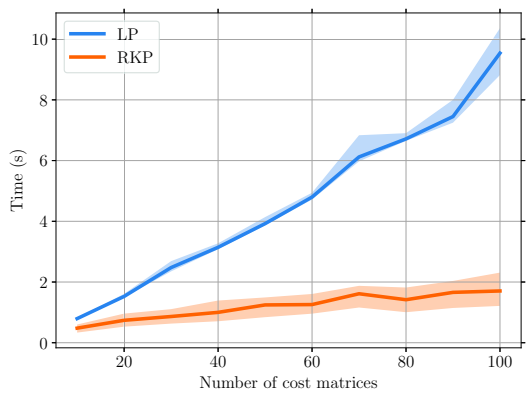
It turns out that this is nothing more than the dual of Problem (5.9). Under this formulation, solving $\text{RKP}(\Pi, \mathfrak{G})$ becomes tractable for $m = m' = 100$ and $|\mathfrak{G}| \in \{10, 20, ..., 90\}$, and allows us to compare between directly solving the LP problem and using our algorithm in Figure 5.4. As the number of candidate matrices grows, our algorithm becomes much more faster than solving the full LP problem Figure 5.4a. This is rather expected since at each iteration, it solves a linear program with much less constraints (the problem is restricted to $\text{Conv}(\mathfrak{P}_t) \times \mathfrak{G}$ instead of $\Pi \times \mathfrak{G}$) and it leverages efficient algorithms for solving problem (5.2). As for the value of the RKP at the solution, in spite of the fact that the LP approach providing a lower value, the relative difference w.r.t. to using Algorithm 1 is negligible with an order of magnitude of 10^{-7} (Figure 5.4b).



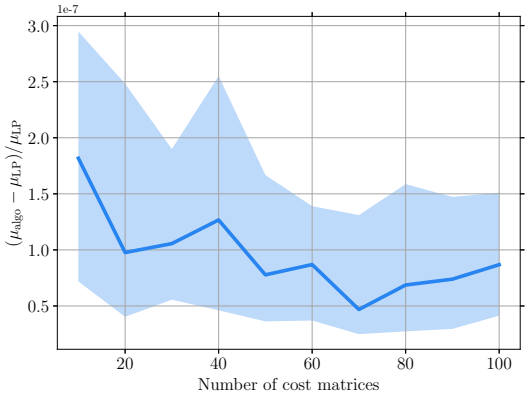
(a) For different numbers of candidate cost matrices.

(b) For different entropic regularization coefficients.

Figure 5.3: Evolution of error ε_t for 100 iterations and $m = m' = 100$. The experiment is repeated 100 times and we report the median along with the interquartile range.



(a) Execution time (in seconds) taken to solve the RKP problem.



(b) Relative difference between the values at the solution: $\frac{\mu_{\text{algo}} - \mu_{\text{LP}}}{\mu_{\text{LP}}}$

Figure 5.4: Comparing our algorithm to solving the original LP problem with $|\mathfrak{G}| \in \{10, 20, ..., 90\}$ and $m' = m = 100$. The experiment is repeated 100 times and we report the median along with the interquartile range.

5.3.2 Comparison to SRW

In this series of experiments, we consider the fragmented hypercube dataset studied in Paty and Cuturi (2019) and earlier in Forrow et al. (2019), and compare RKP to both the SRW and the Wasserstein distances. To proceed, let $\{\mathbf{e}_l\}_{1 \leq l \leq n}$ be the canonical basis of $\mathbb{R}^n$ and let $\mathbb{X} = \{\mathbf{x}_i\}_{i=1}^m$ and $\mathbb{X}' = \{\mathbf{x}'_j\}_{i=1}^{m'}$ be two finite sets drawn i.i.d. from the uniform distribution over the n-dimensional hypercube $\mathcal{U}([-1, 1]^n)$ and its pushforward distribution under the mapping $T : \mathbf{x} \mapsto \mathbf{x} + 2\,\mathrm{sgn}(\mathbf{x}) \odot (\sum_{i=1}^k \mathbf{e}_i)$, where $\odot$ denotes elementwise multiplication and $1 \leq k \leq n$, respectively. Therefore, by construction, there are k relevant features and $d - k$ features that contain no useful information. Depending on the choice of $\mathfrak{G}$, three cases of our algorithm are tested:

1. Squared Euclidean distance after projecting on all vectors of the canonical basis

$$\mathfrak{G} = \{\mathbf{C}_s \in \mathbb{R}^{m \times m'} | (\mathbf{C}_s)_{ij} = ((\mathbf{x}_i - \mathbf{x}'_j)^T \mathbf{e}_s)^2; 1 \leq s \leq n\}$$

2. Squared Euclidean distance after projecting on all combinations of two vectors of the canonical basis

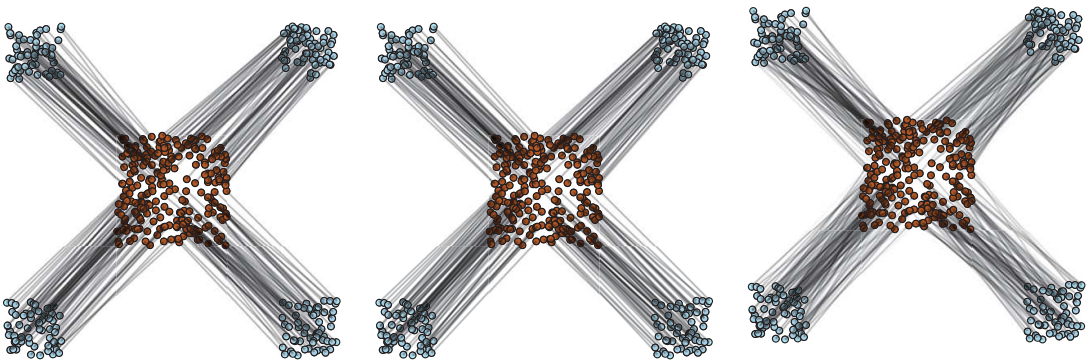
$$\mathfrak{G} = \{\mathbf{C}_{sl} \in \mathbb{R}^{m \times m'} | (\mathbf{C}_{sl})_{ij} = ((\mathbf{x}_i - \mathbf{x}'_j)^T (\mathbf{e}_s + \mathbf{e}_l))^2; 1 \leq s < l \leq n\}$$

3. The Mahalanobis unit ball with the Euclidean norm, centered at $\mathbf{0}$.

Figure 5.5 reproduces the experiments of Paty and Cuturi (2019) and shows that the original OT (Figure 5.5e) is sensitive to noise, while both SRW and RKP (for the 3 configurations considered) are able to recover the true pushforward transformation. However, while SRW requires a hyperparameter k to constrain the rank of the Mahalanobis matrix, our method is parameter-free since k is found automatically as illustrated in Figure 5.6. In this figure, we plot the eigenvalues of $\mathbf{M}^*$ for different values of k and observe that the eigengap between the k largest eigenvalues and the $(k+1)^{th}$ eigenvalue clearly reveals that $\mathrm{rank}(\mathbf{M}^*) = k$.

5.3.3 Stability and Noise Sensitivity

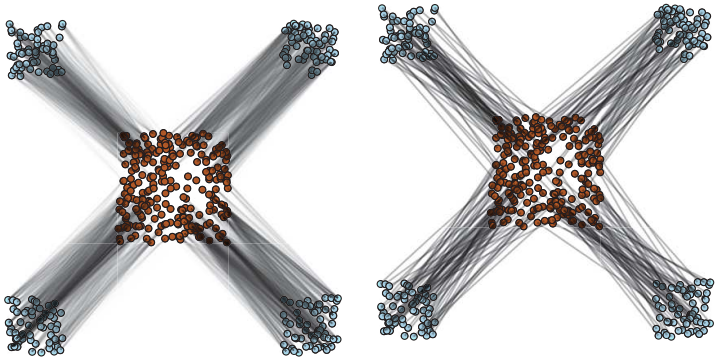
Below, we illustrate the correlation between the cost matrix stability and the sensitivity of the Wasserstein distance to the presence of noise, using both a toy and a real-world dataset.



(a) RKP: 1D projections.

(b) RKP: 2D projections.

(c) RKP: Mahalanobis case with the Euclidean norm.



(d) SRW with $k = 2$.

(e) OT with squared Euclidean distance.

Figure 5.5: Results obtained on the fragmented hypercube for $m = m' = 250$, $d = 30$ and $k = 2$. The lines indicate the connections between points according to the computed transport matrix $\mathbf{P}^*$, and their opacity increasing with the values of $\mathbf{P}^*$'s coefficients.

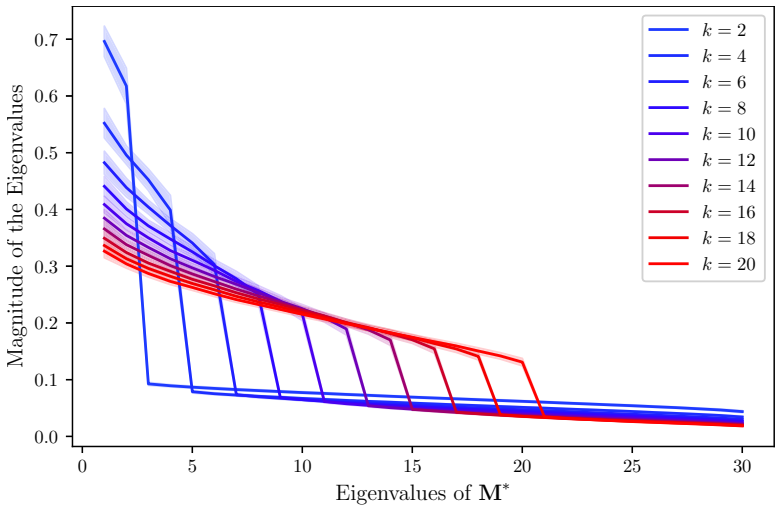


Figure 5.6: Sorted eigenvalues of $\mathbf{M}^*$ obtained using RKP averaged over 100 runs for different values of k reveals a phase transition between k dominant and the $k+1$ eigenvalues

The latter is composed of 100 zeros and 100 ones images coming from the MNIST dataset, after reducing its dimensionality to 10 with UMAP (McInnes et al., 2018). The former consists of 100 points drawn from two 10-dimensional Gaussian distributions centered at $\mathbf{0}_{10}$ and $3 \times \mathbf{1}_{10}$ respectively with unit variance. For both datasets, we generate a family of cost matrices $\{\mathbf{C}_i\}_{i=1}^{50}$ based on random Mahalanobis distances with different norms, normalize them so that their Frobenius norm equals to 1 and compute $\mathrm{WS}_{\mathbf{C}_i, r=0.01}$ from Definition 5.2.1 for all i. To introduce noise to each $\mathbf{C}_i$, we add a random Mahalanobis cost matrix $\mathbf{E}^{\mathbf{N}}$ with $\|\mathbf{N}\|_2 = r$ to it and compute the noise sensitivity defined as:

$$\mathrm{NS}_{\mathbf{C}_i} = \left| \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C}_i \rangle - \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C}_i + \mathbf{E}^{\mathbf{N}} \rangle \right|.$$

Note that we apply a Mahalanobis noise which has the advantage of taking into account the point distributions and can be applied on any matrix $\mathbf{C}_i$. Figure 5.7 presents the results of this experiment averaged over 200 runs and shows a clear correlation between the stability and noise sensitivity indicating that the most stable matrices are more noise tolerant. Other experiments on the MNIST dataset provided in Section D.2 show a similar behavior.

5.3.4 Color Transfer

In this last experiment, we show how we can benefit from the notion of stability to address a color transfer task where the goal is to transfer the colors from a blueish sky image to the reddish ocean image shown in Figure 5.9 (left). Here, we use OT between the sets of pixels in the RGB space extracted from both images. For the sake of efficiency, we consider only 200 pixels from each image and generalize the obtained OT mapping to the remaining pixels following the method detailed in Ferradans et al. (2014). As before, we use $\{\mathbf{C}_i\}_{i=1}^{50}$ as meaningful cost matrices and add 50 completely random matrices that are unrelated to the considered task. The results of this experiment given in Figure 5.9 show a significant gap in terms of stability between the Mahalanobis matrices (the first 50 matrices on the x-axis) and the random ones (the last 50). This tends to highlight the fact that the stability can be used as a criterion to select a good cost matrix, and therefore to induce a relevant Mahalanobis distance. This also holds in terms of visual perception as illustrated in Figure 5.9 (right). Even if the most stable matrix is visually very similar to the Euclidean one, a finer evaluation reveals more discontinuities in the center of the picture, on the water.

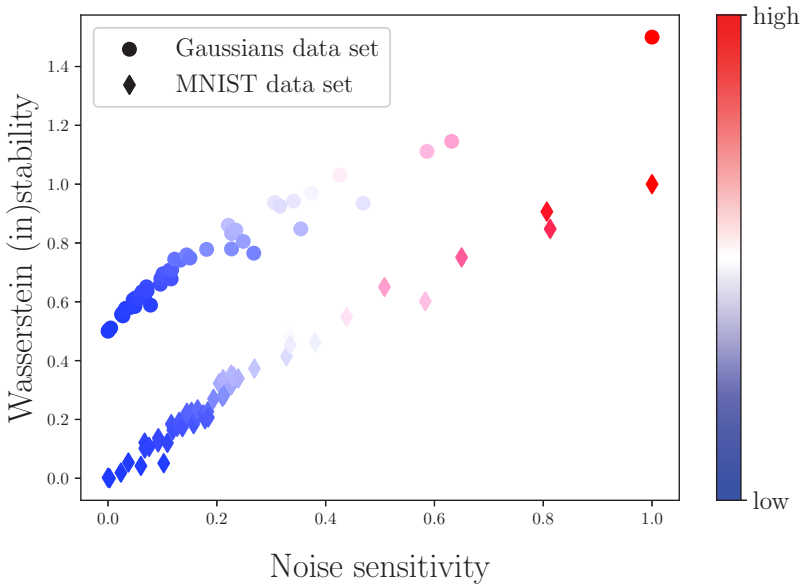


Figure 5.7: Correlation between the stability and the sensitivity to noise.

Figure 5.8: Source (ocean) and target (sky) images considered as probability distributions

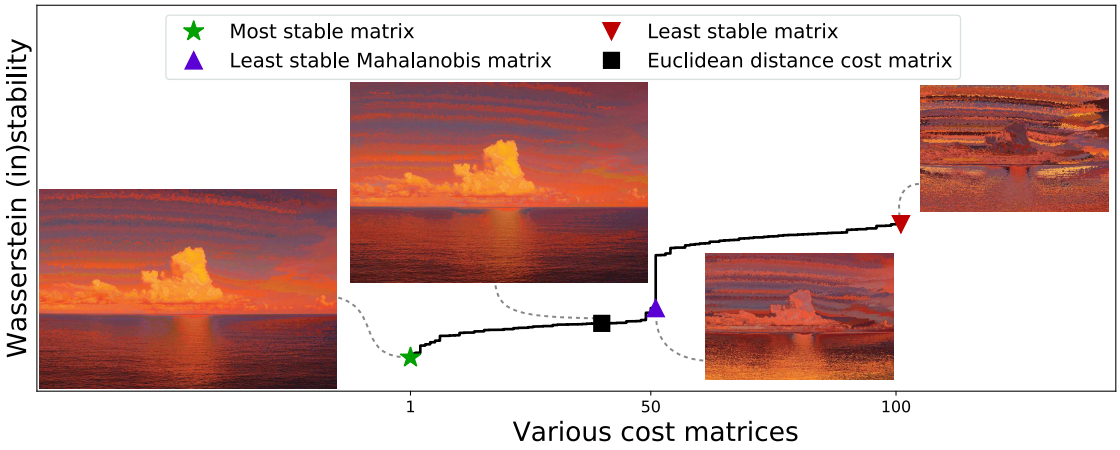


Figure 5.9: Cost matrices sorted by Wasserstein stability. The first 50 are Mahalanobis cost matrices, while the last 50 are random cost matrices.

5.4 Conclusion

In this chapter, we study a general formulation of the minimax OT problem that consists in optimizing over the coupling matrix w.r.t. the worst cost function from a convex set of

cost functions. When the latter is an infinite family of Mahalanobis distances, we highlight several novel properties of the considered problem and characterize the different features of its solutions. We further show how the underlying optimization problem can be solved in practice, using a variation of a cutting set algorithm with theoretical guarantees regarding its convergence speed. Finally, we define a new notion of stability for cost matrices in OT based on the studied minimax problem and reveal a correlation between this stability and the noise resistance of the matrices. This leads to a criterion that can be employed to select a relevant cost function that has been shown to be efficient on both toy and real-world data.

Concerning future perspectives for this work, the generality of the RKP formulation allows us to consider more sets of cost matrices that were not addressed in this chapter. First, one could generalize the case of squared Mahalanobis distances with a bounded norm (Section 5.2.2.1) to include truncated Schatten $p-$norms (Grone, 1980), where only the k largest singular values of a matrix are considered in the norm's definition. In particular, a comparison with SRW is worth an investigation. Second, the dual formulation presented in Proposition 5.2.5 suggests a generalization to the infinite-dimensional case using kernels and the representer theorem of Schölkopf et al. (2001). In this case, even if the compactness of $\mathfrak{G}$ is no longer verified due to the infinite dimension, its convexity is preserved and Sion's minimax theorem (Sion, 1958) still holds. Third, a deeper theoretical understanding of the empirically observed resistance to noise might be achieved via establishing convergence rates of the RKP towards its true distributions counterpart. Finally, a promising line of research consists in extending the idea of choosing a stable cost matrix from a finite set to an infinite and possibly continuous set of candidates.

Conclusion

Throughout this dissertation, we tackled the challenging problem of domain adaptation, and we provided contributions to it from three general viewpoints, namely, learning with similarity functions, solving a large-margin classification task, and using optimal transport and its variations to derive efficient algorithms.

Motivated by the lack of theoretical analysis of similarity-based learning when the training and testing data distributions change, our first contribution concerned the extension of the $(\epsilon, \gamma, \tau)-$good similarity functions theory to domain adaptation. We answered the question concerning the extent of goodness of a similarity function on the target domain in terms of its goodness on the source, a certain divergence measure between the two domains, and also between their potentially different associated landmark distributions. However, with hindsight, we proved the lack of interest in being constrained by the (ϵ, γ, τ) framework, from which we only kept the large-margin classification aspect as the letter is crucial in assessing the confidence a classifier accords to its predictions.

In our second contribution, we proved novel bounds on the margin violation risk for the target domain. Aside from generalizing previous results from the literature, these guarantees led us to defining a new adversarial variation of the optimal transport problem that is more attached to the considered hypothesis spaces, in contrast to previously proposed optimal transport-based domain adaptation approaches. Then, we derived a domain adaptation algorithm that takes the form of an adversarial, yet convex, optimization problem. After noticing the role of similarity induced spaces in verifying the assumption of a low ideal joint error, we demonstrated the interest of the algorithm on a simulated and a real-world high dimensional data set. The method's performance suggests that the attachment of our alignment term to the task at hand, via the considered hypothesis spaces, allows tackling the known limitations of optimal transport in high-dimensional settings.

Finally, with the possibility of generalizing the introduced optimal transport term to broader situations, we defined a new variation of the Monge-Kantorovich problem, called the Robust Kantorovich Problem (RKP), where the transport plan and the cost function are jointly learned in an adversarial fashion. We explored several particular cases of this formulation conditioned by the set of candidate ground cost functions and described how to solve their associated optimization problems. By means of empirical evaluations, we showed that our RKP formulation successfully captures displacements occurring in low-dimensional subspaces and allows us to define a notion of cost matrix stability that helps in picking a cost matrix from a predefined set of candidates.

Our contributions have several possible future research directions that we highlighted at the conclusions of their respective chapters. We now detail the ones we find to have significant potential for future work. For the first one of them, we note that our analysis of learning with $(\epsilon, \gamma, \tau)-$ good similarity functions, intended for learning a classifier, was proven to be redundant in the case of bilinear similarity functions with quadratic regularization. We would like to know when learning a similarity function may be beneficial without being equivalent to directly learning a classifier. A necessary condition, according to our analysis of the bilinear case, is to have a parameter matrix with a rank greater than 1. More generally, we think this condition must be verified for the similarity matrix (even in the non-bilinear case), such as for kernels. Hence, a thorough theoretical study

of $((\epsilon, \gamma, \tau))-$good similarity learning in connection with kernel learning methods is worth exploring. Second, our study of domain adaptation in the large-margin classification context encourages us to attempt to derive dimension-free generalization guarantees (Bartlett and Shawe-Taylor, 1999). Besides, due to the adversarial terms we introduce, the question of whether our adversarial terms may have advantages for deep neural networks arises naturally. In fact, theoretical results have already been used to inspire deep DA approaches, such as Ganin et al. (2016) and Zhang et al. (2019), with the latter considering classification margin. Of course, we need to generalize our results to the multi-class case for the generalization to deep networks to be impactful. Other nonlinear approaches it may be interesting to explore include boosting algorithms. In this case, whether it is possible to gradually construct the set of adversary classifiers, akin to iteratively constructing an optimal hypothesis as a combination of weak classifiers, remains an open question. In line with this, the adversarial optimal transport approach we studied in Chapter 5 deserves an analysis of its sample complexity, which is motivated by the empirically observed resistance to noise. This latter quality suggests a possible improvement of the mappings estimated from optimal transport plans (*e.g.* the barycentric mapping). Also, our empirical observations related to the cost function's stability motivate an in-depth theoretical analysis of this notion too. More generally, it would be interesting to go beyond the domain adaptation special case we considered for classification, *i.e.* single-sourced homogeneous unsupervised domain adaptation. A direct extension would be to include some target domain labels in a semi-supervised setting, whereas the study of the multi-source or the heterogeneous scenario would be even more challenging. Additionally, the similarity learning and optimal transport themes together lead us to think about the possible consideration of the Gromov-Wasserstein distance (Mémoli, 2011) between distributions, if we consider pairs of instances and adapt to them the results of Chapters 4 and 5. In such a case, computational considerations linked to the quadratic size of the set of pairs, as a function of the number of instances, are to be taken into account.

List of Publications

Submissions to journals

Sofien Dhouib and Ievgen Redko. A ridge regression approach for fast bilinear similarity learning with theoretical guarantees. Submitted to Machine Learning Journal 2020.

Part III

Appendices

Appendix A

Some Prerequisites

A.1 Metrics and Norms

Definition A.1.1 (Pseudo-metric, Metric, Metric space)**.** *For a set E, an application $d : E \times E \to \mathbb{R}_+$ is a pseudo-metric if it verifies the following properties:*

1. $\forall \mathbf{x} \in E, \quad d(\mathbf{x}, \mathbf{x}) = 0.$
2. $\forall \mathbf{x}_1, \mathbf{x}_2 \in E, \quad d(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_2, \mathbf{x}_1)$ *(symmetry).*
3. $\forall \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in E, \quad d(\mathbf{x}_1, \mathbf{x}_2) \leq d(\mathbf{x}_1, \mathbf{x}_3) + d(\mathbf{x}_3, \mathbf{x}_2)$ *(Triangle inequality).*

In addition, if $\forall \mathbf{x}_1, \mathbf{x}_2 \in E, \quad d(\mathbf{x}_1, \mathbf{x}_2) = 0 \Rightarrow \mathbf{x}_1 = \mathbf{x}_2$, then d is a metric. In this case, (E, d) is called a metric space.

Definition A.1.2 (Lipschitzness)**.** *Let (E, d_E) and (F, d_F) be two metric spaces. An application $f : E \to F$ verifies the Lipschitz property if there is some $L > 0$ such that for all $\mathbf{x}, \mathbf{x}' \in E$, we have:*

$$d_F(f(\mathbf{x}), f(\mathbf{x}')) \leq L \cdot d_E(\mathbf{x}, \mathbf{x}'). \tag{A.1}$$

In this case we also say that f is $L-$Lipschitz continuous or that f verifies the $L-$Lipschitzness.

Definition A.1.3 (Seminorm, Norm, Normed vector space)**.** *For a real vector space E, an application $\|.\| : E \to \mathbb{R}_+$ is a seminorm if it verifies the following properties:*

1. $\forall \lambda \in \mathbb{R}, \forall \mathbf{x} \in E, \|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|.$
2. $\forall \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in E, \quad \|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\mathbf{x}_1 - \mathbf{x}_3\| + \|\mathbf{x}_3 - \mathbf{x}_2\|$ *(Triangle inequality).*

In addition, if $\forall \mathbf{x} \in E, \quad \|\mathbf{x}\| = 0 \Rightarrow \mathbf{x} = 0$, then $\|.\|$ is a norm. In this case, $(E, \|.\|)$ is called a normed vector space.

Proposition A.1.1. *Any normed space $(E, \|.\|)$ is a metric space for $d : (\mathbf{x}, \mathbf{x}') \mapsto \|\mathbf{x} - \mathbf{x}'\|$.*

Definition A.1.4 (Dual norm)**.** *For a norm $\|.\|$ over a Hilbert space $(E, \langle ., . \rangle)$, its associated dual norm $\|.\|_*$ is defined for $\mathbf{x} \in E$ as:*

$$\|\mathbf{x}\|_* := \sup_{\substack{\mathbf{x}' \in E \\ \|\mathbf{x}'\| \leq 1}} \langle \mathbf{x}, \mathbf{x}' \rangle \tag{A.2}$$

Definition A.1.5 ($p-$norm)**.** *For $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_p) \in \mathbb{R}^n$, the $p-$norm of $\mathbf{x}$ for $p \in [1, \infty]$ is defined as:*

$$\|\mathbf{x}\|_p := \begin{cases} (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}} & \text{if} \quad p < \infty \\ \max_{1 \leq i \leq n}(|x_i|) & \text{if} \quad p = \infty \end{cases} \tag{A.3}$$

Definition A.1.6 (Hölder conjugate)**.** *Two number $p, q \geq 1$ are said to be Hölder conjugates of each other if $\frac{1}{p} + \frac{1}{q} = 1$.*

Proposition A.1.2 (Dual of $p-$norm)**.** *The dual of the $p-$norm is the $q-$norm, where p and q are Hölder conjugates.*

Definition A.1.7.

A.2 Probabilities

We review here some probability notions, and we assume that the reader is familiar with the Lebesgue measure theory and its associated definition of integrals.

Definition A.2.1 (σ-algebra)**.** *Let Ω be a set. A subset $\mathfrak{F}$ of the power set 2^Ω is called a $\sigma-$algebra if it satisfies the following conditions:*

1. *$E \in \mathfrak{F}$.*
2. *$\forall G \in \mathfrak{F}, E \setminus G \in \mathfrak{F}$ (stability by complementary).*
3. *For any sequence $\{G_i\}_{i\in\mathbb{N}}$ such that $G_i \in \mathfrak{F}$, we have*

$$\bigcup_{i\in\mathbb{N}} G_i \in \mathfrak{F}, \tag{A.4}$$

(stability by countable union).

In this case, $(\Omega, \mathfrak{F})$ is called a measurable space.

Definition A.2.2 (Probability)**.** *With the previous notations, let $P : \mathfrak{F} \to \mathbb{R}_+$. P is a probability if:*

1. *$P(\Omega) = 1$.*
2. *For any sequence $\{G_i\}_{i\in\mathbb{N}}$ such that $G_i \in \mathfrak{F}$ and $G_i \cap G_j = \emptyset$ if $i \neq j$, we have:*

$$P\left(\bigcup_{i\in\mathbb{N}} G_i\right) = \sum_{i\in\mathbb{N}} P(G_i). \tag{A.5}$$

In this case, $(\Omega, \mathfrak{F}, P)$ is called a probability space.

Definition A.2.3 (Random variable)**.** *With the previous notations, let $(E, \mathfrak{H})$ be a measurable space. A map $X : \Omega \to E$ is said to be measurable if:*

$$\forall G \in \mathfrak{H}, X^{-1}(G) \in \mathfrak{F}. \tag{A.6}$$

In this case, X is called a random variable.

Definition A.2.4 (Expected value)**.** *With the previous notations, let $X : \Omega \to E$ be a random variable. The expectation or expected value of X over probability P is the following quantity:*

$$\mathop{\mathbb{E}}_{\omega\sim P}[X(\omega)] := \int_\Omega X(\omega)\mathrm{dP}. \tag{A.7}$$

denoted also by $\mathbb{E}_P[X]$ if there is no ambiguity.

Definition A.2.5 (Pushforward measure)**.** *With the previous notations, random variable $X : \Omega \to E$ induces a probability measure on Ω, denoted P_X or $X\#P$ or $P \circ X^{-1}$ and defined for all $A \in \mathfrak{F}$ as:*

$$P_X(A) := P(\{X^{-1}(A)\}). \tag{A.8}$$

Theorem A.2.1 (change of variable formula)**.** *With the previous notations, we have*

$$\mathop{\mathbb{E}}_{\omega\sim P}[X(\omega)] = \mathop{\mathbb{E}}_{x\sim P_X}[x]\,. \tag{A.9}$$

Definition A.2.6 (Absolute continuity)**.** *Let $(\Omega, \mathfrak{F})$ be a measurable space. Let P and Q be two probabilities over Ω. We say that Q is absolutely continuous w.r.t. P, or that P dominates Q, and we write $Q \ll P$, if for all $A \in \mathfrak{F}$:*

$$P(A) = 0 \Rightarrow Q(A) = 0. \tag{A.10}$$

Theorem A.2.2. *(Nikodym, 1930) With the notations of the previous definition, assume that $Q \ll P$. Then there exists a random variable $\frac{\mathrm{d}Q}{\mathrm{d}P} : \Omega \to \mathbb{R}_+$, called the Radon-Nikodym derivative of Q w.r.t. P verifying:*

$$\forall A \in \mathfrak{F}, \quad Q(A) = \mathop{\mathbb{E}}_{\omega\sim P}\left[\frac{\mathrm{d}Q}{\mathrm{d}P}(\omega)[\omega \in A]\right]. \tag{A.11}$$

Theorem A.2.3. *With the previous notations, we have for any random variable $X : \Omega \to E$*

$$\mathop{\mathbb{E}}_{\omega\sim Q}[X(\omega)] = \mathop{\mathbb{E}}_{\omega\sim P}\left[\frac{\mathrm{d}Q}{\mathrm{d}P}\,(\omega)X(\omega)\right]. \tag{A.12}$$

Remark: Continuous and discrete probability distributions

- Probability distributions that are qualified as "continuous", where generally Ω as a subset of $\mathbb{R}^n$ with a non empty interior, are absolutely continuous w.r.t. the Lebesgues measure, and the Radon-Nykodim derivative in this case is the density.
- Probability distributions that are qualified as "discrete", where $\Omega = \{\omega_i\}_{i\in\mathbb{N}}$ is a countable set, are absolutely continuous w.r.t. the counting measure:

  $$\sum_{i\in\mathbb{N}} \delta_{\omega_i},$$

  and the Radon-Nykodim derivative in this case corresponds to the probability masses of the elements of Ω.

Below, we give a definition of the support of probability distributions in the particular case of continuous and discrete ones.

Definition A.2.7 (Support of a probability)**.**

- *For a continuous probability distribution with density function f_P, its support is:*

  $$\operatorname{supp} P := \{\mathbf{x} \in \Omega; f_P(\mathbf{x}) > 0\}$$

- *For a discrete probability distribution P, with probability masses denotes as p_ω, its support is:*

  $$\operatorname{supp} P := \{\omega \in \Omega; p_\omega > 0\}$$

Definition A.2.8. *(Kullback and Leibler, 1951) The Kullback-Leibler divergence between two probability distributions is*

$$\mathrm{KL}(Q\|P) := \begin{cases} \mathop{\mathbb{E}}_{\omega\sim Q}\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}\,(\omega)\right] & \text{if} \quad Q \ll P \\ \infty & \text{otherwise.} \end{cases} \tag{A.13}$$

Definition A.2.9. *(Hero et al., 2001) The Rényi-divergence or $\alpha-$divergence between two probability distributions is defined for $\alpha > 0$ and for $Q \ll P$ by:*

$$D_\alpha(Q\|P) := \begin{cases} \frac{1}{\alpha-1} \log \left(\mathop{\mathbb{E}}_{\omega\sim Q} \left[\frac{\mathrm{d}Q}{\mathrm{d}P} \ (\omega)^{\alpha-1} \right] \right) & \text{if} \quad \alpha \neq 1 \\ \mathrm{KL}(Q\|P) & \text{if} \quad \alpha = 1. \end{cases} \tag{A.14}$$

If $Q \not\ll P$ then $D_\alpha(Q\|P) := \infty$.

Below, we provide a special case of the *Mc-Diarmid* concentration inequality (Doob, 1940), which is a simplification of its statement in Shalev-Shwartz and Ben-David (2014, Lemma 26.4) as below we consider only one probability distribution.

Theorem A.2.4 (Mc Diarmid's inequality)**.** *Let V be some set, and let $f : V^m \to \mathbb{R}$ be a function such that for some $c > 0$, for all $x_1, \cdots, x_m, x_i' \in V$, we have:*

$$\left| f(x_1, \cdots, x_m) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_m) \right| \leq c.$$

Then for any probability distribution $\mathcal{D}$ over V, we have with a probability at least $1-\delta$ over the draw of a sample $S = (x_1, \cdots, x_m) \sim \mathcal{D}^m$:

$$|f(x_1, \cdots, x_m)| - \mathop{\mathbb{E}}_{S'=\{x_1', \cdots, x_m'\}\sim\mathcal{D}^m} \left[f(x_1', \cdots, x_m') \right] \leq c\sqrt{\frac{m}{2} \log \frac{2}{\delta}}$$

A.3 Details on the Bound of Zhang et al. (2019) in the Binary Case

In their work, Zhang et al. (2019) use vector-valued scoring functions that take their values in $\mathbb{R}^2$ in the case of binary classification. To this end, let us denote by $\vec{h} = (h_1, h_2)$ and $\vec{h}' = (h_1', h_2')$ the respective $\mathbb{R}^2$-valued counterparts of our real valued h and h' scoring functions. Similarly, we denote by $\underline{y}$ the label encoded in $\{0, 1\}$, and by y its counterpart in $\{-1, 1\}$. Then, according to the notations used by the authors, we have:

$$\rho_{\vec{h}}(\mathbf{x}, 1) = \frac{1}{2}(h_1(\mathbf{x}) - h_2(\mathbf{x})),$$
$$\rho_{\vec{h}}(\mathbf{x}, 0) = \frac{1}{2}(h_2(\mathbf{x}) - h_1(\mathbf{x})).$$

By linking their notations and ours via the relation $h = 2(h_1 - h_2)$ where > 0 is a fixed constant, the last two equations can be written:

$$\rho_{\vec{h}}(\mathbf{x}, \underline{y}) = yh(\mathbf{x}).$$

In particular, when we consider the class associated to a scoring function, *i.e.* , $\underline{y}(\vec{h}) := [h_1 > h_2]$ corresponding to $y(h) := \mathrm{sgn}\,(h)$, we have:

$$\rho_{\vec{h}'}(\mathbf{x}, \underline{y}_{\vec{h}}) = \mathrm{sgn}\,(h(\mathbf{x})) \cdot h'(\mathbf{x}).$$

Hence, we have for all $\beta > 0$,

$$\ell_{0,\beta} \circ \rho_{\vec{h}'}(\mathbf{x}, y(\vec{h})) = \ell_{0,\beta} \left(\mathrm{sgn}\,(h(\mathbf{x}))\, h'(\mathbf{x}) \right)$$

where $\ell_{0,\beta}$ is denoted Φ_β in Zhang et al. (2019).

Appendix B

Proofs for Chapter 3

Lemma 3.1.1 (same landmarks)**.** *Let K be an $(\epsilon,\gamma)-$good similarity for problem $(\mathcal{S},\mathcal{L})$. Then K is $(\epsilon+\epsilon',\gamma)$-good for problem $(\mathcal{T},\mathcal{L})$, where*

$$\epsilon' = d_{1+,\gamma}(\mathcal{T},\mathcal{S})\cdot\mathfrak{M}_{\mathcal{U},\mathcal{L}}(K)$$

and

$$d_{1+,\gamma}(\mathcal{T},\mathcal{S}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}}\left[\left(\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{U}_{\mathbb{X}}}(\mathbf{x}) - \frac{\mathrm{d}\mathcal{S}_{\mathbb{X}}}{\mathrm{d}\mathcal{U}_{\mathbb{X}}}(\mathbf{x})\right)_+ [y\cdot g_{\mathcal{L}}(\mathbf{x})<\gamma]\right].$$

Moreover, if $\mathcal{T}_{\mathbb{X}}\ll\mathcal{S}_{\mathbb{X}}$ then the obtained results holds with

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

where $d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\left[\left(\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x})-1\right)_+^2 [y\cdot g_{\mathcal{L}}(\mathbf{x})<\gamma]\right]$.

Proof. First, observe that

$$\mathfrak{E}_{\mathcal{T},\mathcal{L}}(K) = \mathfrak{E}_{\mathcal{S},\mathcal{L}}(K) + \mathfrak{E}_{\mathcal{T},\mathcal{L}}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}}(K) \le \epsilon + \mathfrak{E}_{\mathcal{T},\mathcal{L}}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}}(K) \tag{B.1}$$

following from the $(\epsilon,\gamma)-$goodness of K for $(\mathcal{P},\mathcal{L})$. Now we focus on the difference between the last two terms in (B.1). Bearing in mind that $\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}}\ll\mathcal{U}_{\mathbb{X}}$, that $\mathcal{U}_{\mathbb{Y}|\mathbf{x}} = \mathcal{T}_{\mathbb{Y}|\mathbf{x}} = \mathcal{S}_{\mathbb{Y}|\mathbf{x}}$ and denoting $\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{U}_{\mathbb{X}}}$ and $\frac{\mathrm{d}\mathcal{S}_{\mathbb{X}}}{\mathrm{d}\mathcal{U}_{\mathbb{X}}}$ respectively by t and s, we get the following:

$$\begin{aligned}
\mathfrak{E}_{\mathcal{T},\mathcal{L}}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}}(K) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{T}}[l_\gamma(g_{\mathcal{L}}(\mathbf{x}),y)] - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}}[l_\gamma(g_{\mathcal{L}}(\mathbf{x}),y)] \qquad \text{(B.2)}\\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{U}_{\mathbb{X}}}\left[t(\mathbf{x})\mathop{\mathbb{E}}_{y\sim\mathcal{S}_{\mathbb{Y}|\mathbf{x}}}[l_\gamma(g_{\mathcal{L}}(\mathbf{x}),y)]\right] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{U}_{\mathbb{X}}}\left[s(\mathbf{x})\mathop{\mathbb{E}}_{y\sim\mathcal{T}_{\mathbb{Y}|\mathbf{x}}}[l_\gamma(g_{\mathcal{L}}(\mathbf{x}),y)]\right]\\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}}[t(\mathbf{x})\cdot l_\gamma(g_{\mathcal{L}}(\mathbf{x}),y)] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{U}}[s(\mathbf{x})\cdot l_\gamma(g_{\mathcal{L}}(\mathbf{x}),y)]\\
&\le \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}}\left[(t(\mathbf{x})-s(\mathbf{x}))_+\ell_\gamma(y\cdot g_{\mathcal{L}}(\mathbf{x}))[y\cdot g_{\mathcal{L}}(\mathbf{x})<\gamma]\right], \qquad \text{(B.3)}
\end{aligned}$$

where (B.3) is due to the inequality $t\le t_+$ for $t\in\mathbb{R}$, the relation between l_γ and ℓ_γ, and $\forall t\in\mathbb{R}$, the positivity of ℓ_γ and its nullity for $t\ge\gamma$.

Using (B.3), we write:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}}\left[(t(\mathbf{x})-s(\mathbf{x}))_+\ell_\gamma(y\cdot g_{\mathcal{L}}(\mathbf{x}))[y\cdot g_{\mathcal{L}}(\mathbf{x})<\gamma]\right]\\
\le&\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}}\left[(t(\mathbf{x})-s(\mathbf{x}))_+[y\cdot g_{\mathcal{L}}(\mathbf{x})<\gamma]\right]\cdot\mathfrak{M}_{\mathcal{U},\mathcal{L}}(K) \qquad \text{(B.4)}\\
=&d_{1+,\gamma}(\mathcal{T},\mathcal{S})\cdot\mathfrak{M}_{\mathcal{U},\mathcal{L}}(K)
\end{aligned}$$

where we use Hölder's inequality with ℓ_1 and ℓ_∞ norms to obtain (B.4).

For the case when $\mathcal{S}_{\mathbb{X}}$ dominates $\mathcal{T}_{\mathbb{X}}$, we take $\mathcal{U}_{\mathbb{X}} = \mathcal{S}_{\mathbb{X}}$, implying that $t(\mathbf{x}) = \frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}$ and $s(\mathbf{x}) = 1$, and we have:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{U}} \left[(t(\mathbf{x}) - s(\mathbf{x}))_+ l_\gamma(g_{\mathcal{L}}(\mathbf{x}), y)[y \cdot g_{\mathcal{L}}(\mathbf{x}) < \gamma]\right]^2 && \text{(B.5)}\\
=&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}} \left[(t(\mathbf{x}) - 1)_+ l_\gamma(g_{\mathcal{L}}(\mathbf{x}), y)[y \cdot g_{\mathcal{L}}(\mathbf{x}) < \gamma]\right]^2 \\
\leq&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}} \left[(t(\mathbf{x}) - 1)_+^2 [y \cdot g_{\mathcal{L}}(\mathbf{x}) < \gamma]\right] \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}} \left[l_\gamma(g_{\mathcal{L}}(\mathbf{x}), y)^2\right] && \text{(B.6)}\\
=&d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}} \left[l_\gamma(g_{\mathcal{L}}(\mathbf{x}), y)^2\right] \\
\leq&d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) \cdot \mathfrak{M}_{\mathcal{S},\mathcal{L}}(K) \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}} [l_\gamma(g_{\mathcal{L}}(\mathbf{x}), y)] && \text{(B.7)}\\
\leq&d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) \cdot \mathfrak{M}_{\mathcal{S},\mathcal{L}}(K) \cdot \epsilon.
\end{aligned}
$$

To obtain (B.6), we applied the Cauchy-Schwartz inequality. Inequality B.7 is obtained thanks to the boundedness and positivity of l_γ via Hölder inequality for norms ℓ_1 and ℓ_∞. The last line follows from the $(\epsilon, \gamma)-$goodness of K for problem $(\mathcal{S}, \mathcal{L})$. □

B.1 Proof from Section 4

Theorem 3.2.1. *Let K be a similarity function defined on a feature space $\mathbb{X}$. Let $\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K)$ denote its worst performance associated to loss function l_γ and achieved by an example drawn from $\mathcal{S}$, where $\mathcal{L}$ is the landmarks distribution. Assume that $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$ and that the cumulative distribution function F_{l_γ} of the loss function associated with $\mathcal{S}$ and $\hat{L}$ is k times differentiable at $\mathfrak{M}_{\mathcal{S},\hat{L}}(K)$, and that $k > 0$ is the minimum integer such that $F^{(k)}_{l_\gamma}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K)) \neq 0$. Then, for all $\alpha > 1, r \geq 1$, there exists $m_0 \geq 1$ such that for all $m \geq m_0$, we have with probability at least $1 - \delta$:*

$$
\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K) \leq \mathfrak{M}_{\hat{S},\hat{L}}(K) + \frac{2}{\gamma}\operatorname{Rad}_r(\mathbb{H}_1(K)) + \frac{1}{\gamma}\sqrt{2\frac{\log\left(\frac{4}{\delta}\right)}{r}} + \left(\frac{(-1)^{k+1}\log\left(\frac{2\alpha}{\delta}\right)k!}{F^{(k)}_{l_\gamma}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K))m}\right)^{\frac{1}{k}},
$$

where $\mathbb{H}_1(K)$ is the hypothesis class defined by $\mathbb{H}_1(K) := \{h_{\mathbf{x}} : \mathbf{x}' \mapsto K(\mathbf{x}, \mathbf{x}'), \mathbf{x} \in \operatorname{supp}\mathcal{S}_{\mathbb{X}}\}$.

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

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

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

$$
M_1 = \mathfrak{M}_{\mathcal{S},\mathcal{L}}(K) - \mathfrak{M}_{\mathcal{S},\hat{L}}(K), \quad M_2 = \mathfrak{M}_{\mathcal{S},\hat{L}}(K) - \mathfrak{M}_{\hat{S},\hat{L}}(K).
$$

We begin by bounding M_1:

$$
\begin{aligned}
M_1 =& \sup_{(\mathbf{x},y)\in\operatorname{supp}\mathcal{S}} l_\gamma(y \cdot g_{\mathcal{L}}(\mathbf{x})) - \sup_{(\mathbf{x},y)\in\operatorname{supp}\mathcal{S}} l_\gamma(y \cdot g_{\hat{L}}(\mathbf{x})) && \text{(B.8)}\\
\leq& \sup_{(\mathbf{x},y)\in\operatorname{supp}\mathcal{S}} \{l_\gamma(y \cdot g_{\mathcal{L}}(\mathbf{x})) - l_\gamma(y \cdot g_{\hat{L}}(\mathbf{x}))\} && \text{(B.9)}\\
\leq& \frac{1}{\gamma} \sup_{\mathbf{x}\in\operatorname{supp}\mathcal{S}_{\mathbb{X}}} |g_{\mathcal{L}}(\mathbf{x}) - g_{\hat{L}}(\mathbf{x})| && \text{(B.10)}\\
=& \frac{1}{\gamma} \sup_{\mathbf{x}\in\operatorname{supp}\mathcal{S}_{\mathbb{X}}} \left| \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{L}} \left[y' K(\mathbf{x}, \mathbf{x}')\right] - \frac{1}{r}\sum_{i=1}^{r} y'_i K(\mathbf{x}, \mathbf{x}'_i) \right|, && \text{(B.11)}
\end{aligned}
$$

where (B.10) holds due to the $\frac{1}{\gamma}$-Lipschitzness of ℓ_γ[1]. The quantity in (B.11) is known as the representativeness (see, for example, Shalev-Shwartz and Ben-David (2014)) of sample L drawn from $\mathcal{L}$ associated with the hypothesis set $\mathcal{Y} \cdot \mathbb{H}_1(K)$. In what follows, we denote it by $\mathrm{Rep}_{\mathcal{L}}(\mathcal{Y} \cdot \mathbb{H}_1(K), L)$ and notice that its value changes at most by $\dfrac{2}{r}$ if an instance of L is replaced since K takes values in $[-1, 1]$. By applying Mc-Diarmid's inequality (Theorem A.2.4), we have with a probability at least $1 - \frac{\delta}{2}$ for $0 < \delta \leq 1$

$$\mathrm{Rep}_{\mathcal{L}}(\mathcal{Y} \cdot \mathbb{H}_1(K), L) \leq \mathop{\mathbb{E}}_{L \sim \mathcal{L}^m} \left[\mathrm{Rep}_{\mathcal{L}}(\mathcal{Y} \cdot \mathbb{H}_1(K), L)\right] + \sqrt{2\frac{\log\left(\frac{4}{\delta}\right)}{r}}. \tag{B.12}$$

The expectation term in (B.12) can be bounded by twice the Rademacher complexity of hypotheses class $\mathcal{Y} \cdot \mathbb{H}_1(K)$ denoted by $\mathrm{Rad}_r\left(\mathcal{Y} \cdot \mathbb{H}_1(K)\right)$ (see, for example, (Shalev-Shwartz and Ben-David, 2014, Lemma 26.2)), which also equals $\mathrm{Rad}_r\left(\mathbb{H}_1(K)\right)$. Hence, with a probability at least $1 - \frac{\delta}{2}$, we have:

$$M_1 \leq \frac{2}{\gamma} \mathrm{Rad}_r\left(\mathbb{H}_1(K)\right) + \frac{1}{\gamma}\sqrt{2\frac{\log\left(\frac{4}{\delta}\right)}{r}}. \tag{B.13}$$

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

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

where we obtain the last two lines because the elements of S are drawn independently. By the assumptions made on the regularity of F_{l_γ}, setting t to $\dfrac{t}{m^{\frac{1}{k}}}$ yields:

$$\mathop{\mathbb{P}}_{S \sim \mathcal{S}^m}\left[M_2 \geq \frac{t}{m^{\frac{1}{k}}}\right] \tag{B.14}$$

$$= \left(1 + F_{l_\gamma}^{(k)}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K))\frac{(-t)^k}{mk!} + o\left(\frac{t^k}{m}\right)\right)^m \underset{m \to \infty}{\longrightarrow} \exp\left(F_{l_\gamma}^{(k)}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K))\frac{(-t)^k}{k!}\right), \tag{B.15}$$

where the left-hand side in (B.15) is obtained from a Taylor expansion. This implies for any $\alpha > 1$ that there exists $m_0 \in \mathbb{N}^*$ such that for all $m \geq m_0$,

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

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

$$M_2 \leq \left(\frac{(-1)^{k+1}\log\left(\frac{2\alpha}{\delta}\right)k!}{F_{l_\gamma}^{(k)}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K))m}\right)^{\frac{1}{k}}. \tag{B.16}$$

Finally we use a union bound to bound the probability that the two inequalities (B.13) and (B.16) occur simultaneously in order to obtain the desired result. □

[1] We recall that $l_\gamma(y, y') := \ell_\gamma(y \cdot y')$ for $y, y' \in \mathbb{Y}$.

Lemma 3.1.2. *Let K be an $(\epsilon,\gamma)-$good similarity for problem $(\mathcal{S},\mathcal{L})$ and assume that $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$. Then K is $(\epsilon+\epsilon',\gamma)$-good for problem $(\mathcal{T},\mathcal{L})$, with*

$$\epsilon' = \sqrt{2 \sup_{(\mathbf{x},y)\in \operatorname{supp}\mathcal{S}} \left(\left(\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) - 1 \right)_+ l_\gamma(g_{\mathcal{L}}(\mathbf{x}), y) \right) d_1(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}})\epsilon}$$

and

$$d_1(\mathcal{T}_{\mathbb{X}}, \mathcal{S}_{\mathbb{X}}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}} \left[\left| \frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) - 1 \right| \right].$$

Proof. To simplify notations, let X and Y denote the non-negative random variables $\left(\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}(\mathbf{x}) - 1\right)_+$ and $l_\gamma(yg_{\mathcal{L}}(\mathbf{x}))$, where we omit the dependence on $(\mathbf{x},y)\sim\mathcal{S}$.. For any $t>0$, and any $p,q\in[1,\infty)$ such that $\frac{1}{p}+\frac{1}{q}=1$, we have:

$$\begin{aligned}
\mathbb{E}[XY] &= \mathbb{E}[XY[Y\le t]] + \mathbb{E}[XY[Y>t]] \\
&\le \mathbb{E}[X]\,t + \mathbb{E}[|XY|^p]^{\frac{1}{p}}\,\mathbb{E}[[Y>t]^q]^{\frac{1}{q}} && \text{(B.17)}\\
&= \mathbb{E}[X]\,t + \mathbb{E}[|XY|^p]^{\frac{1}{p}} + \mathbb{P}[[Y>t]]^{\frac{1}{q}} && \text{(B.18)}\\
&\le \mathbb{E}[X]\,t + \mathbb{E}[|XY|^p]^{\frac{1}{p}}\,\mathbb{E}[Y]^{\frac{1}{q}}\,t^{-\frac{1}{q}} && \text{(B.19)}
\end{aligned}$$

where the second and fourth lines are obtained respectively by applying the Hölder inequality and the Markov inequality. This bound being true for all $t>0$, we can quickly compute the minimum of the right hand side, yielding the following:

$$\mathbb{E}[XY] \le C_q \cdot \mathbb{E}[(XY)^p]^{\frac{1}{2p-1}} \left(\mathbb{E}[X]\,\mathbb{E}[Y]\right)^{\frac{1}{1+q}} \begin{cases} \xrightarrow{q\to\infty} \mathbb{E}[XY] \\ \xrightarrow{q\to 1} 2\sqrt{\sup_{(\mathbf{x},y)\in\operatorname{supp}\mathcal{S}}(XY)\mathbb{E}[X]\,\mathbb{E}[Y]} \end{cases} \tag{B.20}$$

where $C_q = \left(1+\frac{1}{q}\right) q^{\frac{1}{q+1}}$. Returning to X and Y's expressions, we obtain:

$$\mathbb{E}[XY] \le C_q \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} [(XY)^p]^{\frac{1}{2p-1}}\, d_{1+}(\mathcal{S},\mathcal{T})^{\frac{1}{q+1}}\, \epsilon^{\frac{1}{q+1}} \tag{B.21}$$

$$\tag{B.22}$$

In particular, for $q=1$, we have:

$$\mathbb{E}[XY] \le 2\sqrt{\sup_{(\mathbf{x},y)\in\operatorname{supp}\mathcal{S}}(XY) d_{1+}(\mathcal{S},\mathcal{T})\epsilon},$$

so we choose $q=1$ as we have: $\sqrt{\epsilon} \le \epsilon^{\frac{1}{q+1}}$ for any $q\ge 1$ and any $\epsilon\in(0,1]$. Now, observe that:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[t(\mathbf{x})-1] = 0 &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[(t(\mathbf{x})-1)_+] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[(t(\mathbf{x})-1)_-] \\
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[|t(\mathbf{x})-1|] &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[(t(\mathbf{x})-1)_+] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[(t(\mathbf{x})-1)_-]
\end{aligned}$$

which implies that:

$$d_1(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}}) := \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[|t(\mathbf{x})-1|] = 2\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}\left[(t(\mathbf{x})-1)_+\right] = 2\cdot d_{1+}(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}})$$

Combining the last result with (B.1) yields the result of the theorem. □

Appendix C

Proofs and Supplementary Material for Chapter 4

C.1 Proofs

Proposition C.1.1. *The margin violation loss does not respect the triangle inequality. In particular, for any $\rho \in (0,1)$ there exists $x, y, z \in (0,1)$ such that:*

$$l_{\rho,0}(x,y) > l_{\rho,0}(x,z) + l_{\rho,0}(z,y) \tag{C.1}$$

Proof. Let $\rho, \epsilon_1, \epsilon_2 > 0$ such that $0 < \sqrt{\rho} < \epsilon_1 < \epsilon_2 < 1$. Also, let $x = y = \epsilon_2\sqrt{\rho}$ and $z = \frac{\sqrt{\rho}}{\epsilon_1}$.

We have $x, y, z \in (0,1)$ and :

$$xz = yz = \epsilon_2\sqrt{\rho}\frac{\sqrt{\rho}}{\epsilon_1} = \frac{\epsilon_2}{\epsilon_1}\rho > \rho.$$

Hence,

$$[xz < \rho] = [yz < \rho] = 0.$$

However,

$$xy = \epsilon_2^2\rho < \rho,$$

hence $[xy < \rho] = 1$ and we have:

$$1 = [xy < \rho] > [xz < \rho] + [yz < \rho] = 0.$$

□

Corollary C.1.1. *For any $\rho > 0$, the loss $\rho-$scaled hinge loss defined as $l(y,y') = \left(1 - \frac{y \cdot y'}{\rho}\right)_+$ does not verify the triangle inequality.*

Proof. Taking x, y, z as in the proof of Proposition C.1.1, we have:

$$l(x,z) = l(y,z) = \left(1 - \frac{x \cdot z}{\rho}\right)_+ = \left[1 > \frac{x \cdot z}{\rho}\right]\left(1 - \frac{x \cdot z}{\rho}\right) = 0,$$

because $[xz < \rho] = [yz < \rho] = 0$. Similarly, we have

$$l(x,y) = \left[1 > \frac{x \cdot y}{\rho}\right]\left(1 - \frac{x \cdot y}{\rho}\right) = \left(1 - \frac{x \cdot y}{\rho}\right) > 0.$$

Hence, $l(x,y) > l(x,z) + l(z,y)$. □

Remark: Before proving Theorem 4.2.1, we prove lemma Lemma C.1.1.

Lemma C.1.1. *Let $h, f : \mathbb{X} \to [-1, 1]$, and let $\mathcal{D}$ be a probability distribution over $\mathbb{X} \times \mathbb{Y}$. Assume that $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}[f(\mathbf{x}) = 0] = 0$. Then:*

1. *For any $0 \leq \rho \leq 1$ we have:*

$$\mathfrak{E}_{\mathcal{D}}^{\rho,0}(h) \leq \mathfrak{E}_{\mathcal{D}}^{\rho,0}(h, f) + \mathfrak{E}_{\mathcal{D}}^{0,0}(f).$$

2. *For any $0 \leq \rho < \alpha \leq 1$ we have:*

$$\mathfrak{E}_{\mathcal{D}}^{\rho,0}(h, f) \leq \mathfrak{E}_{\mathcal{D}}^{\frac{\rho}{\alpha},0}(h, f) + \mathfrak{E}_{\mathcal{D}}^{\alpha,0}(f).$$

Proof. Let $\theta \in]0, 1[$. We have

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[yh(\mathbf{x}) < \theta\rho] && \text{(C.2)}\\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[yh(\mathbf{x}) \cdot f(\mathbf{x})^2 < \theta\rho.f(\mathbf{x})^2\right] && \text{(C.3)}\\
&\leq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}[h(\mathbf{x})f(\mathbf{x}) < \rho|f(\mathbf{x})|] + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[yf(\mathbf{x}) < \theta|f(\mathbf{x})|] && \text{(C.4)}\\
&\leq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}[h(\mathbf{x})f(\mathbf{x}) < \rho] + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y\,\mathrm{sgn}(f(\mathbf{x})) < \theta]\\
&= \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}[h(\mathbf{x})f(\mathbf{x}) < \rho] + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y\,\mathrm{sgn}(f(\mathbf{x})) < 0] && \text{(C.5)}\\
&= \mathfrak{E}_{\mathcal{D}}^{\rho,0}(h, f) + \mathfrak{E}_{\mathcal{D}}^{0,0}(f). && \text{(C.6)}
\end{aligned}$$

In the previous developments, we first use the fact that $yh(\mathbf{x})f(\mathbf{x})^2 < \theta\rho f^2(\mathbf{x})$ implies that either $h(\mathbf{x})f(\mathbf{x}) < \rho|f(\mathbf{x})|$ or $yf(\mathbf{x}) < \theta|f(\mathbf{x})|$ to obtain (C.4). The next inequality comes from $|f(\mathbf{x})| \leq 1$ as $f \in \mathbb{H}'$. Then, we use the fact that $\frac{f(\mathbf{x})}{|f(\mathbf{x})|} = \mathrm{sgn}(f(\mathbf{x}))$, where the division by $|f(\mathbf{x})|$ is possible if we restrict the previous developments for the set $\{\mathbf{x} \in \mathbb{X}; f(\mathbf{x}) \neq 0\}$, because the set where $f(\mathbf{x}) = 0$ is negligible by the assumption made on f. Finally, as $0 < \theta < 1$ and $y\,\mathrm{sgn}(f) \in \{-1, 1\}$, the inequality $y\,\mathrm{sgn}(f(\mathbf{x})) < \theta$ is equivalent to $y\,\mathrm{sgn}(f) = -1 < 0$ and to $yf(\mathbf{x}) < 0$, implying (C.5) and (C.6). Taking the limit as $\theta \to 1$ in the previous inequality (by the Beppo-Levi monotonous convergence theorem) finishes the proof of the first point.

For the second point, let $\theta \in]0, 1[$. We have:

$$\begin{aligned}
\mathfrak{E}_{\mathcal{D}}^{\rho,0}(h, f) &= \mathbb{P}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[h(\mathbf{x})f(\mathbf{x}) < \rho]\\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}\left[yh(\mathbf{x}) \cdot yf(\mathbf{x}) < \frac{\rho}{\theta|f(\mathbf{x})|}\theta|f(\mathbf{x})|\right]\\
&\leq \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}\left[yh(\mathbf{x}) < \frac{\rho}{\theta|f(\mathbf{x})|}\right] + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}[yf(\mathbf{x}) < \theta|f(\mathbf{x})|]\\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}\left[yh(\mathbf{x}) < \frac{\rho}{\theta|f(\mathbf{x})|}\right] + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}[y\,\mathrm{sgn}(f(\mathbf{x})) < \theta]\\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}\left[yh(\mathbf{x}) < \frac{\rho}{\theta|f(\mathbf{x})|}\right] + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}[y\,\mathrm{sgn}(f(\mathbf{x})) < 0] && \text{(C.7)}\\
&\xrightarrow{\theta\to 1} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}\left[yh(\mathbf{x}) < \frac{\rho}{|f(\mathbf{x})|}\right] + \mathfrak{E}_{\mathcal{S}}^{0,0}(f(\mathbf{x}))\\
&\leq \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{S}}\left[yh(\mathbf{x}) < \frac{\rho}{\alpha}\right] + \mathbb{P}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[|f(\mathbf{x})| < \alpha] + \mathfrak{E}_{\mathcal{S}}^{0,0}(f(\mathbf{x})) && \text{(C.8)}
\end{aligned}$$

where $\alpha > \rho$ is arbitrarily chosen. To obtain (C.7), we applied the same technique as the one used to prove (C.5). Then, taking the limit as $\theta \to 1$ is justified by the Beppo-Levi monotonous convergence theorem. Finally, we use conditioning on the event $\{|f(\mathbf{x})| \geq \alpha\}$, and the fact that the event $\{yh(\mathbf{x}) < \frac{\rho+\beta}{|f(\mathbf{x})|} \wedge |f(\mathbf{x})| \geq \alpha\}$ implies event $\{yh(\mathbf{x}) < \frac{\rho+\beta}{\alpha}\}$, to obtain (C.8). □

C.2 Empirical Case and Optimization Problem

In this section, we detail the derivation of Equation (4.33). The empirical cost function of Problem (4.28) is:

$$\frac{1}{m}\sum_{1\leq i\leq m} l(\mathbf{w}^T\mathbf{x}_{s,i}, y_{s,i}) + \delta\left\|\sum_{\substack{1\leq i\leq m_s\\1\leq j\leq m_t}} (\mathbf{P})_{ij}\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|\right\|_\infty + \zeta\|\mathbf{w}\|_2^2 \tag{C.9}$$

is a function of $\mathbf{w} \in \mathbb{R}^n$ and $\mathbf{P} \in \Pi$ having elements $(\mathbf{P})_{ij}$.

For a fixed $\mathbf{w} \in \mathbb{R}^n$, we would like to find $\mathbf{P} \in \Pi$ minimizing:

$$\left\|\sum_{\substack{1\leq i\leq m_s\\1\leq j\leq m_t}} (\mathbf{P})_{ij}\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|\right\|_\infty \tag{C.10}$$

However, we have:

$$\begin{aligned}
&\min_{\mathbf{P}\in\Pi}\left\|\sum_{ij}(\mathbf{P})_{ij}\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|\right\|_\infty && \text{(C.11)}\\
=&\min_{\mathbf{P}\in\Pi}\max_{1\leq k\leq d}\sum_{ij}(\mathbf{P})_{ij}\mathbf{e}_k^T\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right| && \text{(C.12)}\\
=&\min_{\mathbf{P}\in\Pi}\max_{\mathbf{q}\in\Delta_n}\sum_{ij}(\mathbf{P})_{ij}\mathbf{q}^T\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right| && \text{(C.13)}\\
=&\max_{\mathbf{q}\in\Delta_n}\min_{\mathbf{P}\in\Pi}\sum_{ij}(\mathbf{P})_{ij}\mathbf{q}^T\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right| && \text{(C.14)}\\
=&-\min_{\mathbf{q}\in\Delta_n}\max_{\mathbf{P}\in\Pi}\left(-\sum_{ij}(\mathbf{P})_{ij}\mathbf{q}^T\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|\right) && \text{(C.15)}
\end{aligned}$$

where $\mathbf{e}_k$ are the vectors of the canonical basis of $\mathbb{R}^n$. Due to the positivity of coordinates of vector $\sum_{ij}(\mathbf{P})_{ij}\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|$, we obtain (C.12). Then, since the function $\mathbf{q} \mapsto \sum_{ij}(\mathbf{P})_{ij}\mathbf{q}^T\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|$ is linear, its maximum is achieved over the vertices of the probability simplex, hence the equality between (C.12) and (C.13). Furthermore, due to the linearity of both $\mathbf{q} \mapsto \sum_{ij}(\mathbf{P})_{ij}\mathbf{q}^T\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|$ and $\mathbf{P} \mapsto \sum_{ij}(\mathbf{P})_{ij}\mathbf{q}^T\left|(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w}\right|$, and to the convexity and compactness of Δ_n and Π, Von-Neumann's minimax theorem allows us to permute the maximum and the minimum to obtain (C.14). Finally, introducing minus sign yields (C.15).

C.3 Experiments

C.3.1 Smooth Proxies used for Optimization

We use the smooth proxies to provide smooth functions for . They all depend on a parameter $\kappa > 0$. For all our experiments, we set $\kappa = 0.1$.

Smooth proxy of the positive part

We define the smooth proxy of the positive part function as:

$$\begin{array}{ll} \mathrm{pos}_\kappa : t \mapsto \dfrac{1}{2\kappa}\left(t + \dfrac{\kappa}{2}\right)^2 & \text{if } -\dfrac{\kappa}{2} \leq t \leq \dfrac{\kappa}{2} \\ (t)_+ & \text{otherwise} \end{array}$$

plotted in figure C.1 (left) and verifying $\mathrm{pos}_\kappa(t) \xrightarrow{\kappa \to 0} (t)_+$ for all $t \in \mathbb{R}$.

Smooth proxy of the absolute value

Since for any $t \in \mathbb{R}$, one has $|t| = (t)_+ + (-t)_+$, we define a smooth proxy of the absolute value in a similar manner:

$$\mathrm{abs}_\kappa(t) = \mathrm{pos}_\kappa(t) + \mathrm{pos}_\kappa(-t)$$

plotted in figure C.1 (right) and verifying $\mathrm{abs}_\kappa(t) \xrightarrow{\kappa \to 0} |t|$ for all $t \in \mathbb{R}$.

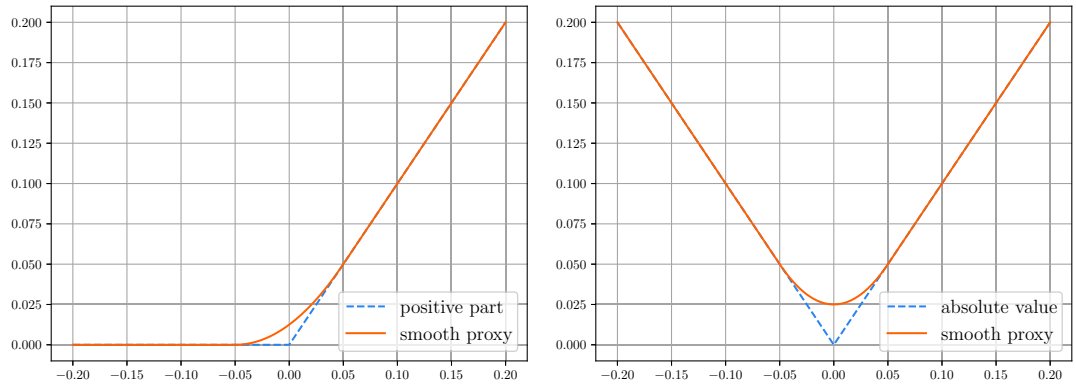


Figure C.1: (**left**) Smooth proxy of the positive part function; (**right**) Smooth proxy of the absolute value function.

Smooth proxy of the infinite norm

As in Problem (4.28), the coordinates of vector $\mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} \left[\left| (\mathbf{x}_s \mathbf{x}_s^T - \mathbf{x}_t \mathbf{x}_t^T) \mathbf{w} \right| \right]$ are positive, computing its infinite norm is simply computing its largest coordinate. Hence, we use the $\mathrm{logsumexp}_\kappa$ function defined for $\mathbf{t} = (t_1, ..., t_n)$ in $\mathbb{R}^n$ as:

$$\mathrm{logsumexp}_\kappa(\mathbf{t}) = \kappa \log\left(\sum_{k=1}^{n} e^{\frac{t_k}{\kappa}} \right)$$

verifying $\mathrm{logsumexp}_\kappa(\mathbf{t}) \xrightarrow{\kappa \to 0} \max(t_1, ..., t_n)$ for all $\mathbf{t} \in \mathbb{R}^n$ with distinct coordinates. While we use the former expression for optimization over $\mathbf{w}$, we use its maximum form proved in (Nesterov, 2005, Lemma 4) for optimization over $\mathbf{P}$:

$$\mathrm{logsumexp}_\kappa(\mathbf{t}) = \max_{\mathbf{q} \in \Delta_n} \langle \mathbf{q}, \mathbf{t} \rangle - \kappa \sum_{i=1}^{n} q_i \log q_i,$$

where $\mathbf{q} = (q_1, \cdots, q_n)$. This is an entropy-regularized variation of the maximum coordinate function $\mathbf{t} \to \max(t_1, \cdots, t_n)$. This form is more suitable for minimization over transport matrix $\mathbf{P}$ as the minimization in this case is analogous to problem (C.15), and the problem becomes:

$$\min_{\mathbf{q}\in\Delta_n} \max_{\mathbf{P}\in\Pi} \left(-\sum_{ij} (\mathbf{P})_{ij} \mathbf{q}^T \left| (\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{w} \right| + \kappa \sum_{i=1}^{n} q_i \log q_i \right). \tag{C.16}$$

C.3.2 Illustrations on the Moons Dataset

In this section, we provide supplementary illustrations of our algorithm on the intertwinning moons data. On Figure C.2, we plotted the decision boundary for $\zeta = 10^{-5}$ and for $\delta \in \{0.1, 0.5, 1, 5, 10\}$ in order to illustrate the influence of the latter hyperparameter on classifier h. As one can notice, for "low" values of δ ($\delta \in \{0.1, 0.5\}$), the decision boundary moderately fits the target data for a rotation angle up to 40°. Conversely, for $\delta \in \{5, 10\}$, the decision boundary is more suited to angles greater than 60°. These observations confirm δ's rule as a weight for the alignment term. Also, the decision boundary has less "curvature" for these values of δ, implying that the alignment term acts like a regularizer in addition to its originally intended role.

C.3.3 Used Libraries

In all of the experiments, we use the `scipy` library L-BFGS optimizer (Virtanen et al., 2020) for optimization over $\mathbf{w}$ and `CVXPY` (Diamond and Boyd, 2016) and `POT` (Flamary and Courty, 2017) for optimization over transport matrix $\mathbf{P}$.

C.4 $\mathbb{H}'$ as the Space of L_p Bounded Linear Classifiers

In this appendix section, we develop some implications of considering $\mathbb{H}'$ to be the set of L_p bounded classifiers for $p \geq 1$ in general. We define the support of transport matrix $\mathbf{P}$ as the set of indices with a non null value, *i.e.*

$$\operatorname{supp} \mathbf{P} := \{(i, j); (\mathbf{P})_{ij} > 0\},$$

then we reindex this set by index k varying from 1 to $|\operatorname{supp} \mathbf{P}|$, and we denote the symmetric matrices $(\mathbf{P})_{ij}(\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)$ having indices $(i, j) \in \operatorname{supp} \mathbf{P}$ as $\{\mathbf{D}_k\}_{k=1}^{|\operatorname{supp} P|}$. Using these latter, we define matrix $\mathbf{A}(\mathbf{w}, \mathbf{P})$ as the matrix having rows $\mathbf{w}^T\mathbf{D}_k$. It follows that:

$$\begin{aligned}
\Delta_{\mathbb{H}'}(h, \mathcal{P}) &= \sup_{\|\mathbf{v}\|_p \leq 1} \sum_{\substack{1\leq i\leq m_s \\ 1\leq j\leq m_t}} (\mathbf{P})_{ij} \left| \mathbf{w}^T (\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{v} \right| && \text{(C.17)} \\
&= \sup_{\|\mathbf{v}\|_p \leq 1} \sum_{(i,j)\in\operatorname{supp}\mathbf{P}} (\mathbf{P})_{ij} \left| \mathbf{w}^T (\mathbf{x}_{s,i}\mathbf{x}_{s,i}^T - \mathbf{x}_{t,j}\mathbf{x}_{t,j}^T)\mathbf{v} \right| && \text{(C.18)} \\
&= \sup_{\|\mathbf{v}\|_p \leq 1} \sum_{k=1}^{|\operatorname{supp}\mathbf{P}|} \left| \mathbf{w}^T \mathbf{D}_k \mathbf{v} \right| = \sup_{\|\mathbf{v}\|_p \leq 1} \|\mathbf{A}(\mathbf{w}, \mathbf{P})\mathbf{v}\|_1 = \|\mathbf{A}(\mathbf{w}, \mathbf{P})\|_{p\to 1} && \text{(C.19)}
\end{aligned}$$

The last quantity is an operator norm of matrix $\mathbf{A}(\mathbf{w}, \mathbf{P})$, and is NP-hard to compute for $p > 1$ (Steinberg, 2005, Chapter 2). In Chapter 4, we already treated the case $p = 1$, which does not suffer from this limitation, and for which the norm has a closed form. Nevertheless, the power method (Boyd, 1974) can be used to compute such a norm. Using duality between the $1-$norm and the $\infty-$norm, it is written:

$$\|\mathbf{A}(\mathbf{w}, \mathbf{P})\|_{p\to 1} = \sup_{\|\mathbf{v}\|_p \leq 1} \|\mathbf{A}(\mathbf{w}, \mathbf{P})\mathbf{v}\|_1 = \sup_{\|\mathbf{u}\|_\infty \leq 1} \sup_{\|\mathbf{v}\|_p \leq 1} \mathbf{u}^T \mathbf{A}(\mathbf{w}, \mathbf{P})\mathbf{v} \tag{C.20}$$

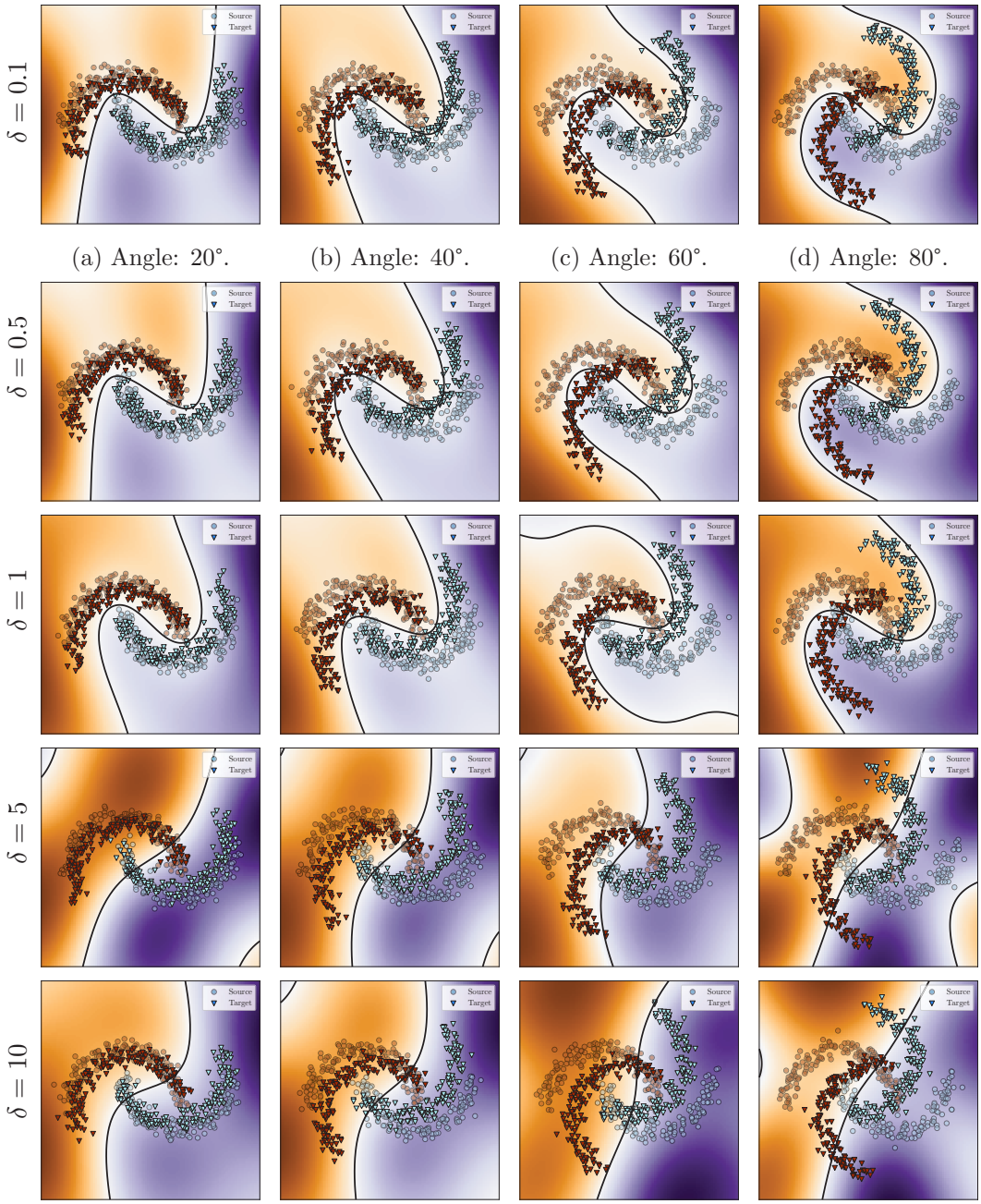


Figure C.2: Decision boundary for the intertwinning moons dataset with different δ values. Parameters ζ's value is set to $\zeta = 10^{-5}$ in all of the experiments.

There exist vectors $\mathbf{u}_*$ and $\mathbf{v}_*$ for which we have:

$$\|\mathbf{A}(\mathbf{w},\mathbf{P})\|_{p\to 1} = \mathbf{u}_*^T \mathbf{A}(\mathbf{w},\mathbf{P})\mathbf{v}_*.$$

The power method provides a way of estimating these two vectors after several initializations, by successively taking the supremum over one while fixing the other.
For a fixed $\mathbf{P}$, the function $\|\mathbf{A}(.,\mathbf{P})\|_{p\to 1}$ is a supremum over a family of convex (linear) functions of $\mathbf{w}$ (Equation (C.20)). Using Danskin's theorem (Danskin, 1966; Bertsekas, 1971), the subgradient of $\|\mathbf{A}(.,\mathbf{P})\|_{p\to 1}$ with respect to $\mathbf{w}$ is given by the following set:

$$\partial_{\mathbf{w}}\|\mathbf{A}(.,\mathbf{P})\|_{p\to 1} = \text{Conv}\left(\left\{\nabla_{\mathbf{w}}\mathbf{u}^T\mathbf{A}(.,\mathbf{P})\mathbf{v}; (\mathbf{u},\mathbf{v}) \in \underset{\|\mathbf{v}\|_p\leq 1, \|\mathbf{u}\|_\infty \leq 1}{\arg\max} \mathbf{u}^T\mathbf{A}(\mathbf{w},\mathbf{P})\mathbf{v}\right\}\right).$$

The same observation can be made for the subgradient with respect to transport matrix $\mathbf{P}$, for which one can apply a Mirror descent algorithm (Nemirovski and Yudin, 1983).

Appendix D

Proofs and Supplementary Material for Chapter 5

This appendix contains the proofs for the different theoretical results of Chapter 5, as well as more details on the experimental part provided for the sake of reproducibility.

D.1 Proofs

Claim (footnote 1 in Chapter 5) $\mathfrak{G}$ defined in (5.5) is a convex compact set.

Proof. Let

$$\begin{aligned}\Phi : \mathbb{R}^{n\times n} &\to \mathbb{R}^{\mathbb{X}\times\mathbb{X}'}\\ \mathbf{M} &\mapsto c^{\mathbf{M}} : (\mathbf{x},\mathbf{x}') \mapsto (\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}')\end{aligned}$$

Notice that the mapping Φ is linear and its domain is a finite dimensional vector space, hence Φ is continuous.
Moreover, we have $\mathfrak{G} = \Phi\left(\mathbb{B}_p^n\right)$ where

$$\mathbb{B}_p^n := \{\mathbf{M} \in \mathbb{R}^{n\times n}; \|\mathbf{M}\|_p \leq 1\}$$

is the Schatten $p-$norm unit ball, which is compact and convex. As a result, $\mathfrak{G}$ is a convex compact set. □

Proposition 5.2.1. *Let $\mathfrak{G}$ be defined as in (5.5) for matrices $\mathbf{M} \in \mathbb{S}_+^{n\times n}$. Then, $\mathfrak{G}$ is a convex compact set of cost functions, and for any $p, q \in [1, +\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$ the following holds:*

1. $\mathrm{RKP}(\Pi, \mathfrak{G}) = \min_{\mathcal{P}\in\Pi} \|\mathbf{V}_\mathcal{P}\|_q$. *In particular, we have:*

$$\mathrm{RKP}(\Pi, \mathfrak{G}) = \begin{cases} W_2^2(\mathcal{D}, \mathcal{D}'), & \text{if } q = 1,\\ \mathcal{S}_1^2(\mathcal{D}, \mathcal{D}'), & \text{if } q = \infty.\end{cases}$$

2. *For any $\mathcal{P} \in \Pi$, $\|\mathbf{M}^*\|_p = 1$ and*

$$\mathbf{M}^* = \underset{\substack{\mathbf{M}\in\mathbb{S}_+^{n\times n}\\ \|\mathbf{M}\|_p\leq 1}}{\arg\max} \langle \mathbf{V}_\mathcal{P}, \mathbf{M}\rangle = \left(\frac{\mathbf{V}_\mathcal{P}}{\|\mathbf{V}_\mathcal{P}\|_q}\right)^{\frac{q}{p}}.$$

In particular, for $p = 2$, one does not need to impose the PSD condition on $\mathbf{M}$, i.e.

$$\mathbf{M}^* = \underset{\|\mathbf{M}\|_2\leq 1}{\arg\max} \langle \mathbf{V}_\mathcal{P}, \mathbf{M}\rangle = \frac{\mathbf{V}_\mathcal{P}}{\|\mathbf{V}_\mathcal{P}\|_2}.$$

Proof. With the notations used to prove that $\mathfrak{G}$ is a convex compact, adding the PSD constraint on $\mathbf{M}$ can be done by considering the image of $\mathbb{B}^n_{p+} := \mathbb{B}^d_p \cap \mathbb{S}^{n\times n}_+$ by mapping Φ. $\mathbb{B}^n_{p+}$ is a convex compact set as the intersection of a convex compact set and a convex cone (the PSD cone). For $\mathcal{P} \in \Pi$, we compute the maximum of $\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}}\left[c^{\mathbf{M}}(\mathbf{x},\mathbf{x}')\right]$ over $\mathbf{M} \in \mathbb{B}^n_{p+}$.

$$\begin{aligned}
&\max_{\mathbf{M}\in\mathbb{B}^n_{p+}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}}\left[c(\mathbf{x},\mathbf{x}')\right] \\
&= \max_{\mathbf{M}\in\mathbb{B}^n_{p+}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}}\left[(\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}')\right] \\
&= \max_{\mathbf{M}\in\mathbb{B}^n_{p+}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}}\left[\mathrm{Tr}\big((\mathbf{x}-\mathbf{x}')(\mathbf{x}-\mathbf{x}')^T\mathbf{M}\big)\right] \\
&= \max_{\mathbf{M}\in\mathbb{B}^n_{p+}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}}\left[\big\langle(\mathbf{x}-\mathbf{x}')(\mathbf{x}-\mathbf{x}')^T,\mathbf{M}\big\rangle\right] \\
&= \max_{\mathbf{M}\in\mathbb{B}^n_{p+}} \langle\mathbf{V}_{\mathcal{P}},\mathbf{M}\rangle\,.
\end{aligned}$$

where we used properties of the trace operator, the linearity of the expectation and the definition of $\mathbf{V}_{\mathcal{P}}$. This maximum is achieved for $\mathbf{M}^*$ verifying $\|\mathbf{M}^*\|_p = \|\mathbf{M}^*\|_p^p = \mathrm{Tr}\{(\mathbf{M}^*)^p\} = 1$. In fact, supposing this is not the case, *i.e.* $\|\mathbf{M}^*\|_p < 1$, then $\mathbf{M}^{**} = \frac{\mathbf{M}^*}{\|\mathbf{M}^*\|}$ verifies $\langle\mathbf{V}_{\mathcal{P}},\mathbf{M}^{**}\rangle > \langle\mathbf{V}_{\mathcal{P}},\mathbf{M}^*\rangle$, which contradicts $\mathbf{M}^*$'s optimality.
Using the equality case of the Hölder inequality for Schatten $p-$norms (Magnus, 1987, Theorem 5), the only PSD matrix achieving this maximum is:

$$\mathbf{M}^* = \left(\frac{\mathbf{V}_{\mathcal{P}}^q}{\mathrm{Tr}\big\{\mathbf{V}_{\mathcal{P}}^q\big\}}\right)^{\frac{1}{p}} = \left(\frac{\mathbf{V}_{\mathcal{P}}}{\|\mathbf{V}_{\mathcal{P}}\|_q}\right)^{\frac{q}{p}}$$

and the value of the maximum is $\|\mathbf{V}_\gamma\|_q$. Taking the minimum over $\mathcal{P} \in \Pi$, we obtain:

$$\min_{\mathcal{P}\in\Pi}\max_{\mathbf{M}\in\mathbb{B}^n_{p+}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}}\left[c(\mathbf{x},\mathbf{x}')\right] = \min_{\mathcal{P}\in\Pi}\|\mathbf{V}_{\mathcal{P}}\|_q.$$

In particular, for $p=\infty$, the corresponding dual norm is $\|\cdot\|_1$, and we have:

$$\begin{aligned}
\min_{\mathcal{P}\in\Pi}\|\mathbf{V}_{\mathcal{P}}\|_1 &= \min_{\mathcal{P}\in\Pi}\mathrm{Tr}\{\mathbf{V}_{\mathcal{P}}\} \\
&= \min_{\mathcal{P}\in\Pi}\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}}\left[\big\|\mathbf{x}-\mathbf{x}'\big\|^2\right] = W_2^2(\mathcal{D},\mathcal{D}').
\end{aligned}$$

and for $p=1$, we the dual norm is the Schatten $\infty-$norm, *i.e.* the maximum singular value, thus corresponding to SRW for the case $k=1$.
As for the last point where $p=2$, we note that $\sup_{\|\mathbf{M}\|_2\leq 1}\langle\mathbf{V}_{\mathcal{P}},\mathbf{M}\rangle$ is achieved, without imposing that $\mathbf{M}$ is PSD, for $\mathbf{M} = \frac{\mathbf{V}_{\mathcal{P}}}{\|\mathbf{V}_{\mathcal{P}}\|}$ (by the equality case of the Cauchy-Schwartz inequality). This matrix is PSD as $\mathbf{V}_{\mathcal{P}}$ is PSD, and has unit norm. This concludes the proof. □

Proposition 5.2.2. *Let $\mathfrak{P}$ be a finite subset of Π. Then, the following holds:*

1. $\mathrm{RKP}(\mathfrak{P},\mathfrak{G}) := \mathrm{RKP}(\mathrm{Conv}(\mathfrak{P}),\mathfrak{G})$ *has a saddle point* $(\mathbf{P}^*,\mathbf{C}^*)$ *verifying:*

$$\langle\mathbf{P}^*,\mathbf{C}^*\rangle = \min_{\mathbf{P}\in\mathrm{Conv}(\mathfrak{P})}\max_{\mathbf{C}\in\mathfrak{G}}\langle\mathbf{P},\mathbf{C}\rangle = \max_{\mathbf{C}\in\mathfrak{G}}\min_{\mathbf{P}\in\mathfrak{P}}\langle\mathbf{P},\mathbf{C}\rangle\,. \tag{5.8}$$

2. $\mathrm{RKP}(\mathfrak{P},\mathfrak{G})$ *is equivalent to*

$$\begin{aligned}
\mathbf{C}^* \in \mathop{\arg\max}_{\mathbf{C}\in\mathfrak{G},\mu\geq 0}\ &\mu \\
s.t.\quad \langle\mathbf{P},\mathbf{C}\rangle \geq\ &\mu,\ \forall\mathbf{P}\in\mathfrak{P}.
\end{aligned} \tag{5.9}$$

3. $\mathbf{P}^* = \sum_{l=1}^{|\mathfrak{P}|} q_l \mathbf{P}_l$*, where* $\mathfrak{Q} = \{q_l\}_{l=1}^{|\mathfrak{P}|}, \sum_i q_i = 1$*, are dual variables of Problem* (5.9)*.*

Proof. Since the set $\mathfrak{P}$ is finite, Conv $(\mathfrak{P})$ is a convex compact set. Also, by definition, $\mathfrak{G}$ is a convex compact set. Moreover, we note that for any $(\mathbf{P}, \mathbf{C}) \in \Pi \times \mathfrak{G}$, the functions $\langle \mathbf{P}, \cdot \rangle$ and $\langle \cdot, \mathbf{C} \rangle$ are linear. By applying Sion's min-max theorem (Sion, 1958), problem RKP$(\mathfrak{P}, \mathfrak{G})$:= RKP(Conv $(\mathfrak{P})$, $\mathfrak{G}$) has at least a saddle point, and any saddle point $(\mathbf{P}^s, \mathbf{C}^s)$ verifies:

$$\langle \mathbf{P}^s, \mathbf{C}^s \rangle = \min_{\mathbf{P} \in \text{Conv}(\mathfrak{P})} \max_{\mathbf{C} \in \mathfrak{G}} \langle \mathbf{P}, \mathbf{C} \rangle = \max_{\mathbf{C} \in \mathfrak{G}} \min_{\mathbf{P} \in \text{Conv}(\mathfrak{P})} \langle \mathbf{P}, \mathbf{C} \rangle . \tag{D.1}$$

However, for any fixed $\mathbf{C} \in \mathfrak{G}$, the linearity of $\langle \cdot, \mathbf{C} \rangle$ implies that its minimum on Conv $(\mathfrak{P})$ is achieved on one of its vertices. More formally, we have:

$$\min_{\mathbf{P} \in \text{Conv}(\mathfrak{P})} \langle \mathbf{P}, \mathbf{C} \rangle = \min_{\mathbf{P} \in \mathfrak{P}} \langle \mathbf{P}, \mathbf{C} \rangle , \forall \mathbf{C} \in \mathfrak{G} \Rightarrow \max_{\mathbf{C} \in \mathfrak{G}} \min_{\mathbf{P} \in \text{Conv}(\mathfrak{P})} \langle \mathbf{P}, \mathbf{C} \rangle = \max_{\mathbf{C} \in \mathfrak{G}} \min_{\mathbf{P} \in \mathfrak{P}} \langle \mathbf{P}, \mathbf{C} \rangle . \tag{D.2}$$

Combining (D.1) and (D.2) yields Equation (5.8).
Moreover, by the saddle point's definition, we have: $\mathbf{C}^* \in \arg\max_{\mathbf{C} \in \mathfrak{G}} \min_{\mathbf{P} \in \mathfrak{P}} \langle \mathbf{P}, \mathbf{C} \rangle$. Using the fact that $\mathfrak{P}$ is finite, we obtain the equivalent Problem (5.9). What is left is computing $\mathbf{P}^*$'s value. To this end, let us introduce $I_{\mathfrak{G}}$, the convex indicator function of set $\mathfrak{G}$, defined by:

$$\begin{aligned} I_{\mathfrak{G}} : \mathbf{C} \mapsto 0 \quad & \text{if} \quad \mathbf{C} \in \mathfrak{G} \\ + \infty \quad & \text{otherwise} \end{aligned}$$

Also, notice that μ is nonnegative even without imposing this condition. In fact, assuming that the cost matrices in $\mathfrak{G}$ have positive values (which is the case in practice), we have $\min_{\mathbf{P} \in \mathfrak{P}} \langle \mathbf{P}, \mathbf{C} \rangle \geq 0$, for all $\mathbf{C} \in \mathfrak{G}$. If μ^*, the value of μ at the solution was negative, its maximality contradicts the condition $\min_{\mathbf{P} \in \mathfrak{P}} \langle \mathbf{P}, \mathbf{C} \rangle \geq 0$. Hence, Problem (5.9) is equivalent to the following:

$$\begin{aligned} \max_{\substack{\mathbf{C} \in \mathbb{R}^{m \times n} \\ \mu \in \mathbb{R}}} \; & \mu - I_{\mathfrak{G}}(\mathbf{C}) \\ \text{s.t} \; & \langle \mathbf{P}, \mathbf{C} \rangle \geq \mu \quad \forall \mathbf{P} \in \mathfrak{P}. \end{aligned} \tag{D.3}$$

The Lagrangian $\mathfrak{L}$ of the previous problem is given by:

$$\mathfrak{L}(\mathbf{q}, \mathbf{C}, \mu) = \mu - I_{\mathfrak{G}}(\mathbf{C}) + \sum_{l=1}^{|\mathfrak{P}|} q_l (\langle \mathbf{P}_l, \mathbf{C} \rangle - \mu), \tag{D.4}$$

where l indexes the finite set of matrices $\mathfrak{P}$ and $q_l \geq 0$ for all $l \in \{1, ..., |\mathfrak{P}|\}$ denote the dual variables of the constraints, and $\mathbf{q} = (q_1, ..., q_{|\mathfrak{P}|})$. A known optimization result (Boyd and Vandenberghe, 2004, Section 5.4.2) implies that that the solution to the primal, $(\mathbf{C}^*, \mu^*)$ and the solution to the dual, $\mathbf{q}^* = (q_1^*, ..., q_l^*)$ form a saddle point of the Lagrangian, which implies:

$$\mathfrak{L}(\mathbf{q}^*, \mathbf{C}^*, \mu^*) = \max_{\mathbf{C}, \mu} \mathfrak{L}(\mathbf{q}^*, \mathbf{C}, \mu) \tag{D.5}$$

Deriving the Lagrangian with respect to μ yields:

$$\sum_l q_l^* = 1. \tag{D.6}$$

In addition to this condition, knowing that the value of the Lagrangian is finite at the solution, we have $I_{\mathfrak{G}}(\mathbf{C}^*) = 0$. Substituting the last two conditions in Equation (D.5) yields:

$$\mathfrak{L}(\mathbf{q}^*, \mathbf{C}^*, \mu^*) = \langle \mathbf{P}^*, \mathbf{C}^* \rangle$$

$$= \max_{\mathbf{C}\in\mathbb{R}^{m\times m'}} \langle \mathbf{P}^*, \mathbf{C}\rangle - I_{\mathfrak{G}}(\mathbf{C})$$
$$= \max_{\mathbf{C}\in\mathfrak{G}} \langle \mathbf{P}^*, \mathbf{C}\rangle \tag{D.7}$$

where $\mathbf{P}^*$ is defined as in the proposition. Also, Equation (D.6) implies that there is at least one $l' \in \{1, ..., |\mathfrak{P}|\}$ verifying $q_{l'} > 0$, and hence its associated constraint from Problem (D.3) is saturated, *i.e.*

$$\mu^* = \langle \mathbf{P}_{l'}, \mathbf{C}^*\rangle = \min_{\mathbf{P}\in\mathfrak{P}} \langle \mathbf{P}, \mathbf{C}^*\rangle = \min_{\mathbf{P}\in \mathrm{Conv}(\mathfrak{P})} \langle \mathbf{P}, \mathbf{C}^*\rangle \tag{D.8}$$

Moreover, by the Lagrangian's definition, we have $\mu^* = \mathfrak{L}(\mathbf{q}^*, \mathbf{C}^*, \mu^*)$. This latter equation combined with (D.8) and (D.7) yields:

$$\langle \mathbf{P}^*, \mathbf{C}^*\rangle = \max_{\mathbf{C}\in\mathfrak{G}} \langle \mathbf{P}^*, \mathbf{C}\rangle = \min_{\mathbf{P}\in \mathrm{Conv}(\mathfrak{P})} \langle \mathbf{P}, \mathbf{C}^*\rangle \tag{D.9}$$

i.e. , $(\mathbf{P}^*, \mathbf{C}^*)$ is a saddle point of $\mathrm{RKP}(\mathfrak{P}, \mathfrak{G})$. □

Algorithm 1 Cutting set method for $\mathrm{RKP}(\Pi, \mathfrak{G})$ with constraint elimination

Input: T, $\mathfrak{G}$, $\mathfrak{P}_0 \subset \Pi$, τ_1, τ_2
$t, \nu_{-1} \leftarrow 0$
$\varepsilon, \mu_{-1} \leftarrow \infty$
while $t < T$ and $\varepsilon_t > \tau_1$ and $\frac{\mu_{t-1}-\mu_t}{\mu_{t-1}} > \tau_1^2$ **do**
  Solve (5.9) to obtain $(\mu_t, \mathbf{C}_t), \mathfrak{Q}$
  for j in $\{0, ..., |\mathfrak{P}_t| - 1\}$ **do**
    if $q_j \leq \tau_2$ **then**
      $\mathfrak{P}_t \leftarrow \mathfrak{P}_t \setminus \{\mathbf{P}_j\}$
      $\mathfrak{Q} \leftarrow \mathfrak{Q} \setminus \{\mathbf{q}_j\}$
    end if
  end for
  Find $\mathbf{P}_t \in \arg\min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C}_t\rangle$
  $\nu_t \leftarrow \max(\nu_{t-1}, \langle \mathbf{P}_t, \mathbf{C}_t\rangle)$
  $\varepsilon_t \leftarrow \mu_t - \nu_t$
  $\mathfrak{P}_{t+1} = \mathfrak{P}_t \cup \{\mathbf{P}_t\}$
end while
return $\sum_{l=0}^{|\mathfrak{P}_t|-1} q_l \mathbf{P}_l$, $\mathbf{C}_t$

Proposition 5.2.3. *Let T be the number of iterations required by Algorithm 1 to reach error $\varepsilon_T \leq \tau_1$. Then*

$$T \leq \left(\frac{\mathrm{diam}_\infty(\mathfrak{G}) + \mathrm{RKP}(\mathfrak{P}_0, \mathfrak{G})}{2\tau_1} + 1 \right)^{\dim(\mathfrak{G})+1}$$

where

$$\mathrm{diam}_\infty(\mathfrak{G}) := \sup_{\substack{\mathbf{C}^1, \mathbf{C}^2 \in \mathfrak{G} \\ i,j}} \left| (\mathbf{C}^1)_{ij} - (\mathbf{C}^2)_{ij} \right|,$$

and $\dim(\mathfrak{G})$ *is the dimension of the affine hull of* $\mathfrak{G}$*. Also, we have*

$$0 \leq \mathrm{RKP}(\mathfrak{P}_t, \mathfrak{G}) - \mathrm{RKP}(\Pi, \mathfrak{G}) \leq \varepsilon_t.$$

Proof. In this proof, we use the notation $\|\mathbf{A}\|_1 = \sum_{ij} |(\mathbf{A})_{ij}|$ and $\|\mathbf{A}\|_\infty = \sup_{ij} |(\mathbf{A})_{ij}|$. We note that these notations are only used in this proof and do not apply to the rest of the chapter, as they do not correspond to the Schatten $1-$norm and $\infty-$norm.

We apply the result of (Mutapcic and Boyd, 2009, Section 5.2) to our case. To this end, since our nominal problem corresponds to $\mathfrak{P}_0$, we define its feasible set $\mathfrak{F}_0$ as:

$$\mathfrak{F}_0 = \{(\mu, \mathbf{C}) \in \mathbb{R}_+ \times \mathfrak{G} | \mu \leq \min_{\mathbf{P} \in \mathfrak{P}_0} \langle \mathbf{P}, \mathbf{C} \rangle\}.$$

Also, we define

$$\|(\mu, \mathbf{C})\|_\infty := |\mu| + \|\mathbf{C}\|_\infty. \tag{D.10}$$

For every $(\mu_1, \mathbf{C}_1), (\mu_2, \mathbf{C}_2) \in \mathfrak{F}_0$ and for every constraint, *i.e.* , for every $\mathbf{P} \in \mathfrak{P}_0$, we have:

$$\begin{aligned}
&|(\langle \mathbf{P}, \mathbf{C}_1 \rangle - \mu_1) - (\langle \mathbf{P}, \mathbf{C}_2 \rangle - \mu_2)| \\
&\qquad \leq |\langle \mathbf{P}, \mathbf{C}_1 \rangle - \langle \mathbf{P}, \mathbf{C}_2 \rangle| + |\mu_1 - \mu_2| && \text{(D.11)} \\
&\qquad \leq \|\mathbf{P}\|_1 \|\mathbf{C}_1 - \mathbf{C}_2\|_\infty + |\mu_1 - \mu_2| && \text{(D.12)} \\
&\qquad = \|\mathbf{C}_1 - \mathbf{C}_2\|_\infty + |\mu_1 - \mu_2| && \text{(D.13)} \\
&\qquad = \|(\mu_1, \mathbf{C}_1) - (\mu_2, \mathbf{C}_2)\|_\infty && \text{(D.14)}
\end{aligned}$$

where $(\mu_1, \mathbf{C}_1) - (\mu_2, \mathbf{C}_2) := (\mu_1 - \mu_2, \mathbf{C}_1 - \mathbf{C}_2)$. (D.11) is due to the triangle inequality, followed by the Hölder inequality to obtain (D.12). Then, since $\mathfrak{P}_0 \subset \Pi$ and any matrix in Π has the sum of its entries equal to 1, we obtain (D.13). Lastly, we used definition (D.10) to obtain (D.14).

To establish the bound as done in Mutapcic and Boyd (2009), we also need to find the radius R of a ball that contains the feasible set $\mathfrak{F}_0$, and we consider the affine hull of $\mathfrak{G}$ instead of $\mathbb{R}^{m \times m'}$ as the space containing $\mathfrak{G}$. It is then sufficient to bound the diameter of $\mathfrak{F}_0$, denoted $\operatorname{diam}_\infty(\mathfrak{F}_0)$ and to take half of the bound for R. To this end, for any $(\mu_1, \mathbf{C}_1), (\mu_2, \mathbf{C}_2) \in \mathfrak{F}_0$,

$$\begin{aligned}
\|(\mu_1, \mathbf{C}_1) - (\mu_2, \mathbf{C}_2)\|_\infty &= \|(\mu_1 - \mu_2, \mathbf{C}_1 - \mathbf{C}_2)\|_\infty \\
&= \|\mathbf{C}_1 - \mathbf{C}_2\|_\infty + |\mu_1 - \mu_2| \\
&\leq \operatorname{diam}_\infty(\mathfrak{G}) + |\mu_1 - \mu_2|.
\end{aligned}$$

By $\mathfrak{F}_0$'s definition, we have for $j \in \{1, 2\}$:

$$0 \leq \mu_j \leq \min_{\mathbf{P} \in \mathfrak{P}_0} \langle \mathbf{P}, \mathbf{C}_j \rangle \leq \mathrm{RKP}(\mathfrak{P}_0, \mathfrak{G})$$

Hence, $|\mu_1 - \mu_2| \leq \mathrm{RKP}(\mathfrak{P}_0, \mathfrak{G})$. Taking the supremum over $(\mu_1, \mathbf{C}_1), (\mu_2, \mathbf{C}_2) \in \mathfrak{F}_0$ (the definition of a diameter), we obtain:

$$\operatorname{diam}_\infty(\mathfrak{F}_0) \leq \operatorname{diam}_\infty(\mathfrak{G}) + \mathrm{RKP}(\mathfrak{P}_0, \mathfrak{G}).$$

We can then set radius R as half of the previous upper bound, leading to the bound on the number of iterations T.
For the second result of the proposition, let $t \geq 0$, we have for any $0 \leq t' \leq t$:

$$\nu_t = \max_{0 \leq t' \leq t} \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C}_t \rangle \leq \max_{\mathbf{C} \in \mathfrak{G}} \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C} \rangle = \mathrm{RKP}(\Pi, \mathfrak{G}) \leq \mathrm{RKP}(\mathfrak{P}_t, \mathfrak{G}) = \mu_t$$

where the left inequality is due to taking the maximum over $\mathfrak{G}$, while the right one is due to the set inclusion $\mathfrak{P}_t \subset \Pi$. Thus,

$$0 \leq \mathrm{RKP}(\mathfrak{P}_t, \mathfrak{G}) - \mathrm{RKP}(\Pi, \mathfrak{G}) \leq \mathrm{RKP}(\mathfrak{P}_t, \mathfrak{G}) - \nu_t = \mu_t - \nu_t = \varepsilon_t,$$

which concludes the proof. □

To prove Proposition 5.2.4, we first prove the following lemma.

Lemma D.1.1. *Let c and d be two positive integers. The dual of the linear program*

$$\begin{aligned} &\max_{\mathbf{p}\in\Delta_d,\mu\geq 0} \mu \\ &s.t.\ \mathbf{G}\mathbf{p} \geq \mu \mathbf{1}_c, \end{aligned} \tag{D.15}$$

is the linear program

$$\begin{aligned} &\min_{\mathbf{q}\in\Delta_c,\eta\geq 0} \eta \\ &s.t. \mathbf{G}^T\mathbf{q} \leq \eta \mathbf{1}_d, \end{aligned}$$

Proof. We will transform (D.15) to a standard LP formulation. To this end, let $\mathbf{v} = (p_1, ..., p_d, \mu)$, i.e the concatenation of $\mathbf{p}$ and μ. Also, we transform the equality condition $\mathbf{1}_d^T\mathbf{p} = 1$ into the two inequalities $\mathbf{1}_d^T\mathbf{p} \leq 1$ and $-\mathbf{1}_d^T\mathbf{p} \leq -1$. We construct the following matrix:

$$\mathbf{F} = \begin{bmatrix} -\mathbf{G} & \mathbf{1}_c \\ \mathbf{1}_d^T & 0 \\ -\mathbf{1}_d^T & 0 \end{bmatrix}$$

Having $c+2$ rows and $d+1$ columns. Then, (D.15) can be re-written under the standard form:

$$\begin{aligned} \max \quad & \mathbf{e}_{d+1}^T\mathbf{v} \\ \text{s.t} \quad & \mathbf{F}\mathbf{v} \leq \mathbf{e}_{c+1} - \mathbf{e}_{c+2} \\ & \mathbf{v} \geq 0 \end{aligned}$$

where $\mathbf{e}_i$ denotes the vectors of $\mathbb{R}^{d+1}$'s canonical basis. This latter problem has the following dual:

$$\begin{aligned} \min \quad & (\mathbf{e}_{c+1} - \mathbf{e}_{c+2})^T\mathbf{w} \\ \text{s.t} \quad & \mathbf{F}^T\mathbf{w} \geq \mathbf{e}_{d+1} \\ & \mathbf{w} \geq 0 \end{aligned}$$

Using the fact that

$$\mathbf{F}^T = \begin{bmatrix} -\mathbf{G}^T & \mathbf{1}_d & -\mathbf{1}_d \\ \mathbf{1}_c^T & 0 & 0 \end{bmatrix}$$

and denoting $\mathbf{w} = (q_1, ..., q_c, \eta_1, \eta_2)$, and $\mathbf{q} = (q_1, ..., q_c)$, the dual is written:

$$\begin{aligned} \min \quad & \eta_1 - \eta_2 && \text{(D.16)} \\ \text{s.t} \quad & \mathbf{G}^T\mathbf{q} \leq (\eta_1 - \eta_2)\mathbf{1}_d && \text{(D.17)} \\ & \mathbf{1}_c^T\mathbf{q} \geq 1 && \text{(D.18)} \\ & \mathbf{q} \geq 0 && \text{(D.19)} \\ & \eta_1, \eta_2 \geq 0 && \text{(D.20)} \end{aligned}$$

Setting $\eta = \eta_1 - \eta_2$, from (D.17) and the fact that $\mathbf{G}$ has positive elements (Frobenius products between cost matrices and transport matrices), we have $\eta \geq 0$. Also, for η^*, $\mathbf{q}^*$ the solution of the dual, we necessarily have $\mathbf{1}_c^T\mathbf{q}^* = 1$. In fact, assuming that $\mathbf{1}_c^T\mathbf{q}^* > 1$ and dividing (D.17) by $\mathbf{1}_c^T\mathbf{q}^*$, we see that $\eta^{**}, \mathbf{q}^{**}$ defined by $\mathbf{q}^{**} = \frac{\mathbf{q}^*}{\mathbf{1}_c^T\mathbf{q}^*}$ and $\eta^{**} = \frac{\eta^*}{\mathbf{1}_c^T\mathbf{q}^*}$ verify all the constraints, whereas $\eta^{**} < \eta^*$. This latter inequality contradicts the minimality of η. Hence, the dual formulation is proven. □

Proposition 5.2.4 (Finite set $\mathfrak{G}$)**.** *Let $\mathfrak{G} = \text{Conv}(\{\mathbf{C}_1, ..., \mathbf{C}_K\})$. Then, for $t \geq 0$, solving the problem given in* (5.9) *over $\mathfrak{P}_t \times \mathfrak{G}$ is equivalent to the following linear program*

$$\begin{aligned} \min_{\mathbf{p} \in \mathbb{R}_+^K} \quad & \mathbf{1}_K^T \mathbf{p} \\ s.t. \quad & \mathbf{G}\mathbf{p} \geq \mathbf{1}_{|\mathfrak{P}_t|}, \end{aligned} \tag{5.11}$$

where $\mathbf{G} \in \mathbb{R}^{|\mathfrak{P}_t| \times K}$ with $\mathbf{G}_{kl} = \langle \mathbf{P}_k, \mathbf{C}_l \rangle$. Moreover, the saddle point $(\mathbf{P}^*, \mathbf{C}^*)$ is given by

$$\mathbf{C}^* = \frac{\sum_{k=1}^K p_k^* \mathbf{C}_k}{\sum_{k=1}^K p_k^*}, \qquad \mathbf{P}^* = \frac{\sum_{l=1}^{|\mathfrak{P}_t|} q_l^* \mathbf{P}_l}{\sum_{l=1}^{|\mathfrak{P}_t|} q_l^*},$$

where $\mathbf{p}^* = (p_1, \cdots, p_K)$ and $\mathbf{q}^* = (q_1, \cdots, q_{|\mathfrak{P}_t|})$ are optimal solutions of (5.11) *and its dual.*

Proof. Since $\mathfrak{G}$ is the convex hull of matrices $\{\mathbf{C}_1, ..., \mathbf{C}_K\}$, i.e the set of their convex combinations, problem (5.9) can be formulated as follows:

$$\begin{aligned} &\max_{\mathbf{p} \in \Delta_K, \mu \geq 0} \mu \\ &\text{s.t. } \mu \leq \sum_l p_l \langle \mathbf{P}_k, \mathbf{C}_l \rangle \qquad \forall 1 \leq k \leq d \end{aligned}$$

Let $\mathbf{G}_{kl}$ be the matrix whose elements are: $\mathbf{G}_{kl} = \langle \mathbf{P}_k, \mathbf{C}_l \rangle$. The previous problem can be re-written:

$$\begin{aligned} &\max_{\mathbf{p} \in \Delta_K, \mu \geq 0} \mu \\ &\text{s.t. } \mathbf{G}\mathbf{p} \geq \mu \mathbf{1}_{|\mathfrak{P}_t|}, \end{aligned} \tag{D.21}$$

Since the probability simplex Δ_K can be expressed as:

$$\Delta_K = \left\{ \frac{\mathbf{p}}{\mathbf{1}_K^T \mathbf{p}}; \mathbf{p} \in \mathbb{R}_+^K \setminus \{0\} \right\}$$

the previous problem is equivalent to

$$\begin{aligned} &\max_{\mathbf{p} \in \mathbb{R}_+^K, \mu \geq 0} \mu \\ &\text{s.t. } \mathbf{G}\mathbf{p} \geq \mu \mathbf{1}_K^T \mathbf{p} \mathbf{1}_{|\mathfrak{P}_t|}. \end{aligned}$$

Setting $\mu \mathbf{1}_K^T \mathbf{p} = 1$ (same technique used to derive primal SVM optimization problem as a constrained norm minimization problem) proves formulation (5.11). Also, from the change of variables that we made on $\mathbf{p}$, we obtain

$$\mathbf{C}^* = \frac{\sum_{k=1}^K p_k^* \mathbf{C}_k}{\sum_{k=1}^K p_k^*},$$

where $\mathbf{p}^*$ is the solution of Problem (5.11).

Now we focus on the second part of the proof, to obtain the expression of $\mathbf{P}^*$, the other component of the saddle point. By the result in Lemma D.1.1, denoting $\tilde{\mathbf{q}}^*$ the dual variables of Problem (D.21), $\tilde{\mathbf{q}}^*$ is a solution to the following dual problem:

$$\begin{aligned} &\min_{\mathbf{q} \in \Delta_{|\mathfrak{P}_t|}, \nu \geq 0} \nu \\ &\text{subject to } \mathbf{G}^T \mathbf{q} \leq \nu \mathbf{1}_d, \end{aligned} \tag{D.22}$$

By the same argument used to obtain the equivalent formulation (5.11), Problem (D.22) is equivalent to:

$$\begin{aligned} &\max_{\mathbf{q}\in\mathbb{R}_+^{|\mathfrak{P}_t|}} \mathbf{1}_{|\mathfrak{P}_t|}^T\mathbf{q} \\ &\text{s.t. } \mathbf{G}^T\mathbf{q} \leq \mathbf{1}_d, \end{aligned} \tag{D.23}$$

where the components of the solution $\tilde{\mathbf{q}}^*$ by normalizing solution $\mathbf{q}^*$ of the previous problem, which yields the expression of $\mathbf{P}^*$. Finally, it is sufficient to notice that Problems (D.23) and (5.11) are each the dual of the other, to conclude the proof. □

For additional details on the experimental evaluations kindly proceed to the following page.

D.2 Experimental Evaluations

In this section, we add the details needed to reproduce the experiments from Chapter 5 using the code submitted in the supplementary material. We also provide more experimental results for the considered evaluation scenarios and full-size figures presented in a reduced size in Chapter 5. For all of the experiments, threshold τ_2 used for constraint elimination is set to 10^{-12}.

Section 4.1: Convergence and Execution Time

Convergence curves for the first two plots are obtained for threshold value $\tau_1 = 0$, as we are interested in the evolution of ε_t with the iterations. The maximum number of iterations T is set to 100. As for the right figure, we set the maximum number of iterations to 1000 and τ_1 to 10^{-8}, and we use MOSEK solver to solve the LP formulation, for which we set all tolerance values to 10^{-8}.

Section 4.2: Hypercube

We set T to 10, $\mathfrak{P}_0$ is set to the uniform distribution, *i.e.* $\mathfrak{P}_0 = \{\mathbf{r}\mathbf{c}^T\}$ and $\tau_1 = 10^{-8}$. The experiment is reproduced 100 times.

Section 4.3: Stability and Noise Sensitivity

The parameters used in this experiment for all additional data sets are the same as for the MNIST 0-to-1 dataset and the two Gaussians. The maximum number of iterations of the cutting set method, T is set to 10. $\mathfrak{P}_0$ is set to the uniform distribution, $\tau_1 = 10^{-20}$. The Mahalanobis ball has the radius $r = 0.01$. The 50 cost matrices are created with random Mahalanobis projections and different norms taking values in $(2, 3, 4, 5, 10)$. We also add the cost matrix associated with the squared Euclidean distance. Each cost matrix is divided by its Frobenius norm. The noise sensitivity is computed over 200 runs.

In all examples of Figure D.1, the sensitivity to noise is correlated to the stability of the cost matrix. The cost matrix associated with the squared Euclidean distance is often stable and robust to noise which is predictable as it is the most used distance in OT. However, it is never the best cost matrix in terms of our notion of stability.

Section 4.4: Color Transfer

We use the same setting as above with the following parameters: T is set to 200, $\tau_1 = 10^{-8}$, $r = 0.001$ and we divide each cost matrix element-wise by its corresponding transport cost. Below, we first provide images from Chapter 5 in a bigger size in order to see more fine-grained details.

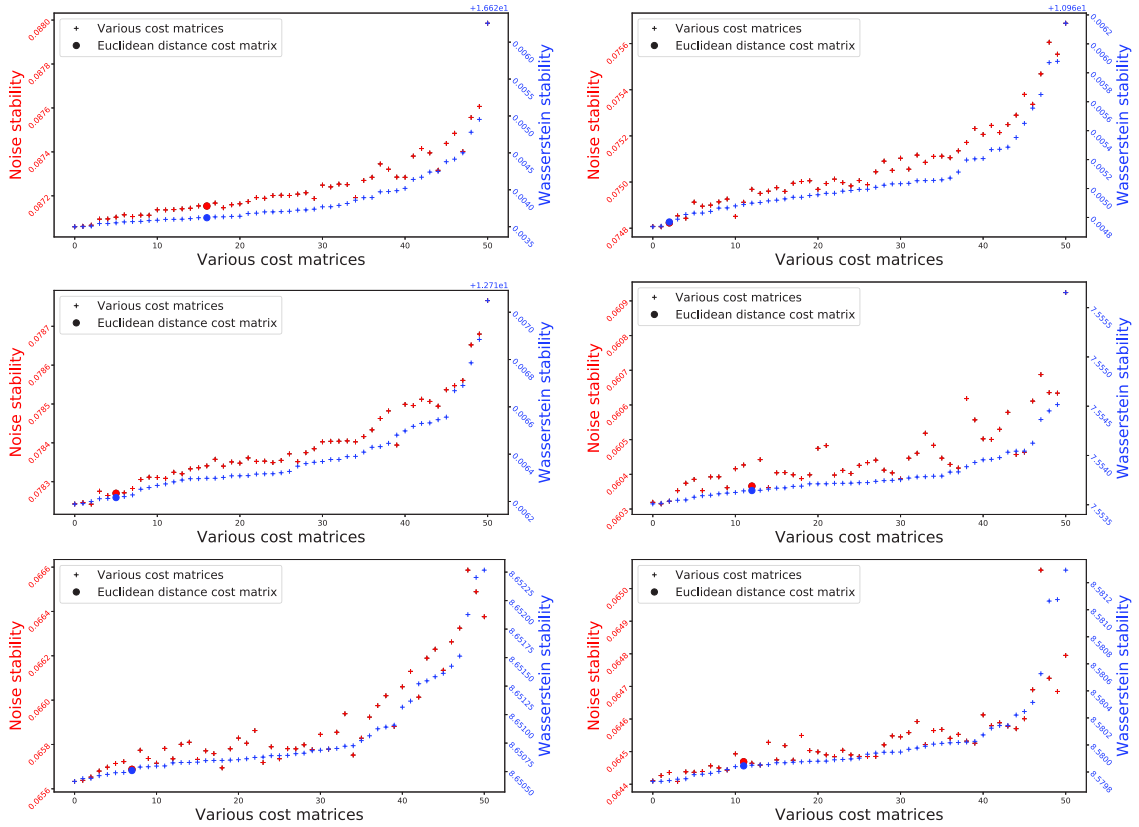


Figure D.1: Left to right, top to bottom: Gaussians, MNIST 0-to-1, MNIST 3-to-0, MNIST 6-to-5, MNIST 7-to-1, MNIST 7-to-4 data sets. Y-axis (left) is the difference between the OT cost with $\mathbf{C}_i$ and $\mathbf{C}_i + \mathbf{E}^{\mathbf{M}}$. Y-axis (right), the Wasserstein stability defined in Section 3.5. Each column is a different cost matrix, the matrices are ordered by the Wasserstein stability.

Figure D.3 presents additional visualizations for a new pair of images. The obtained results are in line with experiments shown in Chapter 5 and exhibit similar behaviour.

Figure D.2: Top row: Original images of ocean sunset and ocean sky. Middle row: (left) most stable cost matrix, (right) squared Euclidean based cost matrix. Bottom row: (left) least stable Mahalanobis cost matrix, (right) least stable cost matrix. Notice the quality difference between the most stable matrix and the squared Euclidean based one in the area just under the cloud.

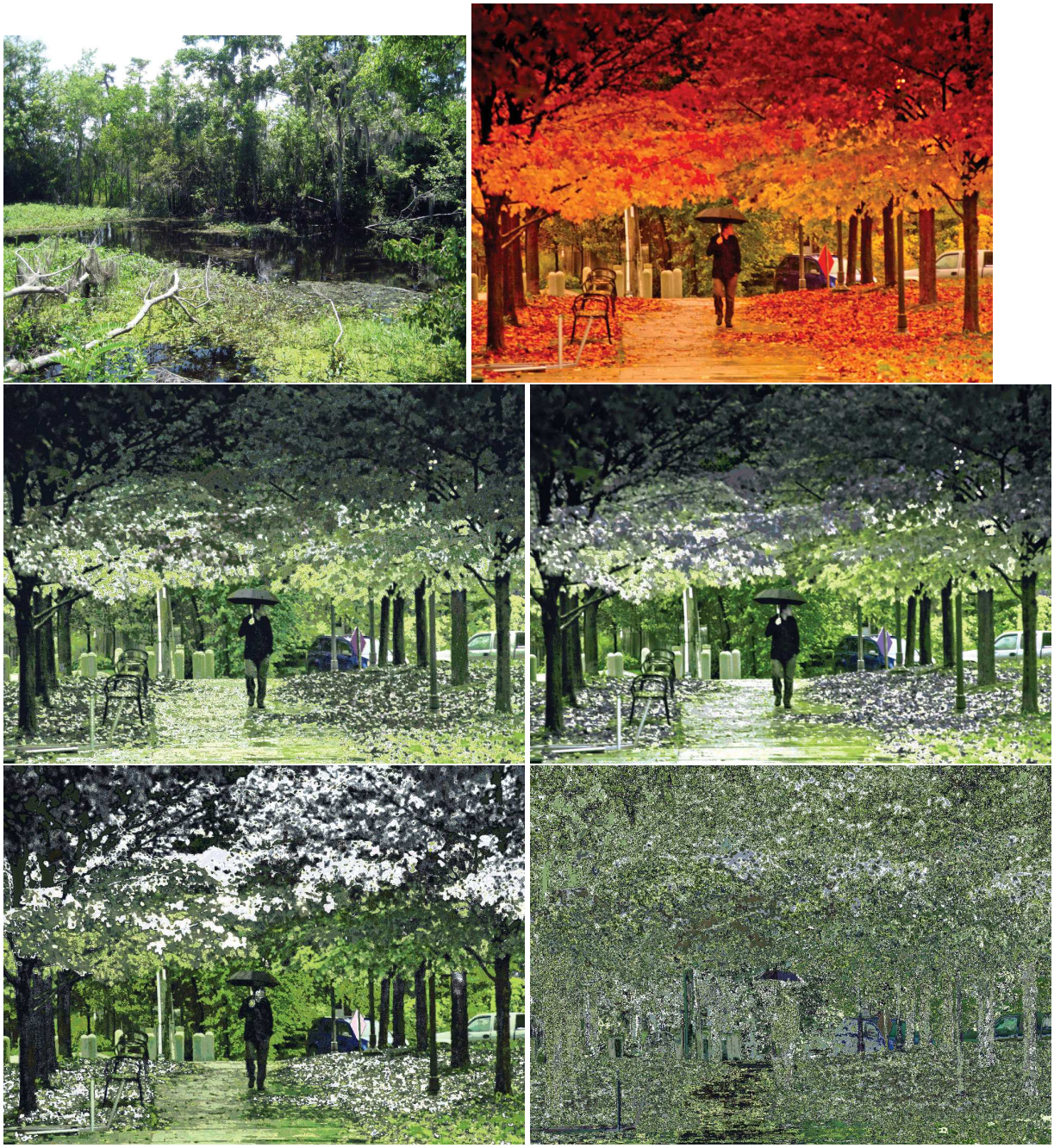


Figure D.3: Top row: Original images of woods and autumn. Middle row: (left) most stable cost matrix, (right) Euclidean based cost matrix. Bottom row: (left) least stable Mahalanobis cost matrix, (right) least stable cost matrix.

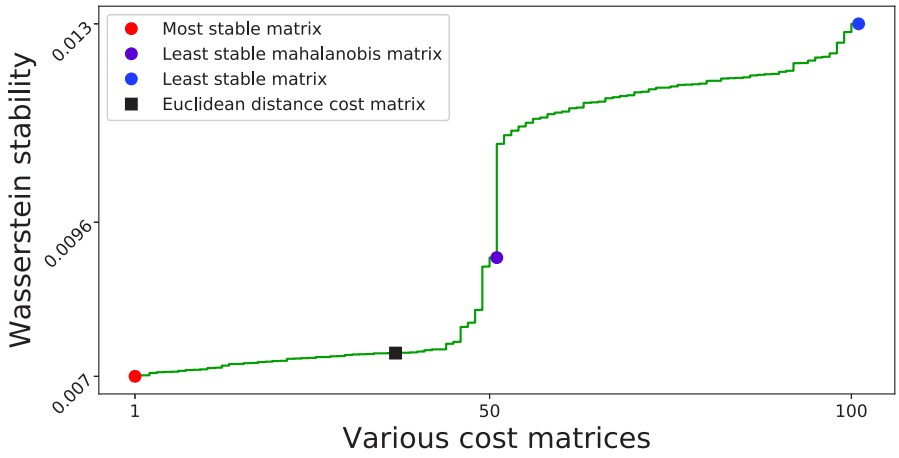


Figure D.4: Cost matrices sorted by the Wasserstein stability. The first 50 are Mahalanobis cost matrices, while the last 50 are random cost matrices.

Appendix E

Learning a Bilinear Similarity Function via Regression

Based on our submission to Journal track of ECML-PKDD 2020. This chapter is an L_2 regression version of Bellet et al. (2012).

Abstract We consider the general $(\epsilon, \gamma, \tau)-$good similarity learning framework that includes several large families of similarity learning approaches and show how to modify it to admit a ridge regression alike closed-form solution. We derive theoretical guarantees for the proposed approach highlighting several properties of the obtained solution as well as its generalization bounds with both data- and problem-dependent terms.

Introduction

In this chapter, we investigate learning of (ϵ, γ, τ)-good similarity functions under a new angle by introducing a different loss function to the original definition of Balcan et al. (2008a), presented in Section 1.4. This novel formulation allows us to derive a closed-form solution of the goodness maximization procedure proposed by Bellet et al. (2012), resembling that of a ridge regression problem. We further use the obtained expression to study its theoretical properties, such as the bound on the norm of the optimal similarity matrix and the value of the objective function corresponding to it. This latter result allows us to further characterize the obtained solution in terms of the original definition of the (ϵ, γ, τ)-good learning framework of Balcan et al. (2008a). Moreover, we provide novel data-dependent and dimension independent generalization bounds for our approach and show that it offers a remarkable computational speed-up on benchmark datasets.

The rest of the chapter is set as follows. Section E.1 presents the main definitions, notations, and preliminary knowledge related to $(\epsilon, \gamma, \tau)-$good similarity functions that we use later. Section E.2 is dedicated to a detailed derivation of the optimal solution to the introduced optimization problem with a quadratic loss function. This result is further used to relate our proposed approach to the original similarity learning framework of Balcan et al. (2008a). Our algorithm's generalization guarantees are presented in Section E.3.

Finally, we conclude the chapter in Section E.5 and give several possible future research directions.

E.1 Preliminary Knowledge

Given a feature space $\mathbb{X} \subseteq \mathbb{R}^n$ and a label space $\mathbb{Y} = \{-1, 1\}$, we assume having access to a labeled data sample $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\} \sim \mathcal{D}^m$. Throughout the chapter, and unless specified otherwise, a matrix's Frobenius norm will be denoted $\|.\|$ (instead of

$\|.\|_2$), and its largest and smallest singular values will be denoted respectively by $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$. As in the work of Balcan et al. (2008a), a similarity function is defined as $K : \mathbb{X} \times \mathbb{X} \to [-1, 1]$ (Section 1.4). Using these notations, $(\epsilon, \gamma, \tau)-$good similarity functions can be defined in two different ways based on either the margin violation loss or with respect to the hinge loss (Definitions 1.4.1 and 1.4.2). For the sake of simplicity, we group both of them in the following definition related to an arbitrary loss function l.

Definition E.1.1 (Balcan et al. 2008a)**.** *A similarity function K is (ϵ, γ, τ)-good in l loss for problem (distribution) $\mathcal{D}$ if there exists a (probabilistic) indicator function R of a set of "reasonable points" (also called landmarks) such that:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[l\left(\mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}} \left[K(\mathbf{x}, \mathbf{x}')y' | R(\mathbf{x}') = 1 \right], y \right) \right] \leq \epsilon, \tag{E.1}$$

$$\mathop{\mathbb{P}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}} \left[R(\mathbf{x}') = 1 \right] \geq \tau. \tag{E.2}$$

Note that the goodness in margin violation loss and in hinge loss (Definition 1.4.1,Definition 1.4.2) correspond respectively to setting l to the margin violation loss or to the hinge loss.

E.2 Learning a Good Bilinear Similarity in a Closed Form

In this section, we present several of our main contributions. We start by formally introducing our objective function and deriving a closed-form solution to it. We use it further to derive a bound on the norm of the optimal similarity function as well as the optimal value of the considered objective function. We end this section by a brief comparison to other existing similarity learning methods in terms of the obtained results and the considered problem setups.

E.2.1 Problem Setup

Let us consider a bilinear similarity function $K_{\mathbf{A}} : (\mathbf{x}, \mathbf{x}') \mapsto \mathbf{x}^T \mathbf{A} \mathbf{x}'$ parametrized by a matrix $\mathbf{A}$, such that $\|\mathbf{A}\| \leq 1$ and $\|\mathbf{x}\|_2 \leq 1$, $\forall \mathbf{x} \in \mathbb{X}$. This last assumption guarantees that $K_{\mathbf{A}}$ takes its values in $[-1, 1]$. We now review the definition of $(\epsilon, \gamma, \tau)-$good similarity functions with the quadratic loss function l_q defined via $\ell_q : t \mapsto (t - s)^2$ where $s > 0$ is a hyperparameter[1]. Let us further denote by $\mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}})$ the l_q-risk defined with this loss function so that:

$$\mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}}) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[l_q\left(\mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}} \left[K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}')y' | R(\mathbf{x}') = 1 \right], y \right) \right].$$

As $\mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}})$ is upper-bounded by ϵ in Equation (E.1), a natural goal is to search for a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ solving the following optimization problem:

$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{n\times n} \\ \|\mathbf{A}\|\leq 1}} \mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}}). \tag{E.3}$$

The solution of this problem thus provides a matrix $\mathbf{A}$ that maximizes the goodness (*i.e.* minimizes the l_q-risk) of the similarity function $K_{\mathbf{A}}$. Instead of solving the constrained problem (E.3), we rewrite it by adding a regularization term given by the norm of the matrix $\mathbf{A}$ as follows:

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n}} \mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}}) + \beta\|\mathbf{A}\|^2, \tag{E.4}$$

[1]We recall that l_q and ℓ_q are related by $l_q(y, y') = \ell_q(y \cdot y')$ (Equation (1.18)).

where $\beta > 0$ is a regularization hyperparameter. This is the formulation used in Bellet et al. (2012) where the authors consider the hinge loss instead of the quadratic one. With our formulation, we can now present our first result giving the closed-form solution of Problem (E.4).

E.2.2 Deriving a Closed-Form Solution for the Similarity Matrix

Before presenting our first result, we define the following quantities that we will use throughout the paper:

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}\left[\mathbf{x}\mathbf{x}^T\right] = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(y\mathbf{x})(y\mathbf{x})^T\right], \\
\boldsymbol{\mu} &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y\mathbf{x}\right], \quad \boldsymbol{\mu}' = \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}}\left[y'\mathbf{x}'|R(\mathbf{x}')=1\right], \\
\beta' &= \frac{\beta}{\|\boldsymbol{\mu}'\|^2}, \quad \mathbf{M}_\beta = (\boldsymbol{\Sigma}+\beta'\mathbf{I})^{-1}.
\end{aligned}$$

Here, $\boldsymbol{\Sigma}$ is a positive semidefinite symmetric matrix, and it is positive definite (PD) if and only if distinct features are linearly independent. From now on, we suppose this is the case for our distribution $\mathcal{D}$. If $\boldsymbol{\Sigma}$ is PD, then $(\boldsymbol{\Sigma}+\beta'\mathbf{I})$ and $\mathbf{M}_\beta$ are also PD. The solution to Equation (E.4) can now be given in a closed-form by the following proposition.

Proposition E.2.1 (Closed-form solution)**.** *The minimization problem Equation* (E.4) *can be rewritten as:*

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\mathbf{x}^T\mathbf{A}\boldsymbol{\mu}' - ys)^2\right] + \beta\|\mathbf{A}\|^2 \tag{E.5}$$

and its solution is given by

$$\mathbf{A}^* = \frac{s}{\|\boldsymbol{\mu}'\|^2}\mathbf{M}_\beta\boldsymbol{\mu}\boldsymbol{\mu}'^T. \tag{E.6}$$

Proof. We denote by $J(\mathbf{A})$ the cost function of the optimization problem Equation (E.4) so that $J(\mathbf{A}) = \mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}}) + \beta\|\mathbf{A}\|^2$. We now rewrite $\mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}})$ as follows:

$$\begin{aligned}
\mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}}) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_q\left(\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}}\left[K(\mathbf{x},\mathbf{x}')yy'|R(\mathbf{x}')=1\right]\right)\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(\mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}}\left[K(\mathbf{x},\mathbf{x}')yy'\right] - s\right)^2 |R(\mathbf{x}')=1\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(\mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}}\left[y\mathbf{x}^T\mathbf{A}\mathbf{x}'y'\right] - s\right)^2 |R(\mathbf{x}')=1\right] && \text{(E.7)} \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(y\mathbf{x}^T\mathbf{A}\boldsymbol{\mu}' - s\right)^2\right] && \text{(E.8)} \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\boldsymbol{\mu}'^T\mathbf{A}^T\mathbf{x}\mathbf{x}^T\mathbf{A}\boldsymbol{\mu}' - 2sy\mathbf{x}^T\mathbf{A}\boldsymbol{\mu}'\right] + s^2 && \text{(E.9)} \\
&= \boldsymbol{\mu}'^T\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}' - 2s\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}' + s^2. && \text{(E.10)}
\end{aligned}$$

Here, Equations (E.7)–(E.8) are obtained from the definitions of $K_{\mathbf{A}}$ and $\boldsymbol{\mu}'$, while Equations (E.9)–(E.10) follow from the linearity of the expectation and the definition of $\boldsymbol{\Sigma}$. It is clear that $\mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}})$ is quadratic in $\mathbf{A}$, hence convex in $\mathbf{A}$, and so is $J(\mathbf{A})$. In this case, $\mathbf{A}^*$ is a minimizer of $J(\mathbf{A})$ iff $\nabla J(\mathbf{A}^*) = \nabla_A\mathfrak{E}^q_{\mathcal{D}}(K^*_{\mathbf{A}}) + \beta\|\mathbf{A}^*\|^2 = 0$, where $\mathbf{K}^*_{\mathbf{A}} = \mathbf{x}^T\mathbf{A}^*\mathbf{x}'$, $\forall\mathbf{x},\mathbf{x}'\sim\mathcal{P}$. For any $\mathbf{H}\in\mathbb{R}^{n\times n}$, we can calculate $\nabla J(\mathbf{A}^*)$ as follows:

$$\begin{aligned}
&\mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}+\mathbf{H}}) - \mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}}) \\
&= (\boldsymbol{\mu}'^T\mathbf{H}^T\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}' + \boldsymbol{\mu}'^T\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\mu}' - 2s\boldsymbol{\mu}^T H\boldsymbol{\mu}') + o(\mathbf{H}) \\
&= \mathrm{Tr}\big(\mathbf{H}^T\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}'\boldsymbol{\mu}'^T\big) + \mathrm{Tr}\big(\mathbf{H}\boldsymbol{\mu}'\boldsymbol{\mu}'^T\mathbf{A}^T\boldsymbol{\Sigma}\big)
\end{aligned}$$

$$- \operatorname{Tr}\big(2s\boldsymbol{\mu}'\boldsymbol{\mu}^T H\big) + o(\mathbf{H}) \tag{E.11}$$
$$= \operatorname{Tr}\big(2\mathbf{H}^T(\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}'\boldsymbol{\mu}'^T - 2s\mathbf{H}^T\boldsymbol{\mu}\boldsymbol{\mu}'^T)\big) + o(\mathbf{H}) \tag{E.12}$$
$$= 2\left\langle \boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}'\boldsymbol{\mu}'^T - s\boldsymbol{\mu}\boldsymbol{\mu}'^T, \mathbf{H}\right\rangle + o(\mathbf{H}), \tag{E.13}$$

where $o(\mathbf{H})$ denotes the remainder that can be omitted and Equations (E.11)–(E.13) follow from the trace's properties and the definition of the Frobenius inner product. Hence,

$$\nabla_A \mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}}) = 2(\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}'\boldsymbol{\mu}'^T - s\boldsymbol{\mu}\boldsymbol{\mu}'^T).$$

Setting $\nabla J(\mathbf{A}^*) = \nabla_A \mathfrak{E}^q_{\mathcal{D}}(K_{\mathbf{A}^*}) + \beta\mathbf{A}^* = 0$ implies

$$\boldsymbol{\Sigma}\mathbf{A}^*\boldsymbol{\mu}'\boldsymbol{\mu}'^T + \beta\mathbf{A}^* = s\boldsymbol{\mu}\boldsymbol{\mu}'^T. \tag{E.14}$$

We now right multiply the obtained expression by $\boldsymbol{\mu}'$ and get:

$$\begin{aligned}
\boldsymbol{\Sigma}\mathbf{A}^*\boldsymbol{\mu}'\|\boldsymbol{\mu}'\|^2 + \beta\mathbf{A}^*\boldsymbol{\mu}' &= s\boldsymbol{\mu}\|\boldsymbol{\mu}'\|^2 \\
\boldsymbol{\Sigma}\mathbf{A}^*\boldsymbol{\mu}' + \frac{\beta}{\|\boldsymbol{\mu}'\|^2}\mathbf{A}^*\boldsymbol{\mu}' &= s\boldsymbol{\mu} \\
\mathbf{A}^*\boldsymbol{\mu}' &= s\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)^{-1}\boldsymbol{\mu}.
\end{aligned}$$

By injecting this last result into Equation (E.14), we obtain:

$$s\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)^{-1}\boldsymbol{\mu}\boldsymbol{\mu}'^T + \beta'\mathbf{A}^* = s\boldsymbol{\mu}\boldsymbol{\mu}'^T.$$

The final expression for $\mathbf{A}^*$ can be now derived as follows:

$$\begin{aligned}
\mathbf{A}^* &= \frac{s}{\beta}(\mathbf{I} - \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)^{-1})\boldsymbol{\mu}\boldsymbol{\mu}'^T \\
&= \frac{s}{\beta}(\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)^{-1} - \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)^{-1})\boldsymbol{\mu}\boldsymbol{\mu}'^T \\
&= \frac{s}{\beta}\frac{\beta}{\|\boldsymbol{\mu}'\|^2}\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)^{-1}\boldsymbol{\mu}\boldsymbol{\mu}'^T \\
&= \frac{s}{\|\boldsymbol{\mu}'\|^2}\left(\boldsymbol{\Sigma} + \beta'\mathbf{I}\right)^{-1}\boldsymbol{\mu}\boldsymbol{\mu}'^T.
\end{aligned}$$

□

From the established result, one may observe that the closed-form solution for $\mathbf{A}^*$ is similar to the one solving the ridge regression problem with the only difference consisting in a right matrix multiplication by $\boldsymbol{\mu}'^T$. In fact, one can equivalently obtain a ridge regression problem by introducing $\operatorname{vec}(\mathbf{A})$ instead of $\mathbf{A}$ into the objective function, where $\mathbf{a} := \operatorname{vec}(\mathbf{A})$ is obtained by stacking all vector columns of $\mathbf{A}^*$. Problem (E.5) can then be rewritten as follows:

$$\min_{\mathbf{a}\in\mathbb{R}^{n^2}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[((\boldsymbol{\mu}'\otimes\mathbf{x})^T\mathbf{a} - ys)^2\right] + \beta\|\mathbf{a}\|^2,$$

where $\otimes$ denotes the Kronecker product[2]. In this case, the solution $\mathbf{a}^*$ can be obtained by solving the following system of linear equations:

$$(\boldsymbol{\mu}'\boldsymbol{\mu}'^T \otimes \boldsymbol{\Sigma} + \beta'\mathbf{I}_{n^2})\mathbf{a}^* = s(\boldsymbol{\mu}'\otimes\boldsymbol{\mu}). \tag{E.15}$$

Opting for this reformulation, however, has several important drawbacks that make our closed-form solution more attractive. First, the matrix $(\boldsymbol{\mu}'\boldsymbol{\mu}'^T\otimes\boldsymbol{\Sigma}+\beta'\mathbf{I}_{n^2})$ has its minimum

[2]The Kronecker product of two vectors $\mathbf{u}$ and $\mathbf{v}$ is the vector obtained by horizontally concatenating the rows of matrix $\mathbf{u}\mathbf{v}^T$

singular value equal to β' due to the fact that $\boldsymbol{\mu}'\boldsymbol{\mu}'^T$ is singular (as its rank is equal to 1) implying that $(\boldsymbol{\mu}'\boldsymbol{\mu}'^T \otimes \boldsymbol{\Sigma})$ is singular as well. This makes the condition number of $(\boldsymbol{\mu}'\boldsymbol{\mu}'^T \otimes \boldsymbol{\Sigma} + \beta'\mathbf{I}_{n^2})$ unbounded when considered as a function of β. On the contrary, for our solution the minimum singular value of $\boldsymbol{\Sigma}+\beta\mathbf{I}$ is $\sigma_{\min}(\boldsymbol{\Sigma})+\beta \geq \sigma_{\min}(\boldsymbol{\Sigma}) > 0$ as $\boldsymbol{\Sigma}$ is PD. We refer the interested reader to (Schacke, 2003) for various properties of the Kronecker product and the vectorization operation that we used to obtain this result. Second, even when β' is large enough to ensure the finiteness of the condition number, solving linear system Equation (E.15) in the best case (when matrix $\boldsymbol{\mu}'\boldsymbol{\mu}'^T \otimes \boldsymbol{\Sigma} + \beta'\mathbf{I}_{n^2}$ is triangular) requires $\mathcal{O}\left((n^2)^2\right) = \mathcal{O}\left(n^4\right)$ operations, while computing our solution always requires only $\mathcal{O}\left(n^3\right)$ operations, as it involves inversion and multiplications of matrices of size $n \times n$ at most. Finally, after inverting the $n^2 \times n^2$ matrix from Equation (E.15), recovering the optimal solution $\mathbf{A}^*$ from its vectorized form $\mathbf{a}^*$ makes it difficult to characterize important properties of this latter related to its rank and its spectrum.

E.2.3 Bounding the Norm of the Optimal Similarity Matrix

The closed-form solution derived above allows us to establish several other interesting theoretical guarantees related to the properties of the optimal matrix $\mathbf{A}^*$. As the original problem introduced in Equation (E.4) imposes a constraint on the norm of the solution, one may want to bound it in order to understand when the inequality $\|\mathbf{A}^*\| \leq 1$ is satisfied. We establish this result in the following proposition.

Proposition E.2.2 (Interval norm bound)**.** *Let* $\mathbf{A}^*$ *be the solution of the optimization problem* (E.4)*, then:*

$$\frac{s\|\boldsymbol{\mu}'\|\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|^2\sigma_{\max}(\boldsymbol{\Sigma}) + \beta} \leq \|\mathbf{A}^*\| \leq \frac{s\|\boldsymbol{\mu}'\|\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|^2\sigma_{\min}(\boldsymbol{\Sigma}) + \beta}.$$

Proof. Taking the square of the Frobenius norm, we have:

$$\begin{aligned}
\|\mathbf{A}\|^2 &= \mathrm{Tr}\big(\mathbf{A}\mathbf{A}^T\big) \\
&= \left(\frac{s}{\|\boldsymbol{\mu}'\|^2}\right)^2 \mathrm{Tr}\big(\mathbf{M}_\beta\boldsymbol{\mu}\boldsymbol{\mu}'^T\boldsymbol{\mu}'\boldsymbol{\mu}^T\mathbf{M}_\beta\big) \\
&= \left(\frac{s}{\|\boldsymbol{\mu}'\|}\right)^2 \boldsymbol{\mu}^T\mathbf{M}_\beta^2\boldsymbol{\mu}.
\end{aligned}$$

Using the fact that for any PSD matrix $\mathbf{Q}$

$$\sigma_{\min}(\mathbf{Q}) = \min_{\|\mathbf{v}\|=1} \mathbf{v}^T\mathbf{Q}\mathbf{v} \text{ and } \sigma_{\max}(\mathbf{Q}) = \max_{\|\mathbf{v}\|=1} \mathbf{v}^T\mathbf{Q}\mathbf{v},$$

we obtain:

$$\begin{aligned}
\|\boldsymbol{\mu}\|^2\sigma_{\min}(\mathbf{M}_\beta^2) \leq \boldsymbol{\mu}^T\mathbf{M}_\beta^2\boldsymbol{\mu} &\leq \|\boldsymbol{\mu}\|^2\sigma_{\max}(\mathbf{M}_\beta^2) \\
\left(s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}\right)^2 \sigma_{\min}(\mathbf{M}_\beta^2) \leq \|\mathbf{A}^*\|^2 &\leq \left(s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}\right)^2 \sigma_{\max}(\mathbf{M}_\beta^2) \\
\left(s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}\right)^2 \sigma_{\min}^2(\mathbf{M}_\beta) \leq \|\mathbf{A}\|^2 &\leq \left(s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}\right)^2 \sigma_{\max}^2(\mathbf{M}_\beta) \\
s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}\sigma_{\min}(\mathbf{M}_\beta) \leq \|\mathbf{A}^*\| &\leq s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}\sigma_{\max}(\mathbf{M}_\beta) \\
\frac{s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}}{\sigma_{\max}(\boldsymbol{\Sigma}) + \beta'} \leq \|\mathbf{A}^*\| &\leq \frac{s\frac{\|\boldsymbol{\mu}\|}{\|\boldsymbol{\mu}'\|}}{\sigma_{\min}(\boldsymbol{\Sigma}) + \beta'}.
\end{aligned}$$

The last step is obtained by the fact that $\sigma_{\min}(\mathbf{M}_\beta) = \frac{1}{\sigma_{\max}(\mathbf{M}_\beta^{-1})}$ and $\sigma_{\max}(\mathbf{M}_\beta) = \frac{1}{\sigma_{\min}(\mathbf{M}_\beta^{-1})}$. □

This result provides a sufficient condition for parameter s that ensures $\|\mathbf{A}^*\| \leq 1$. Consequently, we can derive the following bound for s satisfying $\|\mathbf{A}^*\| \leq 1$:

$$s \leq \frac{\|\boldsymbol{\mu}'\|^2 \sigma_{\min}(\boldsymbol{\Sigma}) + \beta}{\|\boldsymbol{\mu}'\|\|\boldsymbol{\mu}\|}. \tag{E.16}$$

One can note that, for a fixed β, all quantities in this expression can be calculated explicitly in practice so that the value of s can be set beforehand in order to ensure that the constraint $\|\mathbf{A}^*\| \leq 1$ is satisfied.

E.3 Theoretical Analysis

In this section, we provide generalization guarantees for our approach. We do this in two different ways: (1) we relate our approach to the original (ϵ, γ, τ)-good framework in order to ensure that Theorem 1.4.1 holds for it too; (2) we use the closed-form solution of problem Equation (E.4) to derive a data- and problem-dependent generalization bound for our method.

E.3.1 Relation to $(\epsilon, \gamma, \tau)-$goodness in Margin Violation

In order to relate our approach to the original (ϵ, γ, τ)-good framework, we now derive the optimal value of the objective function obtained using Proposition E.2.1 for the optimal bilinear similarity function $K_{\mathbf{A}^*}$. This result is given by the following lemma.

Lemma E.3.1 (Optimal goodness)**.** *Let* $\mathbf{A}^*$ *be the solution of the optimization problem Equation* (E.4)*. The risk at* $\mathbf{A}^*$*,* i.e. *the minimum risk, is*

$$\mathfrak{E}_{\mathcal{D}}^q\left(K_{\mathbf{A}^*}\right) = s^2\left(1 - \boldsymbol{\mu}^T(\boldsymbol{\Sigma} + \beta'\mathbf{I})^{-2}(\boldsymbol{\Sigma} + 2\beta'\mathbf{I})\boldsymbol{\mu}\right).$$

Proof. To obtain the desired result, we need to plug the solution of the optimization problem given by Proposition E.2.1 into the expression of $\mathfrak{E}_{\mathcal{D}}^q\left(K_{\mathbf{A}}\right)$. To this end, we begin by expressing the first two terms appearing in Equation (E.10) as follows:

$$\begin{aligned}\boldsymbol{\mu}'^T\mathbf{A}^{*T}\boldsymbol{\Sigma}\mathbf{A}^*\boldsymbol{\mu}' &= \left(\frac{s}{\|\boldsymbol{\mu}'\|^2}\right)^2 \boldsymbol{\mu}'^T\boldsymbol{\mu}'\boldsymbol{\mu}^T\mathbf{M}_\beta\boldsymbol{\Sigma}\mathbf{M}_\beta\boldsymbol{\mu}\boldsymbol{\mu}'^T\boldsymbol{\mu}' \\ &= s^2\boldsymbol{\mu}^T\mathbf{M}_\beta^2\boldsymbol{\Sigma}\boldsymbol{\mu}. \end{aligned}\tag{E.17}$$

On the other hand, we have that

$$\begin{aligned}\boldsymbol{\mu}^T\mathbf{A}^*\boldsymbol{\mu}' &= \frac{s}{\|\boldsymbol{\mu}'\|^2}\boldsymbol{\mu}^T\mathbf{M}_\beta\boldsymbol{\mu}\boldsymbol{\mu}'^T\boldsymbol{\mu}' \\ &= s\boldsymbol{\mu}^T\mathbf{M}_\beta\boldsymbol{\mu}. \end{aligned}\tag{E.18}$$

Hence, Equation (E.10) can be equivalently written as

$$\begin{aligned}&\boldsymbol{\mu}'^T\mathbf{A}^{*T}\boldsymbol{\Sigma}\mathbf{A}^*\boldsymbol{\mu}' - 2s\boldsymbol{\mu}^T\mathbf{A}^*\boldsymbol{\mu}' + s^2 \\ &\qquad = s^2(\boldsymbol{\mu}^T\mathbf{M}_\beta^2\boldsymbol{\Sigma}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\mathbf{M}_\beta\boldsymbol{\mu} + 1) \\ &\qquad = s^2(\boldsymbol{\mu}^T(\mathbf{M}_\beta^2\boldsymbol{\Sigma} - 2\mathbf{M}_\beta)\boldsymbol{\mu} + 1),\end{aligned}\tag{E.19}$$

where Equation (E.19) is obtained from Equation (E.17) and Equation (E.18). We can now express matrix $\mathbf{M}_\beta^2\boldsymbol{\Sigma} - 2\mathbf{M}_\beta$ as follows:

$$\mathbf{M}_\beta^2\boldsymbol{\Sigma} - 2\mathbf{M}_\beta = \mathbf{M}_\beta^2(\boldsymbol{\Sigma} + \beta'\mathbf{I}) - \beta'\mathbf{M}_\beta^2 - 2\mathbf{M}_\beta$$

$$
\begin{aligned}
&= \mathbf{M}_\beta - \beta' \mathbf{M}_\beta^2 - 2\mathbf{M}_\beta \\
&= -\mathbf{M}_\beta(\mathbf{I} + \beta' \mathbf{M}_\beta). \qquad \text{(E.20)}
\end{aligned}
$$

Finally, we obtain the desired result:

$$
\begin{aligned}
\mathfrak{E}_{\mathcal{D}}^{q}\left(K_{\mathbf{A}^*}\right) &= \boldsymbol{\mu}'^T \mathbf{A}^{*T} \boldsymbol{\Sigma} \mathbf{A}^* \boldsymbol{\mu}' - 2s\boldsymbol{\mu}^T \mathbf{A}^* \boldsymbol{\mu}' + s^2 \\
&= s^2(1 - \boldsymbol{\mu}^T \mathbf{M}_\beta(\mathbf{I} + \beta' \mathbf{M}_\beta)\boldsymbol{\mu}) \qquad \text{(E.21)} \\
&= s^2(1 - \boldsymbol{\mu}^T \mathbf{M}_\beta^2(\mathbf{M}_\beta^{-1} + \beta' \mathbf{I})\boldsymbol{\mu}) \\
&= s^2(1 - \boldsymbol{\mu}^T (\boldsymbol{\Sigma} + \beta' \mathbf{I})^{-2}(\boldsymbol{\Sigma} + 2\beta' \mathbf{I})\boldsymbol{\mu}),
\end{aligned}
$$

where Equation (E.21) is obtained using Equation (E.20). □

This lemma prepares a result showing the possibility to characterize the optimal solution we obtain using our approach in terms of Definition E.1.1. To this end, we investigate its goodness in margin violation loss in the proposition given below.

Proposition E.3.1 (Link to goodness in margin violation loss)**.** *Let* $\mathbf{A}^*$ *denote the solution of the optimization problem* (E.4) *given by Proposition E.2.1. Then, for any* $0 < \gamma < s$, $K_{\mathbf{A}^*}$ *is* (ϵ, γ, τ)*-good in margin violation, where*

$$
\begin{aligned}
\epsilon &= \frac{1}{(1 - \frac{\gamma}{s})^2} \mathfrak{E}_{\mathcal{D}}^{q}\left(K_{\mathbf{A}}^*\right) \\
&\leq \frac{1}{(1 - \frac{\gamma}{s})^2} \left(1 - \|\boldsymbol{\mu}\|^2 \frac{\sigma_{\max}(\boldsymbol{\Sigma}) + 2\beta'}{(\sigma_{\max}(\boldsymbol{\Sigma}) + \beta')^2}\right).
\end{aligned}
$$

Proof. When $\gamma < s$, we consider the margin violation event $y\mathbf{x}^T \mathbf{A}^* \boldsymbol{\mu}' < \gamma$ in order to further bound its probability. We start by writing that

$$
y\mathbf{x}^T \mathbf{A}^* \boldsymbol{\mu}' < \gamma \Rightarrow ((s - y\mathbf{x}^T \mathbf{A}^* \boldsymbol{\mu}'))^2 \geq (s - \gamma)^2. \qquad \text{(E.22)}
$$

The last equation is due to the fact that $t \mapsto (t - s)^2$ is decreasing on $(-\infty, s]$ and that $\gamma < s$. We can now apply Markov's inequality to bound the probability of this event as follows:

$$
\begin{aligned}
\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y\mathbf{x}^T \mathbf{A}^* \boldsymbol{\mu}' < \gamma\right] &\leq \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}\left[(s - y\mathbf{x}^T \mathbf{A}^* \boldsymbol{\mu}')^2 \geq (s - \gamma)^2\right] \\
&\leq \frac{\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(y\mathbf{x}^T \mathbf{A}^* \boldsymbol{\mu}' - s)^2\right]}{(\gamma - s)^2} \\
&= \frac{s^2}{(s - \gamma)^2}(1 - \boldsymbol{\mu}^T (\boldsymbol{\Sigma} + \beta' \mathbf{I})^{-2}(\boldsymbol{\Sigma} + 2\beta' \mathbf{I})\boldsymbol{\mu}).
\end{aligned}
$$

In order to obtain a lower bound for $\boldsymbol{\mu}^T (\boldsymbol{\Sigma} + \beta' \mathbf{I})^{-2}(\boldsymbol{\Sigma} + 2\beta' \mathbf{I})\boldsymbol{\mu}$, we note that $\mathbf{N} := (\boldsymbol{\Sigma} + \beta' \mathbf{I})^{-2}(\boldsymbol{\Sigma} + 2\beta' \mathbf{I})$ is PSD and its singular values belong to the set $\{F(\sigma) : \sigma \text{ is a singular value of } \boldsymbol{\Sigma}\}$, where $F : \sigma \mapsto \dfrac{\sigma + 2\beta'}{(\sigma + \beta')^2}$. A quick study shows that F is a decreasing function of σ, implying that $\sigma_{\min}(N) = F(\sigma_{\max}(\boldsymbol{\Sigma}))$.

Finally, we have

$$
\boldsymbol{\mu}^T \mathbf{N} \boldsymbol{\mu} \geq \|\boldsymbol{\mu}\|^2 \sigma_{\min}(\mathbf{N}) = \|\boldsymbol{\mu}\|^2 F(\sigma_{\max}(\boldsymbol{\Sigma}))
$$

yielding the desired bound on ϵ. □

The immediate consequence of this proposition is that it ensures the existence of a linear separator achieving a low error thanks to Theorem 1.4.1. The existence and the upper-bound on the error (in terms of margin violations) of this linear separator is associated to a projection space ϕ obtained using the optimal bilinear similarity $K_{\mathbf{A}^*}$ produced by our method. Thus, Proposition E.3.1 justifies the use of $K_{\mathbf{A}^*}$ with linear classifiers and allows us to expect that our method should also perform well empirically due to the established theoretical guarantee. Also, we note that this result may suggest picking an arbitrary large s to minimize the value of ϵ. However, this would be counter-productive for two main reasons. First, according to Equation (E.16), s must be bounded to ensure that $\|\mathbf{A}\| \leq 1$. Second, we show below that large s worsens the generalization capacity of the optimal similarity $K_{\mathbf{A}^*}$ established in the next section dedicated to generalization guarantees.

E.3.2 Generalization Guarantees

Our minimization algorithm is applied to an observed data sample drawn from the unknown distribution $\mathcal{D}$. In practice, however, we only have access to the empirical distribution $\hat{S}$ associated to the observed sample S (Equation (1.1)). Below, we investigate under what conditions the true risk of the empirical solution would be close to the minimum achievable true risk if one had full access to the unknown distribution $\mathcal{D}$. In what follows, for any quantity $Q = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x},y)]$, we denote its empirical counterpart by adding S as an index, *i.e.* $Q_S = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\hat{S}}[f(\mathbf{x},y)]$ (where f is some measurable function). In order to establish our generalization result, we further assume that all of the drawn instances are landmarks, *i.e.* $\boldsymbol{\mu}_S = \boldsymbol{\mu}'_S$. This assumption is made due to the fact that the landmark distribution is unknown in practice. To this end, we further note that it is rather well-founded in practice as the landmarks assumed to be drawn from $\boldsymbol{\mu}_{S'}$ are usually chosen from the set of available training points distributed according to $\boldsymbol{\mu}_S$. We are now ready to state our main generalization guarantee[3].

Proposition E.3.2 (Generalization bound)**.** *Assume $\boldsymbol{\mu}_S = \boldsymbol{\mu}'_S$ and let $\mathbf{A}^*$ and $\mathbf{A}^*_S$ denote the true and empirical solutions of the problem Equation* (E.4)*, respectively. Then, with probability at least $1-\delta$, we have:*

$$\mathfrak{E}^q_{\mathcal{D}}\left(K_{\mathbf{A}^*_S}\right) \leq \min_{\|\mathbf{A}\|\leq 1} \mathfrak{E}^q_{\mathcal{D}}\left(K_{\mathbf{A}}\right) + s^2\left(2 + \frac{\|\boldsymbol{\mu}'\|}{C} + \frac{\|\boldsymbol{\mu}'_S\|}{C_S}\right)\left(\frac{3+2C}{CC_S}\sqrt{\frac{2\log\frac{4}{\delta}}{m}} + \frac{4+2C}{CC_S}(1-\tau)\right)$$

where

$$C = \left(\left\|\boldsymbol{\mu}'\right\|^2\sigma_{\min}(\boldsymbol{\Sigma}) + \beta\right) \quad \textit{and} \quad C_S = \left(\left\|\boldsymbol{\mu}'_S\right\|^2\sigma_{\min}(\boldsymbol{\Sigma}_S) + \beta\right) \tag{E.23}$$

Proof. In this proof, we denote the maximum singular value as $\|.\|_\infty$, corresponding to the Schatten $p-$norm as $p \to \infty$. We have

$$\begin{aligned}
&\left|\mathfrak{E}^q_{\mathcal{D}}\left(K_{\mathbf{A}^*}\right) - \mathfrak{E}^q_{\mathcal{D}}\left(K_{\mathbf{A}^*_S}\right)\right| \\
=&\left|\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(sy - \mathbf{x}^T\mathbf{A}^*\boldsymbol{\mu}')^2\right] - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(sy - \mathbf{x}^T\mathbf{A}^*_S\boldsymbol{\mu}')^2\right]\right| \\
=&\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(2sy - \mathbf{x}^T(\mathbf{A}^* + \mathbf{A}^*_S)\boldsymbol{\mu}')\mathbf{x}^T(\mathbf{A}^* - \mathbf{A}^*_S)\boldsymbol{\mu}'\right] &\text{(E.24)}\\
\leq&\sup_{\|\mathbf{x}\|\leq 1}\left|2sy - \mathbf{x}^T(\mathbf{A}^* + \mathbf{A}^*_S)\boldsymbol{\mu}'\right|\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left|\mathbf{x}^T(\mathbf{A}^* - \mathbf{A}^*_S)\boldsymbol{\mu}'\right|\right] &\text{(E.25)}\\
\leq&(2s + \|\mathbf{A}^* + \mathbf{A}^*_S\|_\infty)\|\mathbf{A} - \mathbf{A}_S\|_\infty &\text{(E.26)}
\end{aligned}$$

[3]The proof is provided in Section E.6.

$$\leq (2s + \|\mathbf{A}^*\|_\infty + \|\mathbf{A}^*_S\|_\infty)\|\mathbf{A}^* - \mathbf{A}^*_S\|_\infty,$$

where (E.24) is due to the identity $\|\mathbf{u}\|^2 - \|\mathbf{v}\|^2 = \langle \mathbf{u} - \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, followed by (E.25) where the Hölder inequality is applied. We then use the fact that all of the instances are supposed to have a bounded $2-$norm to obtain (E.26).

In order to bound $\|\mathbf{A}\|^*$, notice that $\mathbf{A}^*$ has rank 1, which implies the following

$$\begin{aligned}
\|\mathbf{A}^*\|_\infty &= \left\| s(\left\|\boldsymbol{\mu}'\right\|^2 \boldsymbol{\Sigma} + \beta \mathbf{I})^{-1} \boldsymbol{\mu} \boldsymbol{\mu}'^T \right\|_\infty \\
&= s \left\| (\left\|\boldsymbol{\mu}'\right\|^2 \boldsymbol{\Sigma} + \beta \mathbf{I})^{-1} \boldsymbol{\mu} \right\| \left\|\boldsymbol{\mu}'\right\| && \text{(E.27)} \\
&\leq \left\| (\left\|\boldsymbol{\mu}'\right\|^2 \boldsymbol{\Sigma} + \beta \mathbf{I})^{-1} \right\|_\infty \left\|\boldsymbol{\mu}'\right\| && \text{(E.28)} \\
&= s \frac{\|\boldsymbol{\mu}'\|}{C}, && \text{(E.29)}
\end{aligned}$$

where we used the fact that for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\left\|\mathbf{u}\mathbf{v}^T\right\|_\infty = \|\mathbf{u}\|\|\mathbf{v}\|$. In the same manner, we have $\|\mathbf{A}_S\| \leq s \frac{\|\boldsymbol{\mu}'_S\|}{C_S}$. Hence,

$$\left| \mathfrak{E}^q_{\mathcal{D}} (\mathbf{A}^*) - \mathfrak{E}^q_{\mathcal{D}} (\mathbf{A}^*_S) \right| \leq s \left(2 + \frac{\|\boldsymbol{\mu}'\|}{C} + \frac{\|\boldsymbol{\mu}'_S\|}{C_S} \right) \|\mathbf{A}^* - \mathbf{A}^*_S\|_\infty.$$

Combining the last equation with the result of Lemma E.6.1 finishes the proof. □

We note that our generalization bound is not a classic PAC bound (Definition 1.2.1), as it depends on the considered probability distribution via expectation vectors $\boldsymbol{\mu}'_S$ and $\boldsymbol{\mu}'$, and second moment matrices $\boldsymbol{\Sigma}_S$ and $\boldsymbol{\Sigma}$. This dependence was possible due to the explicit expression of the solution of the considered optimization problem. Also, the bound is drastically different from that obtained in (Bellet et al., 2012) based on the algorithmic stability theory. This latter can be written in the following (asymptotic) form:

$$\mathfrak{E}^q_{\mathcal{D}} \left(K_{\mathbf{A}^*_S} \right) \leq \mathfrak{E}^q_{\hat{S}} \left(K_{\mathbf{A}^*_S} \right) + \mathcal{O} \left(\sqrt{\frac{\log \frac{1}{\delta}}{m}} \left(\frac{2(\tau + 2\beta\gamma)}{\tau \beta \gamma^2} + 1 \right) \right).$$

First, contrary to the inequality given above, Proposition E.3.2 provides a non-degenerate generalization guarantee for our approach in case when $\beta = 0$, *i.e.* when solving the optimization problem Equation (E.4) without the regularization term, as long as matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_S$ are invertible. This is a rather strong result as it is commonly expected that regularization is required for a method to generalize well. Note that this remark also holds when comparing the generalization bound in (Perrot and Habrard, 2015a, Theorem 2) as the authors use stability theory to derive it. Second, our bound involves problem-dependent terms, such as τ, but also data-dependent terms $\boldsymbol{\mu}_S$, $\boldsymbol{\Sigma}_S$ and the smallest singular value of the $\boldsymbol{\Sigma}_S$ matrix $\sigma_{\min}(\boldsymbol{\Sigma}_S)$ that can be calculated in practice. This latter point is quite important as it allows to gain insights into the characteristics of the available finite-size data sample that ensure good generalization. In addition to the decay in the number of samples in $\frac{1}{\sqrt{m}}$, the term $1 - \tau$ appears as the price that one has to pay for not knowing which points are landmarks. Finally, we note that the value of s can be fixed beforehand due to Proposition E.2.2, and that in practice, we consider that $\tau = 1$. While the disparity between this consideration and the real proportion of landmarks may decrease performance (as non landmark points are used to decide an instance's class), it guarantees that the empirical and real errors are close to each other.

E.4 Comparison to other Existing Methods

We now briefly compare the obtained results to the existing methods proposed in the literature on similarity learning. As mentioned in Section 1.4.4.2, most of the similarity learning methods rely on learning a similarity function based on pair (or triplet)-wise constraints derived from the class labels of the data points. Our proposed approach is different from these methods in three principal ways. First, we aim to learn a similarity function that can be used with linear classifiers, while the above-mentioned methods are specifically related to k-NN classification. Second, our approach is much more computationally attractive as its optimal solution can be calculated in a closed-form. Finally, the generalization guarantees provided for our approach are both data- and problem-dependent and allow to explicitly link the quality of the obtained similarity function to the performance of the linear classifier.

Another important line of research in similarity learning are online algorithms including such methods as POLA (Shalev-shwartz et al., 2004), LEGO (Jain et al., 2009), OASIS (Chechik et al., 2009) and a recent OPML algorithm (Li et al., 2018) among others. As computational efficiency is crucial in online learning, some of these algorithms (namely LEGO, OASIS, and OPML) benefit from a closed-form solution for updating the learned similarity for newly arrived samples, even though the initial similarity function has to be learned iteratively on the training data in the first place. On the contrary, our method provides a closed-form solution for the similarity function directly on the training data and thus can be complementary to these methods. Furthermore, the above-mentioned approaches are designed to be used for online classification, while our approach is a traditional supervised learning algorithm.

In terms of the obtained results, two of the most similar methods to the one proposed in this chapter were presented by Law et al. (2016) and Perrot and Habrard (2015a). The paper of Law et al. (2016) is a follow-up work of several other contributions on supervised clustering (Bach and Jordan, 2004; Lajugie et al., 2014), where the goal is to learn a similarity function in a supervised fashion so as to it in a given clustering algorithm. The authors of this paper derived a discriminative similarity function based on learning the Mahalanobis distance similar to the family of methods presented above. As in our work, this paper also proposes a closed-form solution for the introduced optimization problem and shows its connection to multivariate linear regression. In contrast, their method is restricted to be used in conjunction with the k-means algorithm Lloyd (1982) (and, in general, for the purpose of clustering) only, while our approach produces a similarity function that can be used by any linear classification algorithm at hand.

Finally, the most similar approach to ours is that of Perrot and Habrard (2015a). In their paper, the authors proposed a regression-based similarity learning approach that aims at moving examples of different classes to distinct virtual points constructed beforehand. More precisely, the objective function of RVML is given by:

$$\min_{\mathbf{A}} ||\mathbf{XA} - \mathbf{V}||^2 + \beta||\mathbf{A}||^2,$$

where $\mathbf{V}$ is a set of virtual points. One may notice that this is also a ridge regression problem where, contrary to our method, the response variable is given by the set of virtual points. This latter point makes our method more computationally efficient despite RVML also admitting a closed-form solution given by $\mathbf{A}^* = (\mathbf{X}^T\mathbf{X} + \beta\mathbf{I})^{-1}\mathbf{X}^T\mathbf{V}$. To see this, note that the inversion in the latter solution has a cost of $\mathcal{O}(d^3)$, while computing our optimal solution given in Equation (E.6) requires solving for $\mathbf{u} \in \mathbb{R}^{\mathbf{n}}$ the following system of linear equations

$$(\mathbf{\Sigma} + \beta\mathbf{I})\mathbf{u} = \boldsymbol{\mu}$$

with a complexity of $\mathcal{O}(n^2)$. This is further followed by the calculation of an outer product between $\mathbf{u}$ and $\frac{s}{\|\boldsymbol{\mu}'\|^2}\boldsymbol{\mu}'$, resulting in n^2 real number multiplications, thus reducing the

global complexity of the solution's computation to $\mathcal{O}(n^2)$. Also, contrary to our contribution, their method produces a similarity that is tailored to k-NN classifier and depends on the selection of the virtual points in order to be efficient. From the theoretical point of view, our approach also benefits from the learning guarantees related to the linear classifier's performance that are unavailable for the approach of Perrot and Habrard (2015a).

E.5 Conclusion

In this chapter, we presented several novel contributions to the similarity learning field based on the (ϵ, γ, τ)-good similarity learning framework of Balcan et al. (2008a) with application to linear classifiers. We proposed a ridge regression alike formulation of similarity learning derived from a reformulation of the above-mentioned theoretical framework and highlighted various insightful theoretical properties related to it. We further showed that despite the proposed reformulation, we are still able to benefit from the theoretical guarantees of the original (ϵ, γ, τ)-good similarity learning framework by deriving an upper bound on the probability of margin violations achieved by the optimal similarity function. Furthermore, the obtained closed-form solution allowed us to derive a generalization bound with terms depending on both the problem and the data at hand.

We underline that the closed-form solution presented in this chapter is a matrix having rank 1, and one can see this as the first step towards a more general negative result that we present in Section 3.3. To recall, in this part we proved that learning a bilinear similarity function with a quadratic regularization is redundant with learning a classifier in the similarity space as suggested by Theorem 1.4.1.

E.6 Proof of Proposition E.3.2

Lemma E.6.1. *Let $\mathbf{A}^*$ and $\mathbf{A}_S^*$ be the optimal similarity matrices respectively associated with true distribution $\mathcal{D}$ and empirical distribution $\hat{S}$. Then, with a probability at least $1-\delta$ over the draw of $S \sim \mathcal{D}^m$, we have:*

$$\frac{1}{s}\|\mathbf{A}^* - \mathbf{A}_S^*\|_\infty \leq \frac{3+2C}{CC_S}\sqrt{\frac{2\log\frac{4}{\delta}}{m}} + \frac{2+C}{CC_S}\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|, \tag{E.30}$$

where

$$C = (\|\boldsymbol{\mu}'\|^2\sigma_{\min}(\boldsymbol{\Sigma}) + \beta) \quad \textit{and} \quad C_S = (\|\boldsymbol{\mu}'_S\|^2\sigma_{\min}(\boldsymbol{\Sigma}_S) + \beta). \tag{E.31}$$

Proof. The first part of the proof concerns deterministic bounds. First, we notice that $\mathbf{A}^*$ and $\mathbf{A}_S^*$ can be re-written as follows:

$$\begin{aligned}\mathbf{A}^* &= s(\|\boldsymbol{\mu}'\|^2\boldsymbol{\Sigma} + \beta\mathbf{I})^{-1}\boldsymbol{\mu}\boldsymbol{\mu}'^T\\ \mathbf{A}_S^* &= s(\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S + \beta\mathbf{I})^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S'^T.\end{aligned}$$

This allows us to write:

$$\begin{aligned}&\frac{1}{s}\|\mathbf{A}^* - \mathbf{A}_S^*\|_\infty \qquad\qquad (\text{E.32})\\ =&\left\|(\|\boldsymbol{\mu}'\|^2\boldsymbol{\Sigma} + \beta\mathbf{I})^{-1}\boldsymbol{\mu}\boldsymbol{\mu}'^T - (\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S + \beta\mathbf{I})^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S'^T\right\|_\infty\\ \leq&\left\|((\|\boldsymbol{\mu}'\|^2\boldsymbol{\Sigma} + \beta\mathbf{I})^{-1} - (\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S + \beta\mathbf{I})^{-1})\boldsymbol{\mu}\boldsymbol{\mu}'^T\right\|_\infty\\ +&\left\|(\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S + \beta\mathbf{I})^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}'^T - \boldsymbol{\mu}_S\boldsymbol{\mu}_S'^T)\right\|_\infty\\ \leq&\left\|(\|\boldsymbol{\mu}'\|^2\boldsymbol{\Sigma} + \beta\mathbf{I})^{-1} - (\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S + \beta\mathbf{I})^{-1}\right\|_\infty\end{aligned}$$

$$
\begin{aligned}
&+\left\|(\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S+\beta\mathbf{I})^{-1}\right\|_\infty\left\|\boldsymbol{\mu}\boldsymbol{\mu}'^T-\boldsymbol{\mu}_S\boldsymbol{\mu}'^T+\boldsymbol{\mu}_S\boldsymbol{\mu}'^T-\boldsymbol{\mu}_S\boldsymbol{\mu}'^T_S\right\|_\infty\\
&\leq\left\|(\left\|\boldsymbol{\mu}'\right\|^2\boldsymbol{\Sigma}+\beta\mathbf{I})^{-1}-(\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S+\beta\mathbf{I})^{-1}\right\|\\
&+\left\|(\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S+\beta\mathbf{I})^{-1}\right\|_\infty\|\boldsymbol{\mu}-\boldsymbol{\mu}_S\|\|\boldsymbol{\mu}'\|+\left\|\boldsymbol{\mu}'^T-\boldsymbol{\mu}'^T_S\right\|_\infty\|\boldsymbol{\mu}_S\|_\infty \qquad \text{(E.33)}
\end{aligned}
$$

In the previous developments, we used the triangle inequality and the submultiplicativity of the spectral norm. In the last line, we use the fact that $\left\|\mathbf{a}\mathbf{b}^T\right\|=\|\mathbf{a}\|\|\mathbf{b}\|$, for any rank 1 matrix $\mathbf{a}\mathbf{b}^T$.

Let us first concentrate on the first term $\left\|(\|\boldsymbol{\mu}'\|^2\boldsymbol{\Sigma}+\beta\mathbf{I})^{-1}-(\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S+\beta\mathbf{I})^{-1}\right\|$. This this the difference between the resolvents[4] of matrices $\|\boldsymbol{\mu}'\|^2\boldsymbol{\Sigma}$ and $\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S$. By the second resolvent identity (Hille and Phillips, 1996, Theorem 4.8.2), we have:

$$
\begin{aligned}
&\left\|(\left\|\boldsymbol{\mu}'\right\|^2\boldsymbol{\Sigma}+\beta\mathbf{I})^{-1}-(\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S+\beta\mathbf{I})^{-1}\right\|_\infty && \text{(E.34)}\\
=&\left\|(\left\|\boldsymbol{\mu}'\right\|^2\boldsymbol{\Sigma}+\beta\mathbf{I})^{-1}(\left\|\boldsymbol{\mu}'\right\|^2\boldsymbol{\Sigma}-\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S)(\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S+\beta\mathbf{I})^{-1}\right\|_\infty && \text{(E.35)}\\
\leq&\left\|(\left\|\boldsymbol{\mu}'\right\|^2\boldsymbol{\Sigma}+\beta\mathbf{I})^{-1}\right\|_\infty\left\|(\left\|\boldsymbol{\mu}'\right\|^2\boldsymbol{\Sigma}-\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S)\right\|_\infty\left\|(\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S+\beta\mathbf{I})^{-1}\right\|_\infty && \text{(E.36)}\\
=&\frac{\left\|\|\boldsymbol{\mu}'\|^2\boldsymbol{\Sigma}-\|\boldsymbol{\mu}'_S\|^2\boldsymbol{\Sigma}_S\right\|_\infty}{(\|\boldsymbol{\mu}'\|^2\sigma_{\min}(\boldsymbol{\Sigma})+\beta)(\left\|\boldsymbol{\mu}'_S\right\|^2\sigma_{\min}(\boldsymbol{\Sigma}_S)+\beta)}. && \text{(E.37)}
\end{aligned}
$$

The numerator of the previous equation can be bounded as follows:

$$
\begin{aligned}
&\left\|(\left\|\boldsymbol{\mu}'\right\|^2\boldsymbol{\Sigma}-\left\|\boldsymbol{\mu}'_S\right\|^2\boldsymbol{\Sigma}_S)\right\|\\
\leq&\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_S\|+\left|\left\|\boldsymbol{\mu}'\right\|^2-\left\|\boldsymbol{\mu}'_S\right\|^2\right|\\
\leq&\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_S\|+2\left\|\boldsymbol{\mu}'-\boldsymbol{\mu}'_S\right\|. \qquad \text{(E.38)}
\end{aligned}
$$

Taking into account that the data points verify $\|\mathbf{x}\|\leq 1$, from Equations (E.33) and (E.38), we have

$$
\frac{1}{s}\|\mathbf{A}^*-\mathbf{A}^*_s\|\leq\frac{\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_S\|+2\|\boldsymbol{\mu}'-\boldsymbol{\mu}'_S\|}{(\|\boldsymbol{\mu}'\|^2\sigma_{\min}(\boldsymbol{\Sigma})+\beta)(\left\|\boldsymbol{\mu}'_S\right\|^2\sigma_{\min}(\boldsymbol{\Sigma}_S)+\beta)}+\frac{\|\boldsymbol{\mu}-\boldsymbol{\mu}_S\|+\|\boldsymbol{\mu}'-\boldsymbol{\mu}'_S\|}{\left\|\boldsymbol{\mu}'_S\right\|^2\sigma_{\min}(\boldsymbol{\Sigma}_S)+\beta}. \qquad \text{(E.39)}
$$

In addition, considering all of the available data points as landmarks, implying that $\boldsymbol{\mu}'_S=\boldsymbol{\mu}'$, we get:

$$
\frac{1}{s}\|\mathbf{A}^*-\mathbf{A}^*_s\|\leq\frac{\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_S\|+2(\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|+\|\boldsymbol{\mu}-\boldsymbol{\mu}_S\|)}{CC_S}+\frac{2\|\boldsymbol{\mu}-\boldsymbol{\mu}_S\|+\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|}{C_S}. \qquad \text{(E.40)}
$$

The numerator of the previous equation only contains norms between quantities and their empirical counterparts.

The second step of the proof concerns probabilistic bounds. In the previous expression, the terms $\|\boldsymbol{\mu}-\boldsymbol{\mu}_S\|$ and $\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_S\|$ can be bounded with high probability as they are related to the divergence between the true and empirical quantities defined for point drawn from the same probability distribution. In this work, we bound them using the concentration inequality for bounded independent random variables with values in a separable complex Hilbert space summarized in Rosasco et al. (2010, Section 2.4, (3)). This latter states that if $\xi_1,\ldots,\xi_n$ are zero mean independent random variables with values in a separable complex Hilbert space and such that $\forall i,\|\xi_i\|\leq B, i=1,\ldots,m$, then the following bound

$$
\left\|\frac{1}{m}\sum_{i=1}^{n}\xi_i\right\|\leq B\sqrt{\frac{2\log\frac{2}{\delta}}{m}}, \qquad \text{(E.41)}
$$

[4]The resolvent of a matrix $\mathbf{B}$ is the function $s\mapsto(\mathbf{B}-s\mathbf{I})^{-1}$ defined for all s that is not an eigenvalue of $\mathbf{B}$.

holds with probability at least $1-\delta$. We can apply this result to variables $y_i\mathbf{x}_i - \boldsymbol{\mu}_i$ and $\mathbf{x}_i\mathbf{x}_i^T - \boldsymbol{\Sigma}$, which verify: $\|y_i\mathbf{x}_i - \boldsymbol{\mu}\| \leq 1$ and $\left\|\mathbf{x}_i\mathbf{x}_i^T - \boldsymbol{\Sigma}\right\| \leq 1$ since we suppose that the data points are bounded, *i.e.* $\|\mathbf{x}\| \leq 1\ \forall \mathbf{x} \sim \mathcal{D}_{\mathbb{X}}$. Hence, we have with a probability at least $1-\delta$,

$$\frac{1}{s}\|\mathbf{A}^* - \mathbf{A}_S^*\| \leq \frac{\sqrt{\frac{2\log\frac{4}{\delta}}{m}}}{CC_S} + 2\frac{1+C}{CC_S}\sqrt{\frac{2\log\frac{4}{\delta}}{m}} + \frac{2+C}{CC_S}\left\|\boldsymbol{\mu} - \boldsymbol{\mu}'\right\| \tag{E.42}$$

$$= \frac{3+2C}{CC_S}\sqrt{\frac{2\log\frac{4}{\delta}}{m}} + \frac{2+C}{CC_S}\left\|\boldsymbol{\mu} - \boldsymbol{\mu}'\right\| \tag{E.43}$$

Finally, by noticing that $\boldsymbol{\mu}' = \underset{(\mathbf{x},y)\sim\mathcal{D}}{\mathbb{E}}[y\mathbf{x}|R(\mathbf{x})=1] = \frac{\underset{(\mathbf{x},y)\sim\mathcal{D}}{\mathbb{E}}[R(\mathbf{x})y\mathbf{x}]}{\underset{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}{\mathbb{P}}[R(\mathbf{x})=1]}$ and denoting $\underset{\mathbf{x}\sim\mathcal{D}_{\mathbb{X}}}{\mathbb{P}}[R(\mathbf{x})=1] = p$, we get:

$$\begin{aligned}
\left\|\boldsymbol{\mu} - \boldsymbol{\mu}'\right\| &= \left\| \underset{(\mathbf{x},y)\sim\mathcal{D}}{\mathbb{E}}[y\mathbf{x}] - \frac{\underset{(\mathbf{x},y)\sim\mathcal{D}}{\mathbb{E}}[R(\mathbf{x})y\mathbf{x}]}{p} \right\| \\
&= \frac{1}{p}\left\| \underset{(\mathbf{x},y)\sim\mathcal{D}}{\mathbb{E}}[(p - R(\mathbf{x}))y\mathbf{x}] \right\| \\
&= \frac{2p(1-p)}{p} = 2(1-p) \leq 2(1-\tau).
\end{aligned}$$

This completes the proof. □

Appendix F

Extended Summary in French/Résumé Étendu en Français

Ce chapitre est dédié au résumé en français du manuscrit, comme demandé par l'école doctorale. En plus de traduction de l'abstract, introduction et conclusion rédigées en anglais, nous résumons les chapitres de l'état de l'art (Part I) et nos contributions (Part II).

Abstract

L'explosion de la quantité de données produites chaque jour a fait de l' l'*Apprentissage Automatique*, branche de l'*Intelligence Artificielle*, un outil vital pour extraire des motifs de haute valeur à partir de celles-là. Concrètement, un algorithme d'apprentissage automatique apprend de tels motifs après avoir été entraîné sur un jeu de données appelé *(jeu de)données d'entraînement*, et sa performance est évaluée sur échantillon différent, appelé *(jeu de)données de test*.

L'*Adaptation de Domaine* est un champ de recherche active de l'apprentissage automatique, dans lequel les données d'entraînement et de test ne sont plus supposées provenir de la même distribution de probabilité, à l'opposé de l'*Apprentissage Supervisé*. Dans ce cas, les deux distributions des données d'entraînement et de test correspondent respectivement aux domaines *source* et *cible*. La théorie de l'apprentissage supervisé repose sur la convergence de la distribution empirique des données observées vers son vrai homologue afin d'établir des garanties de généralisation, tandis que celles-ci sont entravées par le changement de distribution et par le manque d'étiquettes dans le cas de l'adaptation de domaine. Par conséquent, des supposition de liaison entre entre les deux domaines sont inévitables afin de garantir le succès du processus d'apprentissage.

Nos contributions se focalisent sur trois aspects théoriques en relation avec l'adaptation de domaine pour les tâches de classification. Le premier est l'apprentissage avec des fonctions de similarité, qui traite les algorithmes de classification basés sur la comparaison d'une instance à d'autres exemples pour décider sa classe. Le deuxième est la classification à vaste marge qui concerne l'apprentissage d'un classifieur maximisant la séparation entre classes. Le troisième aspect est le *Transport Optimal* qui formalise un principe d'effort minimal pour le transport de masses de probabilité entre distributions.

Au début de cette thèse, nous nous intéressions à l'apprentissage avec ce que l'on appelle fonctions de similarités $(\epsilon, \gamma, \tau)-$bonnes dans le cadre de l'adaptation de domaine, puisque ces fonctions ont été introduites dans la littérature dans le cadre classique de l'apprentissage supervisé. C'est le sujet de notre première contribution dans laquelle nous étudions théoriquement la performance d'une fonction de similarité sur une distribution cible, étant donné qu'elle est adéquate pour la source. Puis, nous abordons plus

généralement le thème de la classification à vaste marge pour l'adaptation de domaine, avec des hypothèses de départ plus faibles que celles adoptées dans la première contribution. Dans ce contexte, nous proposons une nouvelle étude théorique et un algorithme d'adaptation de domaine, ce qui constitue notre deuxième contribution. Nous dérivons de nouvelles bornes prenant en compte la marge de classification dans le domaine cible, que nous convexifions en tirant profit de la théorie du Transport Optimal, en vue de dériver un algorithme d'adaptation de domaine présentant une variation adversariale du problème classique de Kantorovitch. Finalement, remarquant que notre formulation adversariale est généralisable pour inclure nombre d'autres cas d'intérêt, nous dédions notre dernière contribution aux variations adversariales ou minimax du problème du transport optimal, où nous démontrons l'adaptabilité de notre approche.

Introduction

L'évolution des technologies de l'information de nos jours engendre une production de données à une cadence continuellement croissante. À titre d'exemple, l'on peut penser aux millions de courriels, inscriptions de sites-web, transactions de compagnies et aux informations médicales de patients dans les hôpitaux, sauvegardés quotidiennement dans le monde entier. Nul ne peut nier l'attrait d'extraire des motifs de valeur à partir de ces données: Un site de e-commerce peut proposer de meilleures recommandations à ses clients sur la base de leurs activités antérieures, et un hôptial mettre à profit les expériences de patients actuels afin d'améliorer les expériences des suivants. Le facteur commun à ces exemples est le souhait de transformer l'expérience acquise à partir des données en expertise (Shalev-Shwartz and Ben-David, 2014), ce qui représente la finalité de l'*Apprentissage Automatique*.Ce sous-champ assez vaste de l'*Intelligence Artificielle* renferme plusieurs branches, et ces dernières dépendent de caractéristiques éventuelles correspondant, par exemple, à la forme des données, la manière dont elles sont accédées, le processus sous-jacent qui les génère, et quel motif l'on cherche.

Le travail présenté dans ce manuscrit concerne l'apprentissage hors-ligne, voulant dire que les données sont accédées en une seule fois sous la forme d'un ensemble fixé (à l'opposition de l'apprentissage en ligne) de paires d'entrées-sorties, et le motif à chercher est une règle qui produit une sortie en fonction d'une entrée donnée. En plus, nous nous focalisons sur le cas où les valeurs possibles des sorties sont finies et représentent des *étiquettes* ou *classes* auxquelles les entrées (appelées aussi exemples) appartiennent. Pour approximer la relation entre entrées et sorties, une *hypothèse* ou un *classifieur* est appris sur les données disponibles, *i.e.* il est raffiné jusqu'à ce qu'il corresponde aux observations. Le but derrière une telle procédure est d'être capable de faire des prédictions correctes sur de nouvelles données non utilisées au cours de l'apprentissage. L'étude du cadre d'apprentissage décrit jusqu'ici est fondé sur l'hypothèse que les données d'entraînement et de test proviennent de distributions de probabilité inconnues. La question si les deux ensembles de données sont générés par la même distribution définit deux disciplines de l'apprentissage statistique: supposer la même distribution correspond à l'*Apprentissage Supervisé*, alors qu'autoriser une différence entre les distributions d'entraînement et de test définit l'*Adaptation de Domaine*. L'*Apprentissage* supervisé a toujours été théoriquement étudié dans le cadre de la théorie d'apprentissage statistique, dans laquelle la principale préoccupation est la généralisation d'une hypothèse apprise à partir d'un échantillon à la vraie distribution qui l'a généré. L'étude de l'adaptation de domaine est plus récente, et est motivée par des situations réelles où les processus produisant les données sont soumis à des changements, mettant en cause l'hypothèse d'une même distribution de probabilité pour l'entraînement et le test, et où l'étiquetage est coûteux en temps et en ressources. Dans ce cas, deux distributions correspondent respectivement aux domaines *source* et *cible*. Typiquement, l'adaptation de domaine est convenable pour des situations où l'on a accès à

un nouveau échantillon de test récemment produit et non étiqueté dont la taille est largement supérieure à celle de l'échantillon source disponible précédemment. Une partie des données cible non étiquetées peut alors être exploitée avec les données sources étiquetées au cours de l'apprentissage. L'adaptation de domaine est un champs d'étude active, et le cadre de ce travail est son cas particulier où l'on considère une seule distribution pour le domaine source (contrairement à l'*Adaptation de Domaine Multi-sources*). L'écart entre les deux distributions des domaines, avec le manque d'informations d'étiquettes pour les données cible, rendent le problème de l'adaptation de domaine bien plus difficile que celui de l'apprentissage supervisé. De plus, intuitivement on ne peut pas espérer apprendre à partir de deux domaines complètement indépendants: par exemple, une personne qui essaie d'apprendre une nouvelle langue à partir d'un groupe linguistique différent du sien ne peut pas réussir sans supervision, alors qu'elle peut correctement deviner le sens de plusieurs mots dans une langue liée à la sienne, en se basant sur la similarité entre les mots. Cette intuition est présente dans la théorie de l'adaptation de domaine qui vise à à déterminer des conditions traduisant la liaison entre les deux domaines et qui aident à apprendre en dépit de la différence entre les distributions. Entre ces conditions, une faible divergence entre les domaines source et cible est un facteur commun pour presque toute la littérature de l'adaptation de domaine, avec des variations qui dépendent du choix d'une telle divergence. D'ailleurs, l'objectif de l'écrasante majorité des algorithmes de l'adaptation de domaine est la réduction de cette divergence via une procédure d'alignement visant à rapprocher les domaines l'un de l'autre, et la proximité entre les domaines est dans ce cas conditionnée par le choix de la divergence. L'un des choix devenant de plus en plus populaire récemment est la *Distance de Wasserstein* entre distributions de probabilité, associée au problème du *Transport Optimal*. Ce dernier est une formalisation du principe des moindres efforts pour le transport de masses de probabilité entre distributions.

Précédemment dans cette introduction, nous avons mis l'accent sur le fait que nous nous focalisons sur les tâches de classification, où les valeurs possibles de la sortie ont une nombre fini. Plusieurs approches existent pour résoudre de telles tâches, parmi lesquelles on cite celles basées sur l'intuition du "Qui se ressemble s'assemble", *i.e.* elles s'appuient sur la similarité d'une instance au reste des données pour décider de sa classe. Deux algorithmes populaire à cet égard sont le *l plus Proches Voisins* et les *Machines à Vecteur Support*: le premier repose sur les distances et le deuxième sur des fonctions spéciales appelées *Noyaux*, où les deux reflètent une ressemblance entre les instances. Les deux algorithmes emploient des fonctions de similarité, où ces dernières sont fixées au préalable, ce qui peut potentiellement causer leur échec à capter des motifs cachés dans les données disponibles. La question si une fonction de similarité convient pour la tâche de classification à accomplir se pose naturellement, et l'un des travaux qui y répond est la théorie des fonctions de similarité $(\epsilon, \gamma, \tau)-$bonnes. En gros, ces fonctions requièrent l'existence d'une distribution de probabilité générant des points repères (*landmark points* en anglais), tels que la plupart des instances sont en moyenne plus similaires à des points repères ayant leur étiquettes qu'à ceux ayant des étiquettes opposées. Dans ce cas, les données peuvent être transformées dans un nouvel espace où les classes sont séparables avec une vaste marge.

Dans cette thèse, nous adressons quelques limitations de l'état de l'art actuel en adaptation de domaine pour les problèmes de classification. En premier lieu, en dépit de l'intuition forte et attractive derrière l'apprentissage avec des fonctions de similarité, on constate un manque de compréhension théorique de ces dernières dans le contexte d'adaptation de domaine. Nous abordons cette limitation dans notre première contribution où nous fournissons de nouveaux résultats qui étendent le cade (ϵ, γ, τ) à l'adaptation de domaine. En deuxième lieu, la plupart des résultats théoriques en adaptation de domaine ne considèrent pas la marge de classification dans le domaine cible. En effet, ils reposent essentiellement sur l'inégalité triangulaire pour la fonction de perte considérée, ce qui n'est pas vérifié pour des fonctions de perte visant à maximiser la marge de séparation.

Nous adressons ce problème dans notre deuxième contribution et nous utilisons notre contribution théorique pour proposer un algorithme d'adaptation de domaine où nous comparons les distributions source et cible via un terme de transport optimal adversarial et dépendant de la tâche en question. Ce dernier est étudié d'une manière plus générale dans notre dernière contribution, où nous résolvons diverses instances du problème de transport optimal adversarial et nous montrons sont intérêt pratique.

Plan

Le manuscrit est divisé en trois parties. La première, "Background", fournit au lecteur l'état de l'art courant sur les différents thèmes que nous adressons et est constitué de deux chapitre.

Chapitre 1 est dédié à l'apprentissage supervisé, avec insistance sur la classification binaire à vaste marge. Nous présentons brièvement quelques résultats importants de la théorie de l'apprentissage statistique et quelques algorithmes de référence. Puis nous dressons un portrait général de l'apprentissage avec les fonction de similarité $(\epsilon, \gamma, \tau)-$bonnes.

Chapitre 2 adresse l'adaptation de domaine, où nous commençons par définir le cadre spécifique auquel nous nous intéressons. Ensuite, nous présentons les hypothèses principales de la théorie de l'adaptation de domaine, avec les résultats de la littérature montrant leur suffisance et leur nécessité pour le succès de l'adaptation. En particulier, nous couvrons diverses mesures de divergence destinées à comparer les domaines source et cible.

La deuxième partie, "Contributions", présente notre travail basé sur des soumissions acceptées à des conférences internationales évaluées par des pairs. Chaque preuve des différents résultats théoriques que nous fournissons est disponible soit dans son chapitre correspondant si sa longueur est d'une demi-page au plus, soit elle est reportée à la partie suivante "Appendices". Dans le second case, nous décrivons brièvement l'idée de la preuve dans le chapitre principal.

Chapitre 3 correspond à nos publications Dhouib and Redko (2018a,b), où nous étudions les similarités $(\epsilon, \gamma, \tau)-$bonnes dans le cadre de l'adaptation de domaine. Nous établissons des résultats théoriques reliant les performance d'une fonction de similarité sur les domaines source et cible. Puis, nous présentons une étude rétrospective dans laquelle nous expliquons pourquoi nous abandonnons le cadre des similarités $(\epsilon, \gamma, \tau)-$bonnes pour le reste de cette thèse.

Chapitre 4 est basé sur nos publications Dhouib et al. (2019, 2020b), où nous affaiblissons les hypothèses du chapitre précédent et nous étudions la performance d'un classifieur sur le domaie cible tout en se concentrant sur la qualité de séparation entre les classes. Nous présentons des résultats théoriques qui généralisent qulques travaux précédents de la littérature, et nous introduisons une variation du problème de l'adaptation de domaine qui dépend de la tâche en question. Ensuite, nous spécialisons l'étude à la classification linéaire et nous montrons ses bienfaits empiriquement.

Chapitre 5 représente le travail donnant lieu à notre publication Dhouib et al. (2020a). Il traite une variation min-max du problème de transport optimal qui est une généralisation du terme introduit au chapitre précédent. Nous proposons une méthode d'optimisation pour le résoudre et nous détaillons ses variations en fonction de différentes instances du problème considéré.

La dernière partie, "Appendices", est pour les annexes contenant des pré-requis pour lire le manuscrit ou des détails supplémentaires sur ses différentes parties.

Annexe A rappelle quelques pré-requis nécessaires pour lire le manuscrit.

Annexes B, C et D fournissent plus de détails pour les chapitre 3, 4 et 5 respectivement, en incluant les preuves des différents résultats théoriques, en plus de détails supplémentaires pour les évaluations empiriques.

Annexe E est basée sur notre soumission pour le *Machine Learning journal*, où nous étudions théoriquement l'apprentissage d'une fonction de similarité $(\epsilon, \gamma, \tau)-$bonne via la régression, un travail qui est à l'origine de l'étude rétrospective présentée dans Chapitre 3.

Annexe F (F comme "français") présente un résumé du présent manuscrit en français, en accord avec les exigences de l'école doctorale EEA.

Notations

Dans un souci de faciliter la lecture du manuscrit et de permettre une reconnaissance rapide des différents types de quantités utilisées, nous adoptons les conventions de notation suivantes. Les ensembles mathématiquement adressés comme "espaces" (*e.g.* espaces vectoriels usuels, espaces d'hypothèses en apprentissage ...) sont notés avec la police `\mathbb`. La police gras est exclusivement réservée pour les vecteurs et les matrices, notés par des lettres latines respectivement minuscules et majuscules. Les scalaires sont notés par des lettres grecques ou latines minuscules. De plus, les probabilités sont désignées seulement par la police `\mathcal`. Ces conventions, en plus d'autres notations, sont résumées dans la page "Nomenclature" juste après le chapitre d'introduction au début du manuscrit. Finalement, nous présentons ci-dessous les traductions en français que nous adoptons pour quelques termes techniques.

Traductions adoptées pour quelques termes techniques en anglais

Anglais	Français
Accuracy	Justesse
scoring function	fonction de score
landmark	point repère
Integral Probability Metric	Métrique Intégrale de Probabilités
Maximum Mean Discrepancy	Ecart Moyen Maximal
Ground cost function	fonction de coût terrain

F.1 Apprentissage Supervisé

Dans cette section, nous présentons les points essentiels du Chapitre 1 portant sur l'apprentissage statistique. Ce dernier est l'un des cadres d'apprentissage statistique les plus connus. D'abord nous présentons la théorie formalisant l'apprentissage supervisé, puis nous nous concentrons sur la classification supervisée. En particulier, nous mettons en exergue l'intérêt des fonctions de score et leur marge de classification et nous couvrons quelques inégalités de généralisation pour la classification binaire. Ensuite, nous décrivons deux algorithmes de référence: les machines à vecteurs de support et les plus proches voisins.

Finalement, nous présentons le cadre des similarités $(\epsilon, \gamma, \tau)-$bonnes et nous le relions à l'apprentissage de métriques et de noyaux.

F.1.1 Cadre théorique

F.1.1.1 Données observées

Dans un problème d'apprentissage supervisé, on a accès à un échantillon $S = (\mathbf{x}_i, y_i)_{i=1}^m$ où les entrée $\mathbf{x}_i$ et sorties $\mathbf{y}_i$ sont respectivement des éléments de l'espace des caractéristiques et celui des sorties. On suppose dorénavant que $\mathbb{X} \subseteq \mathbb{R}^n$. Quant à l'espace de sortie, il peut prendre des valeurs dans un domaine continue, ce qui définit un problème de régression ou être fini, *i.e.* $\mathbb{Y} = \{c_1, \ldots, c_K\}$, pour un problème de classification. Les données sont supposées tirées indépendamment d'une distribution de probabilité $\mathcal{D}$ sur $\mathbb{X} \times \mathbb{Y}$. Dans un souci de homogénéisation, nous définissons également la distribution de probabilité empirique associée à un échantillon S comme une distribution uniforme sur ses éléments:

$$\hat{S} := \frac{1}{m} \sum_{i=1}^{m} \delta_{(\mathbf{x}_i, y_i)}, \tag{F.1}$$

où δ dénote la mesure de Dirac. De plus, nous notons la distribution des sorties conditionnées par une valeur $\mathbf{x} \in \mathbb{X}$ par $\mathcal{D}_{\mathbb{Y}|\mathbf{x}}$, les marginales de $\mathcal{D}$ par $\mathcal{D}_{\mathbb{X}}$ et $\mathcal{D}_{\mathbb{Y}}$.

F.1.1.2 Apprendre la tâche en question

L'apprentissage consiste à chercher une fonction $h : \mathbb{X} \to \mathbb{Y}$ visant à expliquer la relation entre les entrées et les sorties. h est appelée *hypothèse*, et elle est choisie à partir d'un ensemble de candidates défini au préalable, l'*espace des hypothèses* que nous notons $\mathbb{H}$ par la suite. Une fois h est choisie, sa performance sur les données est évaluée en faisant appel à une *fonction de perte* $l : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}_+$. Ainsi, $l(h(\mathbf{x}), y)$ mesure le désaccord entre la prédiction de h en $\mathbf{x}$ et la sortie observée y. Pour étendre cette mesure sur toutes les données, on considère son espérance $\mathfrak{E}^l_{\mathcal{D}}(h) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[l(h(\mathbf{x}), y)]$ pour la distribution $\mathcal{D}$ et sa moyenne empirique $\mathfrak{E}^l_{\hat{S}}(h) := \frac{1}{m}\sum_{i=1}^m l(h(\mathbf{x}_i), y_i)$. Les 2 dernières quantités sont désignées par le *vrai* $l-$*risque* et le $l-$*risque empirique*.

L'apprentissage d'une tâche consiste alors à trouver l'hypothèse qui minimise le vrai risque. Mais comme ce n'est pas possible vu l'ignorance de la distribution $\mathcal{D}$, l'une des premières stratégies consistait à minimiser le risque empirique $\mathfrak{E}^l_{\hat{S}}(h)$, motivée par la loi des grands nombres:

$$\min_{h\in\mathbb{H}} \mathfrak{E}^l_{\hat{S}}(h).$$

Cependant, une telle procédure est sujette au *sur-apprentissage*, un problème central en apprentissage que nous présentons dans ce qui suit.

Le compromis biais-variance Le sur-apprentissage arrive quand une hypothèse h_S est parfaitement en accord avec l'échantillon S observée, mais a une mauvaise performance sur la distribution $\mathcal{D}$. En pratique, ce comportement est constaté par une valeur élevée du $l-$risque sur un échantillon $S' \neq S$, issu de $\mathcal{D}$, et qui n'est pas utilisé pour apprendre h_S. Ceci est dû principalement à deux raisons principales. Premièrement, l'échantillon S n'est pas assez représentatif de la distribution $\mathcal{D}$ dont il provient, et dans cas collecter plus d'exemples peut remédier au problème. Deuxièmement, même lorsque S est assez large, le sur-apprentissage peut résulter d'une richesse excessive de $\mathbb{H}$. Par conséquent, la performance de h_S varie largement pour différents échantillons S tirés selon $\mathcal{D}$, et dans ce cas la performance sur S est potentiellement un mauvais indicateur de la performance sur toute la distribution $\mathcal{D}$. Pour cette raison, le sur-apprentissage est aussi connu sous le nom de problème de *grande variance* ou *grande complexité*.

Éviter le sur-apprentissage est possible par la collection de plus de données et par la restriction de l'espace d'hypothèses considéré $\mathbb{H}$. Cependant, quand une telle restriction est assez forte, on peut faire face au *sous-apprentissage*: la performance de h_S est mauvaise sur S et sur la distribution $\mathcal{D}$. Ce problème est aussi connu sous le nom de *grand biais*, puisque imposer des restrictions sur les hypothèses considérées reflète le biais inductif que l'on a envers $\mathbb{H}$ avant de commencer l'apprentissage.

Apprendre en évitant le sur-apprentissage Deux stratégies sont parmi les plus connues en littérature, la *minimisation structurelle du risque* (SRM), et la *minimisation du risque régularisé* (RRM). La première consiste à se donner une famille d'espaces d'hypothèses emboîtés, ce qui offre un équilibre entre minimiser le risque empirique et augmenter la complexité de l'espace d'hypothèses en question. La deuxième consiste à ajouter un terme $R(h)$ au risque empirique, appelé *régulariseur*, qui prend des valeurs élevées pour les hypothèses complexes et aide ainsi à choisir une hypothèse simple.

F.1.2 Classification supervisée

Définie par le cas $\mathbb{Y} = \{c_1, \ldots, c_K\}$, la classification supervisée est le problème que nous adressons principalement dans ce manuscrit. Une hypothèse est alors appelée un classifieur, et les valeurs de sorties dans $\mathbb{Y}$ sont dites des *étiquettes* ou *classes*. La fonction de perte la plus utilisée est la perte de *classification erronée*, aussi appelée la perte $0-1$, que l'on note l_{01} et elle est définie par

$$l_{01}(y, y') := [y \neq y'].$$

Dans le cas empirique, le risque de cette fonction de perte est la proportion des exemples pour lesquels on s'est trompé de prédiction, et dans le cas de la vraie distribution, il s'agit de la probabilité de classification erronée $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$.

Cependant, la minimisation du risque l_{01} est un problème NP-difficile Arora et al. (1997), en plus de la difficulté introduite par la recherche d'un classifieur prenant un nombre fini de valeurs, et donc qui est discontinu. Ces problèmes sont résolus par la considération d'une *fonction de perte remplaçante*. Nous décrivons ces deux notions dans le cadre de classification binaire par la suite.

F.1.2.1 Classification binaire

Dans ce cas, $\mathbb{Y}$ est formé exactement de deux éléments, et nous choisissons d'encoder les étiquettes par $\{-1, 1\}$, menant à la terminologie des classes positives est négatives pour les exemples.

Fonctions de score et marge de classification Au lieu de chercher un classifieur h ayant des valeurs dans $\{-1, 1\}$, il est commun de chercher une fonction $f : \mathbb{X} \to \mathbb{R}$, appelée *fonction de score*, et de prendre son signe comme classifieur h, *i.e.* $h(\mathbf{x}) := \operatorname{sgn}(() f(\mathbf{x}))$. Dans ce cas, pour un couple $(\mathbf{x}, y)$, on a

$$\operatorname{sgn}(f(\mathbf{x})) = y \quad \Leftrightarrow \quad y \cdot f(\mathbf{x}) > 0. \tag{F.2}$$

En plus de l'avantage des fonctions de score pour les procédures d'optimisation, elles renseignent sure la confiance accordée à une classes. En effet, la condition $y \cdot f(\mathbf{x}) > 0$ (Equation (F.2)) est plus faible que la condition $y \cdot f(\mathbf{x}) > \rho$ pour $\rho > 0$ un réel positif donné. En effet, la dernière condition traduit le fait que f est en accord avec y et qu'elle est confiante en sa prédiction. La quantité $y \cdot f(\mathbf{x})$ est appelée la *marge signée* de f, et traduit la confiance accordée en une prédiction correcte, tandis que $|f(\mathbf{x})|$ correspond à la *marge absolue*, et traduit la confiance accordée par f à sa prédiction indépendamment

de son accord avec l'étiquette y. Notons que la notion de marge de classification n'est pas exclusive à la classification binaire, et en effet Koltchinskii and Panchenko (2002) en propose une généralisation.

Fonctions de perte de remplacement Pour $\mathbb{Y} = \{-1, 1\}$, les fonctions de perte l utilisées en pratique ont la forme suivante:

$$l(h(\mathbf{x}), y) = \ell(y \cdot h(\mathbf{x})), \tag{F.3}$$

où ℓ est une fonction décroissante tendant vers 0 en ∞. Les exemples les plus populaires sont disponibles dans Table 1.2.

F.1.3 Garanties de généralisation en classification binaire

Nous formalisons les notions de richesse ou complexité d'une classe d'hypothèse, et leur impact sur l'écart entre le risque empirique et vrai risque d'une hypothèse donnée.

F.1.3.1 Apprentissage Probablement Approximativement Correct (PAC)

Introduit par Valiant (1984), l'apprentissage *probablement approximativement correct* (PAC) est un cadre théorique qui définit ce que l'on entend par "une hypothèse h est apprenable". Il formalise l'idée qu'étant donné un échantillon de $S \sim \mathcal{D}^m$ d'une taille suffisamment grande, le vrai risque associé à h_S est proche au minimum atteignable du vrai risque.,et ceci avec une grande probabilité sur le tirage d'un de l'échantillon S.

Pour prouver que un algorithme bénéficie de l'apprentissage PAC, une condition plus forte est souvent utilisée: on exige qu'avec une grande probabilité sur le tirage de $S \sim \mathcal{D}^m$, les risques d'une hypothèse h sur S et sur $\mathcal{D}$ sont proches, et ceci uniformément sur le choix de l'hypothèse h.

F.1.3.2 Bornes de généralisation uniformes

Définition F.1.1. *Étant donné un espace d'hypothèses $\mathbb{H}$ et une distribution de probabilité $\mathcal{D}$, une borne de généralisation uniforme a la forme suivante:*

Pour tout $\delta \in (0, 1)$, avec une probabilité d'au moins $1-\delta$ sur le tirage d'un échantillon $S \sim \mathcal{D}^m$, on a

$$\forall h \in \mathbb{H}, \quad \mathfrak{E}^l_{\mathcal{D}}(h) \leq \mathfrak{E}^l_{\hat{S}}(h) + \epsilon(\mathcal{D}, \mathbb{H}, \delta, m), \tag{F.4}$$

où $\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) \xrightarrow{m\to\infty} 0$.

Dans la définition précédente, l'écart entre les risque vrai et empirique de $h \in \mathbb{H}$ est contrôlé par $\epsilon(\mathcal{D}, \mathbb{H}, \delta, m)$. La dépendance de $\mathbb{H}$ est exprimé via une mesure de complexité: plus l'espace $\mathbb{H}$ est complexe, plus ce terme devient grand, causant un ralentissement de la convergence du risque empirique vers le vrai risque. Deux cas particuliers de bornes de généralisation uniformes sont présentées dans la suite, en fonction de la mesure de complexité de $\mathbb{H}$: la dimension de *Vapnik-Chervonenkis* (Vapnik, 1992) et la *complexité de Rademacher* (Koltchinskii and Panchenko, 2000).

Définition F.1.2 (dimension VC)**.** *La dimension de Vapnik-Chervonenkis (VC) d'un espace d'hypothèses binaires $\mathbb{H}$ est la taille du plus grand échantillon d'éléments de $\mathbb{X}$ pouvant être étiqueté de toutes les manières possibles par des hypothèses de $\mathbb{H}$:*

$$VC(\mathbb{H}) := \max\{|A|;\ |A| < \infty;\ A \subset \mathbb{X};\ |\mathbb{H}(A)| = 2^{|A|}\}, \tag{F.5}$$

où

$$\mathbb{H}(A) := \{h(\mathbf{x});\ \mathbf{x} \in A;\ h \in \mathbb{H}\} \tag{F.6}$$

est l'ensemble de tous les étiquetages possibles de A par des éléments de $\mathbb{H}$.

La dimension VC est une mesure de la richesse de l'espace d'hypothèses $\mathbb{H}$ et capte à partir de quelle taille d'échantillon un espace d'hypothèses $\mathbb{H}$ arrête de se comporter comme des fonctions de $\mathbb{Y}^{\mathbb{X}}$, puisque ces dernières peuvent étiqueter tout échantillon fini $A \subseteq \mathbb{X}$ de toutes les $2^{|A|}$ manières possibles. Elle est indépendante de la distribution $\mathcal{D}$, ce qui n'est pas le cas pour la mesure de complexité suivante.

Définition F.1.3 (Complexité de Rademacher)**.** *Soit $r_1, ..., r_m$ des variables aléatoires de Rademacher,* i.e.

$$\mathbb{P}[r_i = 1] = \mathbb{P}[r_i = -1] = \frac{1}{2}, \quad \forall 1 \leq i \leq m.$$

1. *La complexité de Rademacher empirique d'un espace d'hypothèses $\mathbb{H}$ associée à un échantillon $S \subset \mathbb{X}$ est*

$$\mathrm{Rad}_S(\mathbb{H}) := \mathop{\mathbb{E}}_{r_1,...,r_m \sim r} \left[\sup_{h \in \mathbb{H}} \frac{1}{m} \sum_{i=1}^{m} r_i h(\mathbf{x}_i) \right]. \tag{F.7}$$

2. *La complexité de Rademacher d'un espace d'hypothèses $\mathbb{H}$ associée à un échantillon de taille m est*

$$\mathrm{Rad}_m(\mathbb{H}) := \mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\mathbb{X}}^m} [\mathrm{Rad}_S(\mathbb{H})]. \tag{F.8}$$

Pour un échantillon S, la complexité empirique de Rademacher mesure la capacité des hypothèses de $\mathbb{H}$ à se corréler avec le bruit aléatoire défini par les variables aléatoires de Rademacher. Si la corrélation est élevée, alors les hypothèses sont trop souples et peuvent conduire à un sur-ajustement.

Les deux mesures de complexité permettent de quantifier la déviation entre le vrai $l-$risque et l'empirique, comme énoncé par le théorème suivant:

Théorème F.1.1. *Étant donné un espace d'hypothèses binaires $\mathbb{H}$ et la perte de classification erronée l_{01}, une borne de généralisation comme définie dans Equation* (F.4) *est vérifiée pour $\epsilon(\mathcal{D}, \mathbb{H}, \delta, m)$ défini:*

- *avec la dimension VC*

$$\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) = 2\sqrt{\frac{1}{m}\left(VC(\mathbb{H}) \log\left(\frac{2em}{VC(\mathbb{H})}\right) + \log\frac{4}{\delta}\right)}. \tag{F.9}$$

- *avec la complexité de Rademacher empirique*

$$\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) = \mathrm{Rad}_S(\mathbb{H}) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{F.10}$$

- *avec la complexité de Rademacher*

$$\epsilon(\mathcal{D}, \mathbb{H}, \delta, m) = \mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\mathbb{X}}^m} [\mathrm{Rad}_S(\mathbb{H})] + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{F.11}$$

F.1.4 Quelques algorithmes notables

F.1.4.1 Machines à vecteurs de support (SVM)

Introduit dans dans Boser et al. (1992); Cortes and Vapnik (1995), l'algorithme SVM sert à la classification binaire (bien qu'il est possible de l'étendre au cas multi-classes) dont

l'idée est de trouver un hyperplan séparant les exemples des deux classes, tout en restant le plus loin possible d'eux. Formellement, on cherche une hypothèse dans l'ensemble

$$\mathbb{H} = \{h : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b; \mathbf{w} \in \mathbb{R}^n, \|\mathbf{w}\|_2 = 1, b \in \mathbb{R}\}, \tag{F.12}$$

qui maximise la marge signée minimale sur l'échantillon S. On peut trouver une telle hypothèse en résolvant le problème suivant:

$$\begin{aligned} \min_{\substack{\mathbf{w}\in\mathbb{R}^n \\ b\in\mathbb{R}}} \quad & C\sum_{i=1}^{m} \xi_i + \|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & \xi_i \geq 0, \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \qquad \forall 1 \leq i \leq m, \end{aligned} \tag{F.13}$$

qui admet la forme duale suivante:

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in[0,C]^m} \quad & \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}^T\mathbf{y} = 0, \end{aligned} \tag{F.14}$$

Cette dernière forme est un problème d'optimisation quadratique (QP), et peut être résolue en utilisant des librairies d'optimisation convexe. Un algorithme bien spécialisé au problème dual du SVM est la *minimisation séquentielle minimale* (SMO) (Platt, 1998).

La formulation duale a l'atout d'être indépendante de la dimension de l'espace de caractéristiques $\mathbb{X}$: en effet, elle peut être exprimé en faisant appel uniquement aux produits scalaires entre les exemples, ce qui motive l'*astuce du noyau* (Boser et al., 1992). Cette dernière consiste à transformer les données dans un espace de dimension infinie, appelé *Espace de Hilbert à Noyau Reproduisant*(Aronszajn, 1950), et de considérer le produit scalaire, appelé *noyau*, dans cet espace. Ceci permet aussi de traiter des problèmes non linéairement séparables dans l'espace d'origine $\mathbb{X}$. De plus, grâce au théorème de Mercer (Mercer, 1909), on peut reconnaitre un produit scalaire dans un RKHS sans avoir à manipuler des instances de dimension infinie, ce qui est numériquement infaisable. Des exemples connus de noyaux sont donnés dans Table 1.3.

L'algorithme SVM bénéficie de fortes garanties de généralisation théoriques (Shalev-Shwartz and Ben-David, 2014, Section 26.3), et en pratique le compromis biais-variance est contrôlé par l'hyperparamètre $C > 0$, en plus des hyperparamètres spécifiques au noyau utilisé.

F.1.4.2 Algorithme des k-plus proches voisins (k-NN)

L'algorithme k-NN (Cover and Hart, 1967) est extrêmement intuitif: la classe d'une instance est décidée comme la classe de la majorité des k instances qui lui sont les plus proches, où k est un nombre fixé a priori et la distance entre les instances est mesurée par une métrique (souvent la métrique Euclidienne). Cet algorithme n'a pas besoin d'entraînement, mais seulement de mémoriser l'échantillon disponible.

Le paramètre k contrôle le compromis biais-variance, avec plus de variance et moins de biais pour quand on diminue k, et vice-versa.

F.1.5 Sélection de modèle

Dans la majorité des cas d'intérêt, trouver un classifieur en résolvant un problème d'optimisation nécessite de fixer au préalable certains paramètres. Par exemple, pour les algorithmes suivant la règle RRM (Section 1.1.4.3), il faut choisir le coefficient de régularisation, tandis que pour la formulation basée sur le noyau du SVM et les algorithmes k-NN, il faut sélectionner le paramètre du noyau (Table 1.3) et le nombre de voisins k, respectivement. Ces paramètres sont souvent appelés *hyperparamètres*, et ils reflètent une connaissance en

amont du problème en question. Ils peuvent également être considérés comme des degrés de liberté supplémentaires permettant d'affiner le modèle utilisé afin d'obtenir les meilleures performances possibles. Par exemple, en augmentant le multiplicateur de régularisation, on réduit l'espace de recherche vers un classifieur antérieur (égal à zéro dans la plupart des cas).

Pour tout algorithme d'apprentissage, une configuration donnée de ses hyperparamètres résulte en une instanciation différente du problème d'optimisation sous-jacent, et la sélection de la "bonne" configuration par rapport à une certaine mesure de performance est un exemple du problème *sélection de modèle*. L'approche la plus couramment utilisée est la *validation-croisée* (Stone, 1974) qui consiste à diviser les données disponibles en ensembles d'entraînement, de validation et de test. Le premier ensemble est ensuite utilisé pour apprendre une hypothèse, et la performance de cette dernière est évaluée sur le second ensemble qui guide la sélection des hyperparamètres. L'évaluation finale de cette approche se fait sur l'ensemble de tests qui n'a été utilisé ni pour l'entraînement, ni pour la sélection des hyperparamètres. Lorsque les données sont rares et qu'aucun ordre particulier n'est supposé pour l'ensemble d'entraînement, d'autres méthodes de sélection de modèles tirent le meilleur parti des données disponibles en considérant différents fractionnements de validation de la formation après avoir isolé un ensemble de tests. Par exemple, la procédure *à k-blocs* (pour $k \in \mathbb{N}^*$) est largement utilisée et consiste à diviser le reste des données en k blocs, puis, pour chaque bloc, sélectionner l $k^{\text{ème}}$ bloc comme ensemble de validation et à s'entraîner sur l'union des $k-1$ ensembles restants. Les hyperparamètres sont sélectionnés en tenant compte de la performance sur l'ensemble de validation, moyennée sur les k étapes. Pour plus de détails sur la sélection des modèles, nous renvoyons le lecteur intéressé à Hastie et al. (2001, Chapter 7) et Arlot et al. (2010).

F.1.6 Apprendre avec des Fonction de Similarité $(\epsilon, \gamma, \tau)-$Bonnes

Nous résumons la théorie d'apprentissage avec des fonctions de similarité $(\epsilon, \gamma, \tau)-$bonnes, proposée dans Balcan et al. (2008b,a). Cette théorie considère deux fonctions de perte: la violation de marge $l(y, y') := [y \cdot y' < \gamma]$ pour $\gamma > 0$ donné, et la $\gamma-$perte *hinge* $l(y, y') := (1 - \frac{yy'}{\gamma})$. Dans tout ce qui suit, une fonction de similarité est une fonction $K : \mathbb{X}^2 \to [-1, 1]$.

F.1.6.1 Avec la fonction de perte violation de marge

Définition F.1.4. *(Balcan et al., 2008a, Definition 6) Une fonction de similarité K est (ϵ, γ, τ)-bonne pour la distribution $\mathcal{D}$ s'il existe une fonction indicatrice (probabiliste) R d'un ensemble de "points de référence" telle que:*

$$\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}} [y \cdot g(\mathbf{x}) < \gamma] \leq \epsilon, \tag{F.15}$$

$$\mathop{\mathbb{P}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}} \big[R(\mathbf{x}') = 1\big] \geq \tau, \tag{F.16}$$

où $g(\mathbf{x}) = \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}} [y' K(\mathbf{x}, \mathbf{x}') | R(\mathbf{x}') = 1]$.

Pour une distribution équilibrée entre classes positive et négative, on peut montrer que la définition précédente équivaut à dire que pour la majorité des exemples (au moins une proportion $1-\epsilon$), un exemple $\mathbf{x}$ est en moyenne plus similaire (au sens de K) à des points de référence ayant la même étiquette, qu'à ceux avec une étiquette opposée, par une vaste marge (au moins 2γ). Ainsi, la définition ci-dessus reflète rigoureusement l'intuition de "ceux qui se ressemblent s'assemblent".

Le résultat suivant montre comment utiliser de telles fonctions de similarité pour la classification linéaire, d'une manière qui rappelle l'astuce du noyau.

Théorème F.1.2. *(Balcan et al., 2008a, Theorem 8) Soit K une fonction de similarité $(\epsilon, \gamma, \tau)-$bonne pour une distribution $\mathcal{D}$. Pour $\delta > 0$, soit $L = \{\mathbf{x}'_1, ..., \mathbf{x}'_{n'}\} \sim \mathcal{D}_{\mathbb{X}}$ un échantillon non étiqueté de taille $n' = \frac{2}{\tau}\log\left(\frac{2}{\delta}\right)\left(1 + \frac{8}{\gamma^2}\right)$. Considérons l'application l'application:*

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

Alors, avec une probabilité d'au moins $1 - \delta$ sur le tirage de L, il existe $\mathbf{w} \in \mathbb{R}^n$ tel que $\|\mathbf{w}\|_1 = 1$ et

$$\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y\left\langle \mathbf{w}, \phi^L(\mathbf{x})\right\rangle < \gamma\right] \leq \epsilon + \delta. \tag{F.17}$$

En d'autres termes, la distribution de données induite par ϕ^L a un séparateur linéaire réalisant un risque de violation de marge d'au plus $\epsilon + \delta$ à la marge γ.

Le résultat précédent est analogue à l'astuce du noyau, dans le sens où les données sont transformées dans un espèce où l'on espère qu'elles sont linéairement séparables.

F.1.6.2 Convexification avec la perte hinge

D'un point de vue pratique, en accédant à un échantillon S, on peut tirer $L \sim \mathcal{D}_{\mathbb{X}}^{n'}$ comme une sous ensemble non étiqueté de S, et chercher un classifieur qui minimise le taux empirique de violation de marge dans l'espace induit par ϕ^L:

$$\min_{\mathbf{w}\in\mathbb{R}^{n'}} \frac{1}{m}\sum_{i=1}^{m}[y_i\left\langle \mathbf{w}, \phi^L(\mathbf{x}_i)\right\rangle < \gamma]. \tag{F.18}$$

Cependant, résoudre ce problème est NP-difficile comme mentionné précédemment quand nous avons introduit la perte de classification erronée. Pour y remédier, Balcan et al. (2008b,a) ont introduit des notions plus adéquates à l'optimisation, en considérant la perte hinge $l_+(y, y') := (1 - y\dot{y}')_+$.

Définition F.1.5.

Théorème F.1.3. *(Balcan et al., 2008a, Definition 7) Une fonction de similarité K est (ϵ, γ, τ)-bonne en perte hinge pour le problème (distribution) $\mathcal{D}$ s'il existe une fonction indicatrice probabiliste R d'un ensembles de "points de références" tel que:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[l_+\left(\frac{g(\mathbf{x})}{\gamma}, y\right)\right] \leq \epsilon, \tag{F.19}$$

$$\mathop{\mathbb{P}}_{\mathbf{x}'\sim\mathcal{D}_{\mathbb{X}}}\left[R(\mathbf{x}') = 1\right] \geq \tau, \tag{F.20}$$

où $g(\mathbf{x}) = \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}}[y'K(\mathbf{x}, \mathbf{x}')|R(\mathbf{x}') = 1]$.

Comme pour le cas de la perte de violation de marge, les similarité bonnes en perte hinge permettent de transformer les données de manière qu'elles soient linéairement séparable, comme énoncé par le théorème suivant.

Théorème F.1.4. *(Balcan et al., 2008a, Theorem 11) Soit K une fonction de similarité $(\epsilon, \gamma, \tau)-$bonne en perte hinge pour le problème $\mathcal{D}$. Pour tout $\epsilon_1 > 0$ et $0 < \delta < \frac{\gamma\epsilon_1}{4}$, soit $L = \{\mathbf{x}'_1, ..., \mathbf{x}'_{n'}\} \sim \mathcal{D}_{\mathbb{X}}^{n'}$ un échantillon non étiqueté de taille $n' = \frac{2}{\tau}\log\left(\frac{2}{\delta}\right)\left(1 + \frac{16}{(\epsilon_1\gamma)^2}\right)$ de points de référence. Considérons ϕ^L de Théorème F.1.2. Alors, avec une probabilité d'au moins $1 - \delta$ sur le tirage de L, il existe $\mathbf{w} \in \mathbb{R}^n$ tel que $\|\mathbf{w}\|_1 = 1$ et:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[l_+\left(\frac{\left\langle \mathbf{w}, \phi^L(\mathbf{x})\right\rangle}{\gamma}, y\right)\right] \leq \epsilon + \epsilon_1. \tag{F.21}$$

Autrement dit, la distribution induite par ϕ^L an un séparateur linéaire réalisant un risque hinge d'au plus $\epsilon + \epsilon_1$ à la marge γ.

Étant donné une fonction de similarité $(\epsilon, \gamma, \tau)-$bonne en perte hinge, et un échantillon $L \sim \mathcal{D}_{\mathbb{X}}^{n'}$, le théorème suivant justifie l'algorithme suivant pour cherche un classifieur linéaire dans l'espace induit par ϕ^L:

$$\min_{\mathbf{w} \in \mathbb{R}^{n'}} \quad \sum_{i=1}^{m} l_+ \left(\left\langle \mathbf{w}, \phi^L(\mathbf{x}_i) \right\rangle, y_i \right) \tag{F.22}$$

$$\text{soumis à} \quad \|\mathbf{w}\|_1 \leq \frac{1}{\gamma}. \tag{F.23}$$

Il s'agit d'un problème d'optimisation linéaire, où la contrainte sur la norme-1 de $\mathbf{w}$ garantit la parcimonie de la solution.

F.1.6.3 Connexion à l'apprentissage de similarité

Les noyaux et les métriques, utilisés respectivement par les algorithmes SVM et k-NN, reflètent une certaine mesure de similarité entre les exemples. De telles mesures sont souvent fixées au préalable et peuvent être non représentatives de la géométrie du problème de classification considéré. Pour adresser ces limitations, les domaines d'*apprentissage de métriques* (Bellet et al., 2013; Kulis, 2013) et d'*apprentissage de noyaux* (Abbasnejad et al., 2012; Gönen and Alpaydın, 2011) sont apparus dans le but d'étudier l'apprentissage d'une fonction de similarité qui rapprochent les exemples de la même classe et éloignent celles ayant des classes opposées.

Les similarités $(\epsilon, \gamma, \tau)-$bonnes sont liées à ces deux domaines car les conditions qu'elles vérifient visent à ce que les instances de classes opposées restent éloignées les unes des autres par d'une certaine marge. Cependant, elles n'imposent pas de conditions sur la distance entre instances de la même classe, et la contraintes qui leurs sont imposées doivent tenir en moyennes, *i.e.* elles sont globales, à l'opposé de contraintes adressées dans l'apprentissage de métriques. De plus, ce dernier domaine inclut plusieurs approches visant à apprendre une distance de Mahlanobis, ce qui ajoute la contrainte du caractère positif semi-défini (PSD), ce qui n'est pas requis par les similarité $(\epsilon, \gamma, \tau)-$bonnes. Notons que ce caractère doit aussi être vérifié par les approches d'apprentissage de noyaux.

F.1.6.4 Quelques travaux basées sur les similarités $(\epsilon, \gamma, \tau)-$ bonnes

Apprentissage de similarité bilinéaire pour la classification linéaire parcimonieuse Dans Bellet et al. (2012), on présente une approche pour apprendre une fonction de similarité bilinéaire $(\epsilon, \gamma, \tau)-$bonne, avec une régularisation quadratique. Les auteurs dérivent ensuite des garanties de généralisation en se basant sur la théorie de la stabilité Bousquet and Elisseeff (2002).

Classification régularisée garantie Guo and Ying (2014) fournissent une analyse théorique d'un problème similaire à celui de Bellet et al. (2012), avec une régularisation non restreinte à la norme de Frobenius, et où la matrice de la fonction bilinéaire est restreinte à être symétrique. Ils dérivent des garanties de généralisation en se basant sur la complexité de Rademacher pour la similarité apprise et pour le classifieur linéaire qui en résulte.

Apprendre conjointement la fonction de similarité et le classifieur induit Le contexte de ce travail, présenté dans Nicolae et al. (2015), est la classification semi-supervisée, où l'on suppose que l'ensemble d'entraînement n'est que partiellement étiqueté. Les auteurs dérivent un algorithme pour apprendre conjointement une fonction de similarité bilinéaire K et le classifieur dans le nouvel espace induit par K, tout

en restreignant la matrice de la similarité bilinéaire à être diagonale. Des garanties de généralisation sont dérivées en se basant sur la complexité de Rademacher.

F.2 Adaptation de Domaine

Le chapitre précédent concerne l'un des scénarios les plus considérés en apprentissage automatique, à savoir l'apprentissage supervisé. Néanmoins, les fondations théoriques de ce dernier ne couvrent pas quelques problèmes pratiques réels où les processus produisant les données diffèrent entrent les ensembles d'apprentissage et de test, correspondant aux domaines source et cible. Ceci est le cas de l'*Adaptation de Domaine* (DA) que nous présentons dans ce chapitre. Nous commençons par la formalisation théorique du problème d'adaptation de domaine dans le cas particulier où l'on dispose d'un seul domaine source, les espaces de caractéristiques et de sortie restent les mêmes pour les deux domaines, et où les étiquettes du domaine cible sont totalement indisponibles au moment d'apprentissage. De plus, nous couvrons quelques conditions suffisantes et d'autres nécessaires au succès de l'adaptation. Puis, nous traçons d'une manière non exhaustive un portrait de différents algorithmes de DA proposés dans la littérature.

F.2.1 Cadre théorique

On suppose que les domaines source et cible ont des distributions jointes $\mathcal{S}$ et $\mathcal{T}$ sur $\mathbb{X} \times \mathbb{Y}$, où $\mathbb{X}$ et $\mathbb{Y}$ sont les même que ceux du chapitre sur l'apprentissage supervisé. Les seules informations dont l'on dispose à propos de ces distributions sont données par les échantillons qu'elles produisent: $S = \{(\mathbf{x}_{s,i}, y_{s,i})\}_{i=1}^{m_s} \sim \mathcal{S}^{m_s}$ et $T_u = \{\mathbf{x}_{t,j}\}_{j=1}^{m_t} \sim \mathcal{T}_{\mathbb{X}}{}^{m_t}$, où ce dernier est non étiqueté, et donc il est tiré de la marginale $\mathcal{T}_{\mathbb{X}}$. Étant donné un espace d'hypothèses $\mathbb{H}$ et une fonction de perte l, l'objectif est de réaliser un faible $l-$risque sur la distribution cible en dépit de l'absence des données étiquetées.

Intuitivement, même si la source et la cible sont différents, elles devraient être reliées d'une certaines manières pour que l'adaptation soit réussie. Dans la suite, nous formalisons la notion de deux domaines reliés, où leur relation est exprimée via des suppositions sur les marginales et les probabilités conditionnelles des étiquettes. Puis, nous présentons quelques résultats montrant la suffisance ou la nécessité de certaines de ces suppositions pour une adaptation réussie, *i.e.* pour trouver une hypothèse ayant un risque faible sur le domaine cible.

F.2.1.1 Hypothèses reliant les marginales

Dominance de la cible par la source La distribution de la cible est absolument continue par rapport à celle de la source, *i.e.* $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$. Dans ce cas, il est possible de considérer la dérivée de Radon-Nikodym (Nikodym, 1930) $\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}$,égale en pratique au rapport des densités. Elle est la base pour les approches par pondération.

Similarité Plutôt que de supposer une relation de dominance entre les distributions, les marginales des domaines dans ce cas sont supposées proches au sens d'une certaine divergence. Deux familles principales de divergence servant à comparer les distributions de probabilité sont les $\phi-$*divergences* (Csiszár, 1967) et les *métriques intégrales de probabilités* (IPM). La première famille suppose que l'une des distributions domine l'autre pour avoir des valeurs finies, une exigence qui n'est pas nécessaire à la deuxième famille de divergences. De plus, les IPM sont un choix populaire pour établir des bornes théoriques de DA, ainsi que pour proposer des algorithmes.

F.2.1.2 Hypothèses reliant les conditionnelles des étiquettes

"Covariate shift" Revient à dire que le changement de distribution entre source et cible est uniquement due au changement des marginales des caractéristiques, *i.e.*

$$\mathcal{S}_{\mathbb{X}} \neq \mathcal{T}_{\mathbb{X}} \quad \text{and} \quad \mathcal{S}_{\mathbb{Y}|\mathbf{x}} = \mathcal{T}_{\mathbb{Y}|\mathbf{x}}. \tag{F.24}$$

Erreur jointe idéale faible C'est la supposition qu'il existe une hypothèse $h \in \mathbb{H}$ qui a une bonne performance sur les deux domaines, *i.e.* la quantité

$$\min_{h\in\mathbb{H}} \mathfrak{E}^{01}_{\mathcal{S}}(h) + \mathfrak{E}^{01}_{\mathcal{T}}(h),$$

qui est le taux de classification erronée (à un facteur près) est faible.

F.2.1.3 Mesurer la divergence entre les marginales des caractéristiques

La similarité entre domaines est évaluée en comparant leurs distributions de probabilité, plus précisément les marginales des caractéristiques car on dispose pas d'étiquettes cible pendant l'entraînement. En DA, les IPM sont les divergences les plus utilisées, et les définit ci-dessous:

Définition F.2.1 (Métrique Intégrale de Probabilités)**.** *Soit $\mathbb{F}$ une famille de fonctions bornée mesurables par rapport à $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{T}_{\mathbb{X}}$. La Métrique Intégrale de Probabilités (IPM) associée à $\mathbb{F}$ entre $\mathcal{S}_{\mathbb{X}}$ et $\mathcal{T}_{\mathbb{X}}$ est définie par*

$$\mathrm{IPM}_{\mathbb{F}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) = \sup_{f\in\mathbb{F}} \left| \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbb{X}}}[f(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}}}[f(\mathbf{x})] \right|.$$

Pour tout $\mathbb{F}$ comme dans la définition ci-dessus, $IPM_{\mathbb{F}}$ est une fonction symétrique de $\mathcal{S}_{\mathbb{X}}$ et $\mathcal{T}_{\mathbb{X}}$, et elle vérifie l'inégalité triangulaire, ce qui fait d'elle une pseudo-métrique sur l'ensemble des distributions de probabilité sur $\mathbb{X}$. Quand $\mathbb{F}$ est assez riche, $IPM_{\mathbb{F}}$ devient une métrique. Dans la Table F.1, nous présentons 4 IPM très utilisées en DA, qui diffèrent selon le choix de $\mathbb{F}$, et pour ce résumé nous donnons plus détails uniquement pour la $\mathbb{H}-$divergence et la divergence Wasserstein, aussi appelée distance de Wassertein car il s'agit d'une métrique.

Name	Espace de fonctions $\mathbb{F}$	Détails
$\mathbb{H}$-divergence	$\mathbb{H}$: espace de classifieurs binaires	$\mathbb{H}$ a une dimension VC finie, étiquettes dans {0,1}
$l-$écart	$\{\mathbf{x} \mapsto l(h(\mathbf{x}), h'(\mathbf{x})); h, h' \in \mathbb{H}\}$	$l : \mathbb{X} \to \mathbb{R}_+$ symétrique, inégalité triangulaire
MMD	$\{f : \mathbb{X} \to \mathbb{R}; f \in \mathbb{V}; \|f\|_{\mathbb{V}} \leq 1\}$	$\mathbb{V}$ est un RKHS d'un noyau universel
Wasserstein-1	$\{f : \mathbb{X} \to \mathbb{R}; f$ est 1-Lipschitzienne$\}$	$(\mathbb{X}, d)$ est un espace métrique

Table F.1: Quelques Métriques Intégrales de Probabilité notables utilisées en DA.

$\mathbb{H}-$divergence Introduite dans le travail fondateur de Ben-David et al. (2010), c'est une pseudo-métrique entre $\mathcal{S}_{\mathbb{X}}$ et $\mathcal{T}_{\mathbb{X}}$, définie par rapport à un espace d'hypothèses binaires $\mathbb{H}$ ayant des valeurs dans $\{0, 1\}$. Elle peut être estimée à partir d'échantillons de taille finie si $\mathbb{H}$ a une dimension VC finie (Ben-David et al., 2010, Lemma 1). De plus, elle bénéficie d'une propriété intéressante: elle correspond, à quelques constantes additives et multiplicatives près, à la justesse[1] du meilleur classifieur de $\mathbb{H}$ essayant de distinguer entre exemples provenant de $\mathcal{S}_{\mathbb{X}}$ et de $\mathcal{T}_{\mathbb{X}}$ (Ben-David et al., 2010, Lemma 2). Ainsi, quand ce meilleur classifieur échoue à les distinguer, c'est que les deux domaines sont similaires à un certain degré.

[1] 1-taux de classification erronée

$l-$**écart** Cette mesure de divergence, proposée par Mansour et al. (2009b), est aussi liée à l'espace d'hypothèses considéré, à l'instar de la $\mathbb{H}-$divergence, mais elle est plus générale car non restreinte à la perte de classification erronée. En effet, elle concerne même les problèmes de régression, et elle exige une fonction de perte symétrique et vérifiant l'inégalité triangulaire. Elle est estimable à partir d'échantillons de taille finie (Mansour et al., 2009b, Corollary 7), un résultat qui fait appel à la complexité de Rademacher.

Écart Moyen Maximal Il s'agit sans doute de l'une des IPM les plus utilisées dans les approches DA proposées dans la littérature, et a été introduite par Gretton et al. (2012). Elle fait appel aux noyaux universels (Steinwart, 2001), un type particulier de noyaux correspondant à un RKHS dense dans l'espace des fonctions continues de $\mathbb{X}$ dans $\mathbb{R}$. Le supremum dans leur forme bénéficie d'une formule explicite exprimée uniquement en terme de produits scalaires, *i.e.* noyaux, entre exemples (Gretton et al., 2012, Lemma 6), d'une manière rappelant l'astuce du noyau. Elle est estimable à partir d'échantillons finis (Gretton et al., 2012, Theorem 7).

Distance de Wasserstein Cette divergence est étroitement liée au problème du transport optimale (Monge, 1781; Kantorovich, 1942). C'est une IPM pour l'espace des fonctions 1-Lipschitziennes par rapport à une métrique donnée $d : \mathbb{X}^2 \to \mathbb{R}_+$. Sa définition lui permet de refléter la géométrie de l'espace sous-jacent grâce à sa liaison avec la métrique d. Elle est empiriquement estimable (Bolley et al., 2007, Theorem 1.1), mais la vitesse de convergence de son estimation empirique dépend de la dimension de l'espace ambiant: plus la dimension augmente, moins la convergence est rapide.

Le problème du transport optimal constitue la forme primale de la distance de Wasserstein et la forme avec un supremum (dans l'expression des IPM) est sa forme duale. Afin de définir le problème de transport optimal tel que posé par Kantorovitch (Kantorovich, 1942) comme une relaxation de celui de Monge (Monge, 1781), nous définissons d'abord l'ensemble des plans de transport.

Définition F.2.2 (Ensemble de plans de transport)**.** *Soit $\mathfrak{P}(\mathbb{X}^2)$ l'ensemble de distributions de probabilité sur $\mathbb{X}^2$, and let $\pi_1 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_1$ and $\pi_2 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_2$. L'ensemble de plans de transport entre $\mathcal{S}_\mathbb{X}$ et $\mathcal{T}_\mathbb{X}$, noté $\Pi(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X})$, est:*

$$\Pi(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X}) := \{\mathcal{P} \in \mathfrak{P}(\mathbb{X}^2); \pi_1 \# \mathcal{P} = \mathcal{S}_\mathbb{X} \text{ et } \pi_2 \# \mathcal{P} = \mathcal{T}_\mathbb{X}\}. \tag{F.25}$$

En particulier, pour les distributions empiriques $\hat{S}_u$ and $\hat{T}_u$ associées aux échantillons $S_u \sim {\mathcal{S}_\mathbb{X}}^{m_s}$ et $T_u \sim {\mathcal{T}_\mathbb{X}}^{m_t}$, $\Pi(\hat{S}, \hat{T})$ est appelé l'ensemble de matrices de transport, défini par:

$$\Pi(\hat{S}_u, \hat{T}_u) := \left\{ \mathbf{P} \in \mathbb{R}_+^{m_s \times m_t}; \mathbf{P}\mathbf{1}_{m_t} = \frac{\mathbf{1}_{m_s}}{m_s}; \mathbf{P}^T \mathbf{1}_{m_s} = \frac{\mathbf{1}_{m_t}}{m_t} \right\}. \tag{F.26}$$

Avec cette définition, on peut énoncer le résultat de la dualité de Kantorovitch-Rubinstein ci dessous.

Théorème F.2.1 (Dualité de Kantorovich-Rubinstein)**.** *La forme primale de la distance de Wasserstein W_1 est*

$$W_1(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X}) = \inf_{\mathcal{P} \in \Pi(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X})} \mathop{\mathbb{E}}_{(\mathbf{x}_s, \mathbf{x}_t) \sim \mathcal{P}} \left[d(\mathbf{x}_s, \mathbf{x}_t) \right]. \tag{F.27}$$

Pour les distributions empiriques associées à $S_u \sim {\mathcal{S}_\mathbb{X}}^{m_s}$ et $T_u \sim {\mathcal{T}_\mathbb{X}}^{m_t}$, on a:

$$W_1(\hat{S}_u, \hat{T}_u) = \inf_{\mathbf{P} \in \Pi(\hat{S}_u, \hat{T}_u)} \sum_{i=1}^{m_s} \sum_{j=1}^{m_t} (\mathbf{P})_{ij} d(\mathbf{x}_{s,i}, \mathbf{x}_{t,j}) = \inf_{\mathbf{P} \in \Pi(\hat{S}_u, \hat{T}_u)} \langle \mathbf{P}, \mathbf{D} \rangle, \tag{F.28}$$

où $(\mathbf{D})_{ij} := d(\mathbf{x}_{s,i}, \mathbf{x}_{t,j})$

Le problème de transport optimal consiste ainsi à trouver une manière de déplacer des unités de masses (de probabilité) entre les deux distributions concernée qui minimise le coût global du transport $\mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}}[d(\mathbf{x}_s,\mathbf{x}_t)]$, où le transport d'une unité de masse est exprimé par la métrique d considérée.

Autres divergences Bien que nous nous avons uniquement présenté des divergences de la famille IPM, d'autres ont été considérées dans la littérature, comme la divergence de Kullback-Leibler(Kullback and Leibler, 1951) utilisée dans Sugiyama et al. (2007), et plus généralement les divergences de Rényi (Hero et al., 2001) considérées dans Cortes et al. (2010). Du point de vue PAC-Bayesien, le $\rho-$désaccord a été proposé dans Germain et al. (2013, 2016a). Plus récemment, Zhang et al. (2019) propose une divergence semblable comparable en forme à celle de Ben-David et al. (2010), mais qui tient la marge de classification en compte.

F.2.1.4 Suffisance: borner le risque cible

Vu l'absence d'étiquettes pour l'échantillon cible, l'étude de la performance d'un modèle sur celle-ci doit être menée en fonctions des quantités disponibles, à savoir les exemples cible non étiquetés les exemples source étiquetés. Plusieurs travaux dans la littérature fournissent des bornes sur un risque dans le domaine cible ayant la forme générique suivante pour un classifieur $h \in \mathbb{H}$:

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{S}}(h) + \text{divergence}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + t(\mathcal{S}, \mathcal{T}). \tag{F.29}$$

Ainsi, pour avoir une bonne performance sur le domaine cible, il suffit que tous les trois termes du membre de droite dans l'inégalité précédente soient petits. Le premier est un risque sur le domaine source, et il renferme l'information relative aux étiquettes. Le deuxième est une divergence entre les marginales, généralement donnée par une IPM. Les deux termes précédents sont empiriquement estimables, au contraire du troisième terme, pour lequel l'impossibilité d'accès est inévitable, et qui doit être faible pour que l'adaptation soit possible. Dans la suite, nous présentons brièvement quelques bornes en DA ayant la forme générale présentée précédemment.

Bornes basées sur la $\mathbb{H}\Delta\mathbb{H}-$divergence Étant donné un espace d'hypothèses binaires $\mathbb{H}$, Ben-David et al. (2010) définissent l'espace $\mathbb{H}\Delta\mathbb{H}$ comme l'espace de désagréments entre deux hypothèses de $\mathbb{H}$:

$$\mathbb{H}\Delta\mathbb{H} := \{h \oplus h'; h, h' \in \mathbb{H}\} = \{\left|h - h'\right|; h, h' \in \mathbb{H}\} = \{[h \neq h']; h, h' \in \mathbb{H}\}, \tag{F.30}$$

et ils prouvent le théorème suivant

Théorème F.2.2. *(Ben-David et al., 2007, Theorem 2) Étant donné un espace d'hypothèses binaires $\mathbb{H}$ à valeurs dans $\mathbb{Y} = \{0, 1\}$, on a pour tout $h \in \mathbb{H}$:*

$$\mathfrak{E}^{01}_{\mathcal{T}}(h) \leq \mathfrak{E}^{01}_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathbb{H}\Delta\mathbb{H}}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + \lambda_{\mathbb{H}}(\mathcal{S}, \mathcal{T}), \tag{F.31}$$

où

$$\lambda_{\mathbb{H}}(\mathcal{S}, \mathcal{T}) = \min_{h\in\mathbb{H}} \mathfrak{E}^{01}_{\mathcal{S}}(h) + \mathfrak{E}^{01}_{\mathcal{T}}(h). \tag{F.32}$$

Ce dernier résultat montre qu'il suffit d'avoir une divergence $\mathbb{H}\Delta\mathbb{H}$ et une erreur idéale jointe faibles pour réussir à l'adaptation.

Bornes basées sur sur le $l-$écart et la MMD Mansour et al. (2009b, Theorem 8) montre la suffisance d'un faible $l-$écart et d'un faible désagrément entre les meilleurs hypothèses sur la source et la cible par rapport à la distribution source, ainsi que

d'une faible erreur minimale sur la cible, pour l'adaptation. Quant à la MMD, utilisée abondamment pour des algorithmes de DA, son intérêt théorique est prouvé dans Redko (2015, Théorème 6.10), un résultat comparable à celui de Ben-David et al. (2010).

Bornes basées sur la distance de Wasserstein La distance de Wasserstein, récemment de plus en plus utilisé dans le DA, a plusieurs justifications théoriques pour une telle application. La première justification est due à Redko et al. (2017, Théorème 2) et est semblable à la borne de Ben-David et al. (2010). Ce résultat concerne un espace RKHS, alors que celui de Shen et al. (2018, Theorem 1) repose sur le caractère Lipschitz des hypothèses. Finalement, la borne de Courty et al. (2017, Théorème 3.1) n'a pas la même forme que celles que nous avons mentionnées, et repose sur d'autres suppositions pour le succès de l'adaptation.

Autres bornes Des bornes dans le cadre PAC-Bayesien sont dérivées dans Germain et al. (2013, 2016a). En particulier, dans la borne de Germain et al. (2016a, Théorème 3), la divergence entre source et cible multiplie un terme non supervisé visant à avoir une marge absolue aussi grande que possible pour les points du domaine cible. Un autre travail important est celui de Zhang et al. (2019), où les auteurs dérivent une borne d'adaptation de domaine (Zhang et al., 2019, Proposition 3.3) qui fait intervenir la marge de classification, et qui généraliser celle de Ben-David et al. (2010).

F.2.1.5 Nécessité: difficulté de l'adaptation

Dans Ben-David et al. (2010), les auteurs montrent rigoureusement la nécessité que la $\mathbb{H}\Delta\mathbb{H}-$divergence et l'erreur idéale jointe soient toutes les deux faibles, en plus de leur suffisance prouvée dans Ben-David et al. (2010), pour le succès de l'adaptation. En particulier, ils définissent une notion d'apprenabilité (ϵ, δ, m, n) pour les tâches d'adaptation et se basent dessus pour prouver des résultats négatifs dans Ben-David et al. (2010, Théorème 1, Théorème 2)

F.2.2 Avancées algorithmiques

La plupart des approches DA existantes reposent sur l'idée commune d'aligner les deux domaines afin de réduire l'écart entre leurs distributions. En effet, après un tel alignement, on ramène la tâche à accomplir à un problème de classification supervisée. D'autres approches reposent sur d'autres hypothèses telles que la *cluster assumption* (Ben-David and Urner, 2014).

Le moins que l'on puisse dire est que le nombre d'algorithmes de DA est énorme. En effet, selon quelle mesure de divergence on utilise entre les domaines, quelles variables on modifie pour les aligner, et si l'alignement et la classification sont faits simultanément, plusieurs approches ont été proposées.

F.2.2.1 Approches par re-pondération

Ces approches reposent sur la supposition *covariate shift* et sur la domination $\mathcal{S}_{\mathbb{X}} \gg \mathcal{T}_{\mathbb{X}}$. Combiner ces deux hypothèses permet de prouver que le risque sur la cible est le risque sur la source après avoir re-pondéré les exemples. Pour un échantillon source non étiqueté, ça consiste en la transformation suivante

$$\hat{S}_u := \frac{1}{m}\sum_{i=1}^{m}\delta_{\mathbf{x}_i} \xrightarrow{\text{Re-pondération}} \begin{cases} \hat{S}_u^{\mathbf{w}} := \frac{1}{m}\sum_{i=1}^{m} w_i\delta_{\mathbf{x}_i}. \\ \hat{S}_u^{\mathbf{p}} := \sum_{i=1}^{m} p_i\delta_{\mathbf{x}_i} \quad \text{où } \mathbf{p} \in \Delta_{m_s}, \end{cases} \tag{F.33}$$

visant à rapprocher les deux distributions. Cette approche est parmi les premières, utilisée par exemple dans Shimodaira (2000); Huang et al. (2007); Sugiyama et al. (2007). Concrètement, ces approches se basent sur une estimation du rapport $\frac{\mathrm{d}\mathcal{T}_{\mathbb{X}}}{\mathrm{d}\mathcal{S}_{\mathbb{X}}}$, qui peut être

indirecte Les poids de chaque de domaine sont estimés, puis on prends le rapport (Shimodaira, 2000; Sugiyama et al., 2005; Baktashmotlagh et al., 2014).

directe Le rapport est estimé directement Tsuboi et al. (2009); Kanamori et al. (2009). En particulier, la MMD a servi pour ces approches dans Huang et al. (2007); Gretton et al. (2009b); Chu et al. (2013); Gong et al. (2013); Yan et al. (2017), grâce à sa forme explicite.

F.2.2.2 Approches par transformation d'espaces de caractéristiques

Au lieu de repondérer les instances, on agit sur les caractéristiques dans ces approches. On distingues des approches symétriques où une transformation est appliquée aux instances source et cible pour les aligner, et les approches asymétriques où la source est transformée pour correspondre à la cible (Table F.2). L'alignement en agissant sur les caractéristiques

Approach	**Asymétrique**		**Symétrique**	
Transformation	$\phi : \mathbb{X} \to \mathbb{X}$		$\phi : \mathbb{X} \to \mathbb{U}$	
Représentation de domaine	distribution	autre	distribution	autre
Objectif	$\phi\#\mathcal{S} \approx \mathcal{T}$	$F(\phi\#\mathcal{S}) \approx F(\mathcal{T})$	$\phi\#\mathcal{S} \approx \phi\#\mathcal{T}$	$F(\phi\#\mathcal{S}) \approx F(\phi\#\mathcal{T})$
Restriction	Respecter l'information des classes			

Table F.2: Les deux principales familles d'approches DA basées sur la transformation de caractéristiques.

peut se faire via des heuristiques (Blitzer et al., 2006, 2007), en minimisant une certaine divergence empirique entre distributions, ou en utilisant d'autres représentations pour les domaines.

Aligner les distribution empiriques Une approche asymétrique importante est celle proposée par Courty et al. (2016), où une matrice de transport optimal est utilisée pour appliquer une transformation aux exemples du domaine source. Concrètement, le problème suivant est résolu:

$$\min_{\mathbf{P}\in\Pi(\hat{S}_u,\hat{T}_u)} \underbrace{\sum_{i,j}(\mathbf{P})_{ij}c(\mathbf{x}_{s,i},\mathbf{x}_{t,j})}_{\text{Coût de transport}} + \lambda \underbrace{\sum_{i,j}(\mathbf{P})_{ij}(-\log((\mathbf{P})_{ij}))}_{\text{Régularisation entropique}} + \eta \underbrace{\Omega_c(\mathbf{P})}_{\substack{\text{Régularisation} \\ \text{liée aux classes}}}, \quad \text{(F.34)}$$

Quant aux approches symétriques, la MMD est un choix très répandu (Pan et al., 2008, 2011; Baktashmotlagh et al., 2016).

Autres formes d'alignement Plûto que de représenter les domaines par des distributions, quelques approches les représentent par des sous-espaces, comme Gong et al. (2012); Gopalan et al. (2011); Fernando et al. (2013). L'idée de ces approches est d'agir sur les vecteurs de base de tels espaces. Une approche qui y est reliée est proposée dans Sun et al. (2016), où chaque domaine est représenté par sa matrice de covariance.

F.2.2.3 Aligner et classifier simultanément

Dans ces approches, l'alignement et l'apprentissage du classifieur se font simultanément, produisant un bon classifieur pour le domaine cible. Ces approches sont utilisées dans l'apprentissage superficiel, mais on les retrouve majoritairement dans les approches par réseaux artificiels profonds. Les plus simples de ces approches consistent à résoudre un problème de la forme suivante:

$$\min_{h\in\mathbb{H}} \mathfrak{E}^l_{\hat{S}}(h) + \eta\,\text{divergence}_h(\hat{S}_u,\hat{T}_u), \quad \text{(F.35)}$$

où $\eta > 0$ est un hyper-paramètre contrôlant un compromis entre un premier terme poussant à avoir une bonne classification sur le domaine source, et un deuxième terme minimisant une certaine divergence et traduisant l'alignement des domaines. Ce dernier terme dépend potentiellement de l'hypothèse h en question. Dans ce cas, la fonction coût a une ressemblance frappante avec la forme générale de bornes en DA, présentée précédemment. D'ailleurs, quelques approches dans la littérature sont basées sur la minimisation de la partie estimable de la borne (Germain et al., 2013, 2016a; Courty et al., 2017).

Apprentissage superficiel Ceci veut dire que l'on utilise les caractéristiques d'origine, ou une transformation fixée a priori avant d'apprendre un classifieur. Parmi les approches dans ce cadre, on cite Morvant et al. (2012); Germain et al. (2013, 2016a); Courty et al. (2017). Dans le dernier travail par exemple, on propose une minimisation d'une distance de Wasserstein entre les distributions des domaines en pseudo-étiquetant le domaine cible:

$$\min_{\substack{h\in\mathbb{H}\\ \mathbf{P}\in\Pi(\hat{S}_u,\hat{T}_u)}} \sum_{i=1}^{m_s}\sum_{j=1}^{m_t}(\mathbf{P})_{ij}\left(\underbrace{l(h(\mathbf{x}_{t,j}),y_{s,i})}_{\text{aligne les étiquettes}} + \alpha \underbrace{c(\mathbf{x}_{s,i},\mathbf{x}_{t,j})}_{\text{aligner les marginales}}\right). \tag{F.36}$$

Cette fonction objectif est la version empirique de la borne de Courty et al. (2017, Théorème 3.1). D'autres approches de bénéficient pas de telles garanties théoriques mais se basent plutôt sur une forte intuition (Quanz and Huan, 2009; Long et al., 2014) qui utilisent une version projeté de la MMD. Les algorithmes de boosting méritent aussi une mention vu leur capacité de découvrir les structures non linéaires (Habrard et al., 2013a).

Approches par apprentissage profond Les réseaux de neurones artificiels (Goodfellow et al., 2016) sont extrêmement puissants pour extraire de nouvelles caractéristiques à partir de données brutes, et on été appliqués avec succès en vision par ordinateur (Lecun et al., 1998; Voulodimos et al., 2018; Grigorescu et al., 2020) en traitement du langage naturel (Zhang et al., 2018; Young et al., 2018), pour ne citer que quelques applications. Ce pouvoir d'extraction est exploité en approches DA pour apprendre des caractéristiques qui, en plus de séparer les classes dans le domaine source, permettent d'aligner les deux domaines. À titre d'exemple, l'approche de Ganin et al. (2016) consiste à apprendre des caractéristiques qui entravent le meilleur classifieur visant à distinguer entre la source et la cible, une idée inspirée du résultat de Ben-David et al. (2010, Lemma 2).

F.2.2.4 Autres approches

La majorité des approches en DA font appel de manière explicite à un alignement des deux domaines,mais pas toutes. Par exemple, plusieurs approches sont basées sur l'auto-étiquetage: le classifieur est initialisé par un qui a une bonne performance sur le domaine source, puis ils est modifié graduellement par l'addition d'exemples cibles et l'élimination d'exemples source selon certains critères (Bruzzone and Marconcini, 2010; Chen et al., 2011b; Habrard et al., 2013b). D'autres approches sont hybrides, et combinent les les idées des différentes catégories déjà présentées (Morvant et al., 2012; Aljundi et al., 2015; Sun and Saenko, 2015).

F.2.2.5 Sélection de modèle

Plusieurs procédures de validation croisées ont été proposées dans la littérature, parmi lesquelles on cite la *validation croisée inversée* (Bruzzone and Marconcini, 2010; Zhong et al., 2010). Elles reviennent à considérer le problème inversé suivant: Étant donné

l'échantillon source étiqueté S et celui cible non étiqueté T_u, soit T^h l'échantillon cible étiqueté par h, et S_u la partie non étiquetée de S. Lorsque h est appris par un algorithme de DA,la validation inversée repose sur la supposition suivante: si une hypothèse h_r, appelée *hypothèse inverse*, apprise de T^h et S_u a une bonne performance sur S_u (ce que l'on peut mesurer vu que l'on dispose d'étiquettes source), alors h est convenable au problème initial. Cette approche, bien qu'elle semble bien correspondre au cadre non supervisé de l'absence d'étiquettes cible, peut ne pas bien fonctionner en pratique (Bousmalis et al., 2016b). Par conséquent, elle n'est pas utilisée dans la plupart des approches DA (Wilson and Cook, 2019, Section 8.2). Plutôt que de l'utiliser, nombre de travaux utilisent soit un sous-ensemble des exemples cible pour sélectionner les hyperparamètres (Bousmalis et al., 2016b, Section 4) ou présentent les résultats pour des hyperparamètres fixés pour plusieurs tâches (Courty et al., 2017, Section 5).

F.2.2.6 Quelques remarques pour conclure

Notre présentation de l'état de l'art de l'adaptation de domaine est certainement pas exhaustive, vu l'arborescence remarquable des variations de cette branche de l'apprentissage statistique. Pour plus de détailsn, nous référons à Kouw and Loog (2019); Redko et al. (2020) pour les aspects théoriques, et à Pan and Yang (2010); Margolis (2011); Weiss et al. (2016); Kouw and Loog (2019); Zhang (2019); Wilson and Cook (2019); Zhuang et al. (2019).

F.3 Fonctions de Similarité (ϵ, γ, τ)-bonnes pour l'Adaptation de Domaine

Dans cette section, nous résumons le chapitre 3, basé sur nos contributions Dhouib and Redko (2018a,b). D'abord, nous étendons l'analyse théorique d'apprentissage avec des fonctions de similarité au cadre d'adaptation de domaine, en introduisant une nouvelle définition d'une fonction de similarité $(\epsilon, \gamma)-$bonne pour l'adaptation de domaine, et en prouvant quelques résultat quantifiant la performance d'une fonction de similarité sur le domaine cible, étant donné sa performance sur le domaine source. En particuliers, nous montrons que si la distribution source domine la cible, alors de nouvelles bornes de DA peuvent être établies. Finalement, nous fournissons une étude rétrospective où nous prouvons théoriquement qu'en dépit de l'attrait théorique et algorithmique, apprendre une fonction de similarité bilinéaire avec pour régularisation une norme de Frobenius avant d'apprendre un classifieur linéaire est redondant, puisque les deux sont essentiellement équivalents.

F.3.1 Fonctions de similarité $(\epsilon, \gamma)-$bonnes en DA

Avec les notations précédentes, on suppose de plus la vérification de la "covariate shift" pour la source et la cible, *i.e.* $\mathcal{S}_{\mathbb{Y}|\mathbf{x}} = \mathcal{T}_{\mathbb{Y}|\mathbf{x}}$ pour $\mathbf{x} \in \mathbb{X}$. Comme suggéré dans Balcan et al. (2008a, Note 2, Theorem 14), les exemples et les points de référence sont potentiellement tirés de distributions différentes. Ainsi, nous proposons une modification de la définition d'une similarité bonne en perte hinge de la manière suivante

Définition F.3.1. *Une fonction de similarité K est $(\epsilon, \gamma)-$bonne en perte hinge pour le problème $(\mathcal{D}, \mathcal{L})$ (où $\mathcal{D}$ et $\mathcal{L}$ sont les distributions respectives des données et des points de référence) si:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[l_+ \left(\frac{g_{\mathcal{L}}(\mathbf{x})}{\gamma}, y \right) \right] \leq \epsilon,$$

où $g_{\mathcal{L}}(\mathbf{x}) := \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{L}} [y' \cdot K(\mathbf{x}, \mathbf{x}')]$.

Cette définition est une généralisation de celle de Balcan et al. (2008a) et traduit l'intuition souvent utilisée pour concevoir des algorithmes de DA, car la distribution $\mathcal{L}$ peut être vue comme un "domaine universel de points de référence" qui consiste en pratique en exemples venant du domaine source ou cible. C'est le cas de quelques contributions en DA (Morvant et al., 2012; Gong et al., 2013; Aljundi et al., 2015), où l'objectif est de réduire le changement de domaine dans l'espace induit par la similarité.

Dans la suite, nous aurons besoin de la fonction de perte

$$l_\gamma : y, y' \mapsto \left(1 - \frac{y \cdot y'}{\gamma}\right)_+,$$

et d'une distribution de probabilité $\mathcal{U}$ sur $\mathbb{X} \times \mathbb{Y}$ telle que $\mathcal{U}_\mathbb{X} \gg \mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X}$ et $\mathcal{U}_{\mathbb{Y}|\mathbf{x}} = \mathcal{S}_{\mathbb{Y}|\mathbf{x}} = \mathcal{T}_{\mathbb{Y}|\mathbf{x}}$. De plus, $\mathfrak{M}_{\mathcal{D},\mathcal{L}}(K)$ dénote la pire marge réalisée par un élément $\mathbf{x} \in \operatorname{supp}\mathcal{D}$ associée à un distribution de points de référence $\mathcal{L}$, i.e:

$$\mathfrak{M}_{\mathcal{D},\mathcal{L}}(K) := \sup_{\mathbf{x} \in \operatorname{supp}\mathcal{D}} l_\gamma(g_\mathcal{L}(\mathbf{x}), y).$$

F.3.1.1 Relier les $l_\gamma-$risques sur la source et la cible

Étant donné une fonction de similarité *e.g.* $-$bonne pour un problème $(\mathcal{S}, \mathcal{L}^s)$, on vise à borner son $l_\gamma-$risque sur le domaine cible pour $(\mathcal{T}, \mathcal{L}^t)$.

Points de référence partagés On traite d'abord le cas particulier où $\mathcal{L}^s = \mathcal{L}^t$ dans le lemme suivant.

Lemme F.3.1 (Mêmes points de référence)**.** *Soit K une fonction de similarité $(\epsilon, \gamma)-$bonne pour le problème $(\mathcal{S}, \mathcal{L})$. Alors K est $(\epsilon + \epsilon', \gamma)$-bonne pour le problème $(\mathcal{T}, \mathcal{L})$, où*

$$\epsilon' = d_{1+,\gamma}(\mathcal{T}, \mathcal{S}) \cdot \mathfrak{M}_{\mathcal{U},\mathcal{L}}(K)$$

et

$$d_{1+,\gamma}(\mathcal{T}, \mathcal{S}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}} \left[\left(\frac{\mathrm{d}\mathcal{T}_\mathbb{X}}{\mathrm{d}\mathcal{U}_\mathbb{X}}(\mathbf{x}) - \frac{\mathrm{d}\mathcal{S}_\mathbb{X}}{\mathrm{d}\mathcal{U}_\mathbb{X}}(\mathbf{x}) \right)_+ [y \cdot g_\mathcal{L}(\mathbf{x}) < \gamma] \right].$$

De plus, si $\mathcal{T}_\mathbb{X} \ll \mathcal{S}_\mathbb{X}$ alors le résultat obtenu vérifie

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

où $d_{\chi^2_+,\gamma}(\mathcal{T}, \mathcal{S}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left[\left(\frac{\mathrm{d}\mathcal{T}_\mathbb{X}}{\mathrm{d}\mathcal{S}_\mathbb{X}}(\mathbf{x}) - 1 \right)_+^2 [y \cdot g_\mathcal{L}(\mathbf{x}) < \gamma] \right].$

Notons que dans le résultat précédent, les espérances définissant les termes de divergence portent uniquement sur le support de la perte hinge, ce qui renforce la dépendance du problème considéré. De plus, dans les deux cas défini par $\mathcal{T}_\mathbb{X} \ll \mathcal{S}_\mathbb{X}$ ou pas, les bornes contiennent $\mathfrak{M}_{\mathcal{U},\mathcal{L}}(K)$ qui est la pire marge réalisée par K sur un exemple de $\operatorname{supp}\mathcal{U}$. Pour plusieurs choix de $\mathcal{U}_\mathbb{X}$, ce terme peut être difficile à contrôler car on ne peut l'estimer qu'à partir des données venant de $\mathcal{S}$. Cette limitation est adressée dans le cas où $\mathcal{T} \ll \mathcal{S}_\mathbb{X}$: le terme de distance entre distributions est multiplié par $\sqrt{\epsilon}$, et donc une fonction de similarité avec une erreur basse peut exploiter la différence entre les distributions des deux domaines.

Distributions différentes pour les points de référence Afin de procéder, nous réécrivons la différence entre $\mathfrak{E}_{\mathcal{T},\mathcal{L}^t}(K)$ et $\mathfrak{E}_{\mathcal{S},\mathcal{L}^s}(K)$ comme suit:

$$\mathfrak{E}_{\mathcal{T},\mathcal{L}^t}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}^s}(K) = \mathfrak{E}_{\mathcal{T},\mathcal{L}^s}(K) - \mathfrak{E}_{\mathcal{S},\mathcal{L}^s}(K) + \mathfrak{E}_{\mathcal{T},\mathcal{L}^t}(K) - \mathfrak{E}_{\mathcal{T},\mathcal{L}^s}(K).$$

La différence entre les deux premiers termes est traitée par le lemme précédent, et il ne reste qu'à borner le reste pour introduire le rôle de la différence entre distributions de points de référence, ce qui donne le théorème suivant.

Théorème F.3.1. *Soit K be une fonction de similarité $(\epsilon, \gamma)-$bonne pour le problème $(\mathcal{S}, \mathcal{L}^s)$. Supposons qu'il existe une probabilité $\mathcal{U}'$ sur $\mathbb{X} \times \mathbb{Y}$ telle que $\mathcal{U}'_{\mathbb{X}}$ domine $\mathcal{L}^s_{\mathbb{X}}$ et $\mathcal{L}^t_{\mathbb{X}}$, et que $\mathcal{U}'_{\mathbb{Y}|\mathbf{x}} = \mathcal{L}_{\mathbb{Y}|\mathbf{x}}{}^s = \mathcal{L}_{\mathbb{Y}|\mathbf{x}}{}^t$. Alors K est $(\epsilon + \epsilon' + \epsilon'', \gamma)$-bonne pour le problème $(\mathcal{T}, \mathcal{L}^t)$, avec*

$$\epsilon'' = \frac{1}{\gamma} d_K(\mathcal{L}^s, \mathcal{L}^t) \text{ et } \epsilon' = d_{1+,\gamma}(\mathcal{T}, \mathcal{S}) \mathfrak{M}_{\mathcal{U},\mathcal{L}^s}(K),$$

où

$$d_K(\mathcal{L}^s, \mathcal{L}^t) := \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{T}_{\mathbb{X}} \\ \mathbf{x}'\sim\mathcal{U}'_{\mathbb{X}}}} \left[\left| \frac{\mathrm{d}\mathcal{L}^s_{\mathbb{X}}}{\mathrm{d}\mathcal{U}'_{\mathbb{X}}}(\mathbf{x}') - \frac{\mathrm{d}\mathcal{L}^t_{\mathbb{X}}}{\mathrm{d}\mathcal{U}'_{\mathbb{X}}}(\mathbf{x}') \right| \left| K(\mathbf{x}, \mathbf{x}') \right| \right].$$

De plus, si $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$, alors le résultat obtenu est vérifié pour

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

Cette borne fait apparaître une divergence entre les distributions de points de référence qui dépend de la fonction de similarité considérée, et suggère que le choix $\mathcal{L}^s = \mathcal{L}^t_{\mathbb{X}}$ est le meilleur pour la minimiser. Cette condition est plutôt intuitive: dans plusieurs algorithmes de DA, les domaines source et cible sont alignés en utilisant un ensemble partagé de composantes invariantes, et les points de référence peuvent être vus comme des points invariants permettant d'adapter la fonction de similarité efficacement à travers les domaines.

F.3.2 Comparaison à d'autres résultats

La plupart des résultat en DA ont la forme suivante (discutée précédemment):

$$\mathfrak{E}^l_{\mathcal{T}}(h) \leq \mathfrak{E}^l_{\mathcal{S}}(h) + \text{divergence}(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}) + t(\mathcal{S}, \mathcal{T}). \tag{F.37}$$

Notre résultat avec la divergence χ^2 est drastiquement différent des bornes avec la forme précédente car il suggère que le risque source impacte directement tous les termes dans la borne. En effet, notre résultat montre l'erreur source peut partiellement tirer parti de la divergence entre les domaines puisqu'elle le multiplie (à une racine carrée près). À notre connaissance, les seuls résultats dans la littérature ayant une telle dépendance multiplicative entre l'erreur source et le terme de divergence sont Mansour et al. (2009c) et Germain et al. (2016a) où des variations de la divergence de Rényi ont été considérées. Contrairement à leurs contributions, notre borne implique un terme de divergence restreint à l'ensemble $[y \cdot g\mathcal{L}(\mathbf{x}) < \gamma]$, ce qui le rend intrinsèquement lié à l'espace d'hypothèses considéré. En outre, Mansour et al. (2009c) suppose que la distribution cible est une combinaison de domaines sources pour obtenir une telle dépendance multiplicative, alors que nous ne prenons pas une telle supposition. De surcroît, la borne de Germain et al. (2016a) fait intervenir un terme non estimable qui est similaire à λ dans l'équation dessus, alors que le terme de pire marge dans notre cas est sujet de l'analyse fournie dans la section suivante.

F.3.3 Analyse empirique du terme de la pire marge

Nous établissons une borne de généralisation sur le terme de pire marge dans le théorème suivant.

Théorème F.3.2. *Soit K une fonction de similarité. Soit $\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K)$ définie comme précédemment. Supposons que $\mathcal{T}_{\mathbb{X}} \ll \mathcal{S}_{\mathbb{X}}$ et que la fonction de répartition F_{l_γ} de la fonction de perte associée à $\mathcal{S}$ and $\hat{L}$ est k fois dérivable en $\mathfrak{M}_{\mathcal{S},\hat{L}}(K)$, and que $k > 0$ l'entier*

naturel minimal tel que $F_{l_\gamma}^{(k)}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K)) \neq 0$. *Alors, pour tout* $\alpha > 1, r \geq 1$, *il existe* $m_0 \geq 1$ *tel que pour tout* $m \geq m_0$, *on avec une probabilité au moins* $1 - \delta$*:*

$$\mathfrak{M}_{\mathcal{S},\mathcal{L}}(K) \leq \mathfrak{M}_{\hat{S},\hat{L}}(K) + \frac{2}{\gamma}\,\mathrm{Rad}_r\left(\mathbb{H}_1(K)\right) + \frac{1}{\gamma}\sqrt{2\frac{\log\left(\frac{4}{\delta}\right)}{r}} + \left(\frac{(-1)^{k+1}\log\left(\frac{2\alpha}{\delta}\right)k!}{F_{l_\gamma}^{(k)}(\mathfrak{M}_{\mathcal{S},\hat{L}}(K))m}\right)^{\frac{1}{k}},$$

où $\mathbb{H}_1(K)$ *l'espace d'hypothèses défini par* $\mathbb{H}_1(K) := \{h_{\mathbf{x}} : \mathbf{x}' \mapsto K(\mathbf{x},\mathbf{x}'), \mathbf{x} \in \operatorname{supp}\mathcal{S}_{\mathbb{X}}\}$.

Ce théorème montre que la vitesse de convergence du terme de pire marge converge vers son homologue théorique est gouvernée par deux termes: un terme de complexité de Rademacher d'une espace induit par la fonction de similarités K et surtout la régularité de F_{l_γ}. Cette dernière contrôle la vitesse de convergence pour $k > 2$.

F.3.4 Limites de l'apprentissage avec des fonctions (ϵ, γ, τ)-bonnes

Initialement, le but de ce chapitre était de fournir des fondations théoriques pour un potentiel algorithme d'adaptation de domaine basé sur les similarités $(\epsilon, \gamma, \tau)-$bonnes. A cet égard, nous notons que l'analyse établie suggère l'utilisation des mêmes points de référence pour les deux domaines puisqu'elle résulte en des bornes plus serrées. D'autre part, elle dépend crucialement de l'apprentissage d'une fonction de similarité, ce qui peut être fait en suivant l'approche de Bellet et al. (2012). Nous avons choisi de nous concentrer sur cette approche en particulier car nous avons été attirés par ses propriétés comme la parcimonie du classifieur résultant et l'extension non linéaire avec la KPCA (Schölkopf et al., 1997). Dans cette section, nous faisons une étude rétrospective sur l'intérêt potentiel d'apprendre une fonction d e similarité bilinéaire *e.g.* $-$bonne, *i.e.* une fonction de similarité dans l'ensemble suivant:

$$\{K_{\mathbf{A}} : (\mathbf{x},\mathbf{x}') \mapsto \mathbf{x}^T\mathbf{A}\mathbf{x}'; \mathbf{A} \in \mathbb{R}^{n\times n}; \|\mathbf{A}\|_2 \leq 1\}.$$

Nous montrons alors que chercher une fonction de similarité $(\epsilon, \gamma)-$bonne paramétré par une matrice $\mathbf{A}$ avec $\|\mathbf{A}\|_2 \leq 1$ est exactement équivalent à chercher un classifieur linéaire $\mathbf{w}$ vérifiant $\|\mathbf{w}\| \leq \|\mu'\|$, indépendamment de la fonction de coût considérée et de la distribution de points de référence $\mathcal{L}$. Ce résultat inclut Bellet et al. (2012) dans lequel les auteurs observent que résoudre leur problème avec différents choix de points de référence ne change pas la performance du classifieur obtenu, ce qui est en accord avec ce que nous venons de dire. De plus, trouver une fonction de similarité restreinte en norme de Frobenius en pratique revient à résoudre un problème d'optimisation régularisé strictement convexe (vu la convexité stricte de la norme de Frobenius). Dans ce cas, la matrice solution du problème a un rang 1 et peut être facilement déduite en résolvant un problème où l'on cherche un classifieur, ce qui réduit le nombre de variables à chercher de d^2 à d. Finalement, avec une matrice de rang 1, l'espace de similarité induit sera de dimension 1, et y chercher un classifieur linéaire, en dépit de sa justification théorique (Balcan et al., 2008a), résulte en un problème d'optimisation mal conditionné.

F.4 Adaptation de domaine tenant compte la marge

Nous proposons une nouvelle analyse théorique de l'adaptation de domaine non supervisée (DA) qui regroupe les notions de vaste marge de classification et d'apprentissage adversarial. Cette analyse généralise des travaux précédents sur le DA en fournissant une borne sur le taux violation de marge sur le domaine cible, ce qui reflète mieux la qualité de séparation entre les classes dans le domaine cible que simplement utiliser le taux de classification erronée. La borne met en exergue le bénéfice d'avoir une vaste de marge de

séparation sur le domaine source pour l'adaptation, et introduit une distance entre domaines basée sur le transport optimal (OT), qui a le mérite d'être dépendante de la tâche en question. À partir des résultats théoriques établis, nous dérivons une solution algorithmique pour l'adaptation de domaine qui introduit une nouvelle approche d'apprentissage adversariale en apprentissage non profond, basée sur le DA, et surpasse d'autres approches basées sur l'OT sur plusieurs tâches de classification avec des données simulées et réelles.

Dans la suite, on se servira de la fonction de perte suivante:

$$l_{\rho,\beta}(y,y') := \begin{cases} 1 - \frac{(y\cdot y' - \rho)}{\beta}, & \text{if } \beta > 0 \text{ and } \rho \leq y\cdot y' \leq \beta + \rho \\ [y\cdot y' < \rho], & \text{otherwise.} \end{cases} \tag{F.38}$$

qui est un perte de rampe (avec une pente $\frac{1}{\beta}$), translatée de $\rho > 0$.

F.4.1 Bornes portant sur la marge de classification dans le domaine cible

F.4.1.1 Une première borne sur le risque hinge

Cette section porte sur la première borne que nous avons établi dans Dhouib et al. (2019), qui est notre première tentative d'établir une garantie théorique en adaptation de domaine en affaiblissant les hypothèses utilisées précédemment: dorénavant, on suppose que ni la "covariate shift", ni la dominance de la cible par la source ne sont vérifiées. Initialement, notre résultat porte sur un classifieur défini par une fonction de similarité $(\epsilon,\gamma,\tau)-$bonne, mais ici nous le généralisons à tout classifieur ayant des valeurs dans $[-1,1]$, dans la proposition suivante.

Proposition F.4.1. *Soit l_ρ la fonction de perte définie par $l(y,y') := \left(1 - \frac{y\cdot y'}{\rho}\right)_+$. Alors, pour tout $h \in \mathbb{H}$, et tout $0 < \rho \leq 1$, on a:*

$$\mathfrak{E}^\rho_{\mathcal{T}}(h) \leq \mathfrak{E}^\rho_{\mathcal{S}}(h) + \frac{1}{\rho}\tilde{\Delta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}}) + \frac{1}{\rho}\tilde{\lambda}(\mathcal{S},\mathcal{T}), \tag{F.39}$$

où

$$\tilde{\Delta}_{h,\mathbb{H}'}(\mathcal{S}_{\mathbb{X}},\mathcal{T}_{\mathbb{X}}) := \sup_{h'\in\mathbb{H}'} \mathop{\mathbb{E}}_{\mathbf{x}_t\sim\mathcal{T}_{\mathbb{X}}} \left[\left| h(\mathbf{x}_t)h'(\mathbf{x}_t) - \mathop{\mathbb{E}}_{\mathbf{x}_s\sim\mathcal{S}_{\mathbb{X}}} \left[h(\mathbf{x}_s)h'(\mathbf{x}_s)\right]\right|\right], \tag{F.40}$$

$$\tilde{\lambda}(\mathcal{S},\mathcal{T}) := \inf_{f\in\mathbb{H}'} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left[|y - f(\mathbf{x})|\right] + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \left[|y - f(\mathbf{x})|\right]. \tag{F.41}$$

Cette borne comprend trois termes, et donc elle est similaire en forme aux résultats classiques en DA. Cependant, elle porte sur une version mise à l'échelle de la perte hinge (par le biais de $\rho > 0$), et donc elle reflète mieux la qualité de séparation entre classes sur le domaine cible, contrairement aux bornes de Ben-David et al. (2010); Zhang et al. (2019) qui ne concernent que le taux de classification erronée. La partie estimable de notre borne est convexe en h, lui conférant un avantage en cas de dérivation d'un algorithme de DA. Néanmoins, ce résultat a quelques limitations majeures. Premièrement, le terme d'alignement ne disparaît pas pour $\mathcal{S}_{\mathbb{X}} = \mathcal{T}_{\mathbb{X}}$. Deuxièmement, dans ce même terme, prendre l'espérance sur le domaine source établit une correspondance entre tous les points cible et un seul point source (correspondant à la moyenne), ce qui n'est pas la meilleure correspondance. Finalement, la valeur absolue dans le terme non estimable n'est pas convenable pour comparer une fonction de score et un classifieur à valeurs dans $\{-1,1\}$.

F.4.1.2 Borner le taux de violation de marge dans le domaine cible

Nous passons au résumé des contributions de Dhouib et al. (2020b), où nous adressons les limites de la borne précédente tout en gardant ses avantages, et nous la convexifions en se basant sur le transport optimal.

Une borne avec une divergence non-convexe entre distributions Dans le théorème suivant, nous énonçons notre borne portant sur le taux de violation de marge sur le domaine cible.

Théorème F.4.1. *Supposons que pour tout $h' \in \mathbb{H}'$, nous avons $\mathbb{P}_{\mathbf{x}\sim\mathcal{S}_\mathbb{X}}[h'(\mathbf{x}) = 0] = \mathbb{P}_{\mathbf{x}\sim\mathcal{T}_\mathbb{X}}[h'(\mathbf{x}) = 0] = 0$. Soit $\rho, \beta, \alpha > 0$ tel que $\rho + \beta < \alpha < 1$ et soit $\rho' := \frac{\rho+\beta}{\alpha}$. Alors tout $h \in \mathbb{H}$ vérifie la borne suivante:*

$$\mathfrak{E}^{\rho,0}_{\mathcal{T}}(h) \leq \mathfrak{E}^{\rho',0}_{\mathcal{S}}(h) + d^{\rho,\beta}_{h,\mathbb{H}'}(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X}) + \lambda_\alpha,$$

où

$$d^{\rho,\beta}_{h,\mathbb{H}'}(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X}) := \sup_{h'\in\mathbb{H}'} \left| \mathfrak{E}^{\rho,\beta}_{\mathcal{S}}(h, h') - \mathfrak{E}^{\rho,\beta}_{\mathcal{T}}(h, h') \right|$$

and

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

Ce théorème concerne le taux de violation de la marge de classification, et donc tout comme la proposition précédente, il reflète la séparation entre classes mieux que le taux de classification erronée. Le terme d'alignement devient nul pour des distributions de marginales égales, et il a la forme d'une métrique intégrale de probabilité (IPM) (Zolotarev, 1984). De plus, contrairement à la borne de Zhang et al. (2019, Proposition 3.3), notre terme est continu en h et ne fait pas intervenir le classifieur correspondant à son signe, ce qui le rend mieux adapté aux algorithmes d'optimisation. Troisièmement, notre terme non estimable a la particularité d'être non symétrique en les deux domaines, contrairement aux termes de Ben-David et al. (2010); Zhang et al. (2019), et il indique qu'il est avantageux d'avoir un taux de violation de marge absolue faible sur le domaine source. Notons qu'à partir du théorème précédent, il est immédiat de déduire le résultat de Ben-David et al. (2010, Theorem 2).

Une divergence convexe entre domaines en utilisant le transport optimal Le théorème précédent fait apparaître la fonction de violation de marge $[. < \rho]$ dont l'optimisation est NP-difficile (Arora et al., 1997). Afin de surmonter un telle difficulté, nous utilisons une fonction de perte de remplacement. Quant au terme de divergence, il est non convexe bien que continu. Nous le convexifions en faisant appel au transport optimal dans la proposition suivante.

Proposition F.4.2. *Pour tout $\rho, \beta > 0$, on a*

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

où

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

et $\Pi(\mathcal{S}_\mathbb{X}, \mathcal{T}_\mathbb{X})$ est l'ensemble de plans de transport entre $\mathcal{S}_\mathbb{X}$ et $\mathcal{T}_\mathbb{X}$.

Pour simplifier les notations, nous désignons par Π l'ensemble de plans de transport entre marginales. Dans notre nouvelle divergence, le terme $\Delta_{\mathbb{H}'}(h, \mathcal{P})$ est convexe en h pour un couplage $\mathcal{P}$ fixé, et vice-versa. Elle a la forme d'une version robuste de la distance de Wasserstein entre les distributions unidimensionnelles $hh'\#\mathcal{S}_{\mathbb{X}}$ et $hh'\#\mathcal{T}_{\mathbb{X}}$, et admet l'interprétation adversariale suivante: le classifieur h' est un adversaire visant à séparer les domaines, alors que le couplage $\mathcal{P}$ réalisant le minimum résiste à cette séparation.

En combinant nos deux derniers résultats, on aboutit à la proposition suivante.

Proposition F.4.3. *Avec les suppositions des deux derniers résultats, supposons de plus que l est une fonction de perte définie par $l(h(\mathbf{x}), y) := \ell(y \cdot h(\mathbf{x}))$, où ℓ est croissante et vérifie $\ell(\rho') \neq 0$. Alors, pour tout $h \in \mathbb{H}$:*

$$\mathfrak{E}^{\rho,0}_{\mathcal{T}}(h) \leq \frac{1}{\ell(\rho')}\mathfrak{E}^{l}_{\mathcal{S}}(h) + \frac{1}{\beta}\inf_{\mathcal{P}\in\Pi}\Delta_{\mathbb{H}'}(h, \mathcal{P}) + \lambda_{\alpha}. \tag{F.42}$$

Cette borne est convexe en h, et donc sert mieux pour dériver un algorithme de DA. De plus, le transport optimal fait mieux correspondre les exemples des deux domaines, en contraste avec le résultat de notre contribution Dhouib et al. (2019), mentionné plus haut.

Comparée à d'autres bornes en DA incluant la distance de Wasserstein (Redko et al., 2017; Courty et al., 2017; Shen et al., 2018), notre terme de divergence prend en compte les espaces d'hypothèses considérés, ce qui fait de lui une pseudo-métrique moins stricte que la distance de Wasserstein. Afin de soutenir cette affirmation, nous bornons notre terme basé sur le transport optimal dans la proposition suivante.

Proposition F.4.4. *Soit $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ une métrique, et supposons que toutes les hypothèses de $\mathbb{H}$ and $\mathbb{H}'$ sont $L-$Lipschitz par rapport à la métrique d pour un certain $L > 0$. Alors l'on a*

$$\sup_{h\in\mathbb{H}}\left(\inf_{\mathcal{P}\in\Pi}\Delta_{\mathbb{H}'}(h, \mathcal{P})\right) \leq \inf_{\mathcal{P}\in\Pi}\sup_{\substack{h\in\mathbb{H}\\ h'\in\mathbb{H}'}}\mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}}\left[\left|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)\right|\right] \leq 2LW_1(\mathcal{S}_{\mathbb{X}}, \mathcal{T}_{\mathbb{X}}),$$

où

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

est la distance de Wasserstein associée à la métrique d.

F.4.2 Algorithme d'adaptation de domaine

Nous rappelons que l'objectif d'un algorithme de DA est de trouver une hypothèse avec un taux de classification erronée faible sur le domaine cible, et que trouver une avec un taux de violation de marge faible est une exigence plus forte. Nous visons à satisfaire cette dernière en minimisant la partie estimable de notre borne, une approche que l'on retrouve dans Germain et al. (2013, 2016a); Courty et al. (2017) par exemple.

F.4.2.1 Minimisation de la partie estimable de la borne

Minimiser la partie estimable de notre borne revient à résoudre le problème suivant

$$\min_{\substack{h\in\mathbb{H}\\ \mathcal{P}\in\Pi}}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\left[l(h(\mathbf{x}), y)\right] + \frac{\ell(\rho')}{\beta}\Delta_{\mathbb{H}'}(h, \mathcal{P}). \tag{F.43}$$

Bien que ce problème soit convexe, notre terme d'alignement contient un supremum sur $\mathbb{H}'$ qui est un ensemble potentiellement infiniment et pour lequel le calcul du supremum n'est pas facile en général. Dans ce qui suit, nous nous focalisons sur un choix particulier de $\mathbb{H}$ et $\mathbb{H}'$ qui facilite l'optimisation du problème précédent.

F.4.2.2 Application à la classification linéaire

Nous nous intéressons à la classification linéaire, et dans ce cas la fonction objectif précédente devient plus explicite comme l'énonce la proposition suivante.

Proposition F.4.5. *Soient* $\mathbb{H}$ *et* $\mathbb{H}'$ *les espaces de classifieurs linéaires bornés en norme Euclidienne et* $1-$*norme, respectivement. Alors, le problème* (4.27) *a l'expression suivante:*

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^n \\ \mathcal{P}\in\Pi}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left[l(\mathbf{w}^T\mathbf{x}, y)\right] + \delta \left\| \mathop{\mathbb{E}}_{(\mathbf{x}_s,\mathbf{x}_t)\sim\mathcal{P}} \left[\left|(\mathbf{x}_s\mathbf{x}_s^T - \mathbf{x}_t\mathbf{x}_t^T)\mathbf{w}\right|\right] \right\|_\infty + \zeta\|\mathbf{w}\|_2^2, \tag{F.44}$$

où l *est une fonction de perte convexe et* $\delta, \zeta > 0$ *sont deux hyperparamètres liés aux bornes sur* $\mathbb{H}$ *et* $\mathbb{H}'$.

Ce résultat introduit un problème d'optimisation fortement convexe grâce à la régularisation en norme euclidienne carrée, et le supremum y admet une forme explicite donnée par la $\infty-$norme. De plus, δ contrôle l'importance donnée au terme d'alignement: pour $delta = 0$, on résout la classification supervisée sur domaine source en négligeant le domaine cible, et pour $\delta \to \infty$, la fonction est minimisée pour $\mathbf{w} = 0$, ce qui aligne bien les domaines mais d'une manière triviale mettant à côté toute information utile sur les classes.

F.4.2.3 Procédure d'optimisation pour le cas discret

Afin de résoudre notre problème dans le cas discret, nous suivons la procédure de Courty et al. (2017), consistant à alterner les minimisations sur $\mathbf{P}$ la matrice de transport (version discrète du couplage $\mathcal{P}$) et sur $\mathbf{w} \in \mathbb{R}^p$, un cas particulier de la descente de coordonnées (Grippo and Sciandrone, 2000). Pour minimiser sur $\mathbf{P}$, nous utilisons l'algorithme de Blankenship and Falk (1976, Algorithme 2.2) et pour $\mathbf{w}$, nous mettons en œuvre la méthode L-BFGS de la famille quasi-Newton.

F.4.2.4 Apprendre dans l'espace induit par une similarité

Notre approche minimisant la partie estimable de notre borne convexe sur le risque cible est basée sur l'existence d'une hypothèse minimisant le terme non estimable. En gros, il suffit d'avoir une hypothèse qui a une bonne performance de classification sur l'union des deux domaines. Si ce n'est pas le cas, alors une transformation des données pourrait servir à vérifier cette hypothèse. Parmi de telles transformations, on choisit celles basées sur les fonctions de similarité $(\epsilon, \gamma, \tau)-$bonnes: en utilisant la transformation $\phi^L(\mathbf{x}) = (K(\mathbf{x}, \tilde{\mathbf{x}}_1), ..., K(\mathbf{x}, \tilde{\mathbf{x}}_{n'}))$, la théorie de Balcan et al. (2008a) garantit l'existence d'un classifieur séparant les classes pourvu que K est une fonction de similarité convenable. Ainsi, pour un problème de DA fixé, nous fixons un noyau (qui est une fonction de similarité) et nous projetons les données dans son espace induit à l'aide de ϕ^L.

F.4.3 Évaluation empirique

Nous évaluons notre approche dans le cas de la classification linéaire sur un jeu de données simulées et un jeu de données réelles. Nous désignons par **MADAOT** notre algorithme (acronyme venant des initiales du titre de notre contribution Dhouib et al. (2020b)). Notons que pour sélectionner les hyperparamètres, nous utilisons un ensemble étiqueté de données cible qui ne sert pas à tester l'algorithme, et bien évidemment les étiquettes ne sont pas utilisées dans la procédure d'entraînement.

F.4.3.1 Données simulées des croissants entrelacés

Il s'agit d'un jeu de données très utilisé pour évaluer la performance d'algorithmes de DA (Courty et al., 2016, 2017), grâce à la possibilité d'augmenter la difficulté de l'adaptation. En effet, le jeu de données est bi-dimensionnel et le domaine cible est obtenu par une rotation des exemples du domaine source: plus l'angle de rotation est élevé, et plus l'adaptation est difficile. D'une manière similaire à Courty et al. (2016), nous adressons la non-linéarité en utilisant un noyau Gaussien comme fonction de similarité, avec un paramètre d'échelle égal à la moyenne de la distance euclidienne entre les exemples du domaines source (Kar and Jain, 2011). Nous comparons notre algorithme à un SVM supervisé (sans adaptation), OT-GL(Courty et al., 2016) (transport optimal avec régularisation de group-Lasso liée aux classes) et JDOT (Courty et al., 2017), dans le tableau 4.1. Nous remarquons en l'efficacité de notre méthode comparée aux autres pour les angles à partir de 40°, correspondant à des problèmes d'adaptation difficiles. De plus, afin d'illustrer le l'hyperparamètre δ, nous avons représenté son influence sur la justesse de classification dans le domaine cible sur la Figure 4.3.

F.4.3.2 Données d'analyse de sentiments

Nous considérons le jeu de données de revues de produits Amazon (Blitzer et al., 2007), correspondant à une tâche d'analyse de sentiments. Les revues correspondent à 4 catégories de produits: DVD, livres, électroniques et cuisine, constituant 4 domaines et donc 12 tâches possibles d'adaptation de domaine avec une seule source. Nous suivons le prétraitement de Chen et al. (2011a) afin de réduire le nombre de caractéristiques de 100000 à 40000, et nous fixons le produit scalaire canonique comme fonction de similarité. Nous comparons notre méthode à un SVM supervisé, OT-GL et JDOT (méthodes utilisées pour le jeu de données simulées), mais aussi à DANN (Ganin et al., 2016) et JDOT avec un réseau de neurones comme classifieur, dans les tableaux 4.2 et 4.3. On y constate que notre méthode surpasse les autres pour 8 tâches sur 12, en dépit du pouvoir d'extraction de caractéristiques des méthodes basées sur l'apprentissage profond. Nous estimons que ce succès est dû à l'attachement de notre terme d'alignement aux espaces d'hypothèses considérés, ce qui n'est pas le cas de la distance de Wasserstein classique associée à la distance euclidienne, dont la vitesse de convergence vers sa version théorique évolue exponentiellement en la dimension de l'espace ambiant.

F.5 Transport Optimal Minimax

Le problème de transport optimal (OT) et sa distance de Wasserstein associée sont récemment devenus un sujet de grand intérêt dans la communauté de l'apprentissage automatique pour la comparaison de distributions de probabilités. Cependant, calculer le OT en pratique nécessite le choix de la métrique terrain qui reflète au mieux la connaissance de l'utilisateur sur le problème en question. Dans ce chapitre, inspirés par le problème de transport optimal adversarial dérivé dans le chapitre précédent, nous proposons une formulation générale d'un problème minimax OT qui optimise conjointement la métrique terrain et le plan de transport, nous permettant de définir une distance robuste entre distributions. Nous analysons la formulation proposée théoriquement dans plusieurs cas d'intérêt pratique, nous montrons ses avantages comparée à des travaux précédemment proposés, et nous dérivons des algorithmes pour chaque cas considéré. En plus, nous utilisons cette méthode pour définir une notion de stabilité, ce qui nous permet de sélectionner une métrique terrain robuste à des perturbations bornées. Finalement, nous menons une étude expérimentale soulignant l'intérêt de notre approche.

F.5.1 Préliminaires

F.5.1.1 Transport optimal

Nous avons déjà présenté le transport optimal dans la deuxième section de ce chapitre, dédiée à l'adaptation de domaine. Cette présentation là concerne principalement le cas où le coût de transport est une métrique, alors que le problème de transport de Monge-Kantorovitch (Monge, 1781; Kantorovich, 1942) est plus général: étant donné une fonction de coût $c : \mathbb{X} \times \mathbb{X}' \to \mathbb{R}_+$, avec $\mathbb{X}$ et $\mathbb{X}'$ des espaces euclidiens dans la plupart des cas. Pour deux distributions $\mathcal{D}$ et $\mathcal{D}'$ (respectivement définies sur $\mathbb{X}$ et $\mathbb{X}'$), la formulation de l'OT par Kantorovich (1942) est la suivante:

$$W_c(\mathcal{D}, \mathcal{D}') = \inf_{\mathcal{P} \in \Pi(\mathcal{D}, \mathcal{D}')} \mathop{\mathbb{E}}_{(\mathbf{x}, \mathbf{x}') \sim \mathcal{P}} \left[c(\mathbf{x}, \mathbf{x}') \right]. \tag{F.45}$$

La version discrète de ce problème est le transport entre deux mesures de probabilités discrètes, à supports correspondant respectivement à deux ensembles de vecteurs $\{\mathbf{x}_i\}_{i=1}^m$, $\{\mathbf{x}'_j\}_{j=1}^{m'}$, et données par $\mathbf{r} \in \Delta_m$ and $\mathbf{c} \in \Delta_{m'}$. Contrairement à la formulation discrète présentée dans la section dédiée au DA, $\mathbf{r}$ et $\mathbf{c}$ peuvent bien être des poids non uniformes. Pour simplifier les notations, $\Pi(\mathcal{D}, \mathcal{D}')$ et $\Pi(\mathbf{r}, \mathbf{c})$ seront dénotés Π, et la distinction est à faire selon le contexte.

F.5.1.2 Transport optimal minimax

Deux autres études ont considéré une formulation minimax pour le transport optimale. Paty and Cuturi (2019) se sont intéressés au transport robuste aux projections dans des sous-espaces d'une dimension donnée k. Leur formulation peut s'écrire comme suit:

$$\mathcal{S}_k^2(\mathcal{D}, \mathcal{D}') := \min_{\mathcal{P} \in \Pi} \max_{\substack{\mathbf{0} \preccurlyeq \mathbf{M} \preccurlyeq \mathbf{I_n} \\ \mathrm{Tr}\{\mathbf{M}\} = k}} \langle \mathbf{V}_\mathcal{P}, \mathbf{M} \rangle = \min_{\mathcal{P} \in \Pi} \sum_{i=1}^{k} \lambda_i(\mathbf{V}_\mathcal{P}), \tag{F.46}$$

$$\text{avec } \mathbf{V}_\mathcal{P} := \mathop{\mathbb{E}}_{(\mathbf{x}, \mathbf{x}') \sim \mathcal{P}} \left[(\mathbf{x} - \mathbf{x}')(\mathbf{x} - \mathbf{x}')^T \right], \tag{F.47}$$

où $\preccurlyeq$ dénote l'ordre de Loewener pour les matrices semi-définies positives (PSD) et $\{\lambda_i\}_{i=1}^k$. Leur formulation permet d'adresser le fléau de la dimension dont souffre le transport optimale.

Dans Alvarez-Melis et al. (2018), les auteurs considère une fonction de coût sous-modulaire appartenant à un ensemble F. Ce dernier admet un polytope de base $\mathfrak{B}_F$ qui leur sert à définir leur problème:

$$\mathrm{StrOT}(\mathcal{D}, \mathcal{D}') := \min_{\mathbf{P} \in \Pi} \max_{\mathbf{C} \in \mathfrak{B}_F} \langle \mathbf{P}, \mathbf{C} \rangle,$$

F.5.2 Transport robuste avec un ensemble convexe de matrices de coût

F.5.2.1 Formulation du problème

Soit $\mathfrak{G}$ un ensemble de fonctions de coût sur $\mathbb{X} \times \mathbb{X}'$. Nous n'imposons aucune conditions sur ces fonctions, hormis qu'elles garantissent l'existence d'un plan de transport. Notre problème d'intérêt est le suivant:

$$\mathrm{RKP}(\Pi, \mathfrak{G}) = \min_{\mathcal{P} \in \Pi} \max_{c \in \mathfrak{G}} \mathop{\mathbb{E}}_{(\mathbf{x}, \mathbf{x}') \sim \mathcal{P}} \left[c(\mathbf{x}, \mathbf{x}') \right], \tag{F.48}$$

où l'on cherche une couplage $\mathcal{P}^*$ robuste au choix de la fonction de coût $c \in \mathfrak{G}$, en considérant le pire coût de transport atteignable. Nous dénotons la valeur de ce problème par $\mathrm{RKP}(\Pi, \mathfrak{G})$ (RKP pour *Robust Kantorovich Problem*). Par abus de notation, nous utilisons $\mathrm{RKP}(\mathfrak{P}, \mathfrak{G})$ pour tout ensemble $\mathfrak{P} \subseteq \Pi$ (même non convexe) pour désigner $\mathrm{RKP}(\mathrm{Conv}(\mathfrak{P}), \mathfrak{G})$, où $\mathrm{Conv}(.)$ dénote l'enveloppe convexe. De plus, nos étendons la notation W_c à $W_\mathfrak{G} := RKP(\Pi, \mathfrak{G})$.

F.5.2.2 Choix de $\mathfrak{G}$

Ensemble infini de distances de Mahalanobis Pour toute matrice $\mathbf{M} \in \mathbb{R}^{n\times n}$, nous rappelons que sa $p-$norme de Schatten est donnée par

$$\|\mathbf{M}\|_p^p = \sum_{1\leq i\leq n} \sigma_i^p(\mathbf{M}),$$

où $p \in [1, +\infty]$ et $\{\sigma_i(\mathbf{M})\}$ sont les valeurs singulières $\mathbf{M}$.

Nous définissons l'ensemble $\mathfrak{G}$ comme suit:

$$\mathfrak{G} = \{c^{\mathbf{M}} : (\mathbf{x}, \mathbf{x}') \mapsto (\mathbf{x} - \mathbf{x}')^T \mathbf{M}(\mathbf{x} - \mathbf{x}'); \|\mathbf{M}\|_p \leq 1\}. \tag{F.49}$$

Nous avons alors la proposition suivante

Proposition F.5.1. *Soit $\mathfrak{G}$ défini par* (F.49) *pour des matrices $\mathbf{M} \in \mathbb{S}_+^{n\times n}$. Alors, $\mathfrak{G}$ est un ensemble convexe compact de fonctions de coût, et pour tous $p, q \in [1, +\infty]$ tels que $\frac{1}{p} + \frac{1}{q} = 1$, on a les résultats suivant:*

1. $\mathrm{RKP}(\Pi, \mathfrak{G}) = \min_{\mathcal{P}\in\Pi} \|\mathbf{V}_{\mathcal{P}}\|_q$. *En particulier, on a:*

$$\mathrm{RKP}(\Pi, \mathfrak{G}) = \begin{cases} W_2^2(\mathcal{D}, \mathcal{D}'), & \text{if } q = 1, \\ \mathcal{S}_1^2(\mathcal{D}, \mathcal{D}'), & \text{if } q = \infty. \end{cases}$$

2. *Pour tout $\mathcal{P} \in \Pi$, $\|\mathbf{M}^*\|_p = 1$ et*

$$\mathbf{M}^* = \underset{\substack{\mathbf{M}\in\mathbb{S}_+^{n\times n} \\ \|\mathbf{M}\|_p\leq 1}}{\arg\max} \langle \mathbf{V}_{\mathcal{P}}, \mathbf{M}\rangle = \left(\frac{\mathbf{V}_{\mathcal{P}}}{\|\mathbf{V}_{\mathcal{P}}\|_q}\right)^{\frac{q}{p}}.$$

En particulier, pour $p = 2$, l'on n'a pas besoin d'imposer la condition PSD à $\mathbf{M}$, i.e.

$$\mathbf{M}^* = \underset{\|\mathbf{M}\|_2\leq 1}{\arg\max} \langle \mathbf{V}_{\mathcal{P}}, \mathbf{M}\rangle = \frac{\mathbf{V}_{\mathcal{P}}}{\|\mathbf{V}_{\mathcal{P}}\|_2}.$$

Ce théorème interpole entre SRW (Paty and Cuturi, 2019) pour $k = 1$, et le problème de Kantorovitch qui peut être vu comme un problème min-max. La Figure 5.1 illustre cette remarque pour plusieurs choix de q. En outre, l'expression de $\mathbf{M}^*$ montre qu'elle est proportionnelle à $\mathbf{V}_{\mathcal{P}}$,et si cette dernière capte les directions de déplacement en dimensions inférieurs, alors ce sera le cas pour $\mathbf{M}^*$. Le dernier point de la proposition montre l'avantage du cas $p = 2$ en pratique en raison de la non-nécessité d'imposer la condition PSD. Pour conclure, nous présentons le corollaire suivant.

Corollaire F.5.1. *Avec les hypothèses de Proposition F.5.1, on a l'inégalité suivante pour tout $p \in [1, +\infty]$:*

$$\frac{1}{d^{\frac{1}{p}}} W_2^2(\mathcal{D}, \mathcal{D}') \leq W_{\mathfrak{G}}(\mathcal{D}, \mathcal{D}') \leq W_2^2(\mathcal{D}, \mathcal{D}').$$

Ce résultat est comparable à celui fourni par Paty and Cuturi (2019, Proposition 2) pour la SRW, mais il ne fait pas intervenir k puisque nous n'imposons pas de contrainte sur le range de $\mathbf{M}$.

Ensemble fini de fonctions de coût Soit $\{c_1, ..., c_K\}$ une famille de fonctions coût candidates, et soit $\mathfrak{G} = \text{Conv}(\{c_1, ..., c_K\})$, ce qui implique que $\mathfrak{G}$ est un ensemble convexe compact puisqu'il s'agit de la combinaison convexe d'un nombre fini d'éléments, *i.e.* un polytope. C'est le cas de StrOT Alvarez-Melis et al. (2018) où $\mathfrak{G} = \mathfrak{B}_F$, avec $\mathfrak{B}_F$ le polytope de base d'un ensemble de fonctions sous-modulaires. Un autre cas où $\mathfrak{G}$ est fini est dans Equation (F.44): le membre de droite est donnée par le problème suivant:

$$\min_{\mathcal{P}\in\Pi} \left\| \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{P}} \left[\left| (\mathbf{x}\mathbf{x}^T - \mathbf{x}'\mathbf{x}'^T)\mathbf{w} \right| \right] \right\|_\infty , \tag{F.50}$$

qui est un cas particulier du RKP pour l'ensemble de fonctions de coût suivant:

$$\mathfrak{G} := \text{Conv}(\{c_1, \cdots, c_n\}) \quad \text{avec} \quad c_k : (\mathbf{x}, \mathbf{x}') \mapsto \left| \mathbf{e}_k^T (\mathbf{x}\mathbf{x}^T - \mathbf{x}'\mathbf{x}'^T)\mathbf{w} \right|, \tag{F.51}$$

où $\mathbf{e}_k$ dénote le $k-$ième vecteur de la base canonique de $\mathbb{R}^n$.

F.5.2.3 Stratégie d'optimisation proposée

Nous considérons le cas discret défini par deux ensembles de vecteurs $\{\mathbf{x}_i\}_{i=1}^m$ et $\{\mathbf{x}'_j\}_{j=1}^{m'}$ avec des distributions empiriques données par $\mathbf{r}$ et $\mathbf{c}$. $\mathfrak{G}$ correspond dans ce cas à un ensemble convexe compact de matrices de coût.

Nous adaptons la méthode des ensembles sécants (*cutting set method*) présentée dans Mutapcic and Boyd (2009) à notre problème. Cette méthode consiste essentiellement en l'alternation entre deux étapes: résoudre un problème de pire cas et résoudre un autre avec un nombre de contraintes évoluant linéairement avec les itérations. Pour l'appliquer à notre problème, nous le considérons comme un problème max-min, ce qui est justifiable par le théorème minimax de Von-Neumann vu la compacité et la convexité de Π et $\mathfrak{G}$. Cependant, résoudre le problème max-min sans se soucier du point selle fournit uniquement la solution $\mathbf{C}^*$ qui maximise la fonction $\min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, . \rangle$. Nous modifions alors l'algorithme pour aussi obtenir la solution $\mathbf{P}^*$ minimisant la fonction $\max \mathbf{C} \in \mathfrak{G} \langle ., \mathbf{C} \rangle$. Bref, nous obtenons le point selle de notre problème RKP. La manière exacte de l'obtenir est décrite dans la proposition suivante:

Proposition F.5.2. *Soit $\mathfrak{P}$ un sous-ensemble fini de Π. Alors on a:*

1. $\text{RKP}(\mathfrak{P}, \mathfrak{G}) := \text{RKP}(\text{Conv}(\mathfrak{P}), \mathfrak{G})$ *a un point selle* $(\mathbf{P}^*, \mathbf{C}^*)$ *vérifiant:*

$$\langle \mathbf{P}^*, \mathbf{C}^* \rangle = \min_{\mathbf{P}\in\text{Conv}(\mathfrak{P})} \max_{\mathbf{C}\in\mathfrak{G}} \langle \mathbf{P}, \mathbf{C} \rangle = \max_{\mathbf{C}\in\mathfrak{G}} \min_{\mathbf{P}\in\mathfrak{P}} \langle \mathbf{P}, \mathbf{C} \rangle . \tag{F.52}$$

2. $\text{RKP}(\mathfrak{P}, \mathfrak{G})$ *est équivalent à*

$$\begin{aligned} \mathbf{C}^* \in \mathop{\arg\max}_{\mathbf{C}\in\mathfrak{G}, \mu\geq 0} \ & \mu \\ s.t. \quad \langle \mathbf{P}, \mathbf{C} \rangle \geq \ & \mu, \ \forall \mathbf{P} \in \mathfrak{P}. \end{aligned} \tag{F.53}$$

3. $\mathbf{P}^* = \sum_{l=1}^{|\mathfrak{P}|} q_l \mathbf{P}_l$, *où* $\mathfrak{Q} = \{q_l\}_{l=1}^{|\mathfrak{P}|}, \sum_i q_i = 1$, *sont les variables duales de* (F.53).

Ce résultat est valable pour tout ensemble $\mathfrak{P} \subset \Pi$, et donc il s'applique $\text{RKP}(\Pi, \mathfrak{G})$ lorsque $\mathfrak{P}$ est l'ensemble des sommets du polytope Π. De plus, il reste valable même en cas de régularisation du transport (*e.g.* régularisation entropique (Cuturi, 2013)) puisqu'il suffit de considérer $\text{RKP}(\tilde{\Pi}, \mathfrak{G})$, où $\tilde{\Pi}$ est un sous-ensemble convexe compact de Π. En effet, le problème régularisé a la forme suivante

$$\min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C} \rangle + \lambda R(\mathbf{P}) \tag{F.54}$$

où R est une fonction convexe, et il peut s'écrire comme problème de transport où $\mathbf{P}$ est contraint à rester dans un sous-ensemble de Π en invoquant les conditions de Karush-Kuhn-Tucker (Karush, 1939; Kuhn and Tucker, 1951).

Notre algorithme est donné par Algorithm 1: à chaque étape, on résout un problème auxiliaire avec un sous ensemble $\mathfrak{P} \subseteq \Pi$ qui croit (pour l'inclusion) jusqu'à ce que l'écart de dualité devient inférieur à un seuil donné. Cet algorithme incorpore la stratégie de réduction du nombre de contraintes décrite dans Mutapcic and Boyd (2009, Section 5.3.2). Cet algorithme est générique et peut être utilisé pour SRW (Paty and Cuturi, 2019) et StrOT (Alvarez-Melis et al., 2018), et bénéficie de la garantie théorique suivante sur le nombre d'itérations.

Proposition F.5.3. *Soit T le nombre d'itérations requises par Algorithm 1 pour atteindre une erreur $\varepsilon_t \leq \tau_1$. Alors*

$$T \leq \left(\frac{\operatorname{diam}_\infty(\mathfrak{G}) + \operatorname{RKP}(\mathfrak{P}_0, \mathfrak{G})}{2\tau_1} + 1 \right)^{\dim(\mathfrak{G})+1}$$

where

$$\operatorname{diam}_\infty(\mathfrak{G}) := \sup_{\substack{\mathbf{C}^1, \mathbf{C}^2 \in \mathfrak{G} \\ i,j}} \left| (\mathbf{C}^1)_{ij} - (\mathbf{C}^2)_{ij} \right|,$$

et $\dim(\mathfrak{G})$ *est la dimension de l'enveloppe affine de* $\mathfrak{G}$*. De plus,* $\forall t \geq 0$*, on a*

$$0 \leq \operatorname{RKP}(\mathfrak{P}_t, \mathfrak{G}) - \operatorname{RKP}(\Pi, \mathfrak{G}) \leq \varepsilon_t.$$

Ce résultat introduit $\operatorname{diam}_\infty(\mathfrak{G})$ pouvant être interprété comme un degré de désaccord entre les matrices de coût de $\mathfrak{G}$: plus le désaccord est grand et plus on a besoin d'itérations pour converger vers la solution. On note aussi la dépendance de l'initialisation par la présence de $\operatorname{RKP}(\mathfrak{P}_0, \mathfrak{G})$ et de la dimension intrinsèque de $\mathfrak{G}$ (dimension affine). Ainsi, si $\mathfrak{G}$ se trouve dans un sous-espace affine de petite dimension alors le nombre d'itérations est réduit, ce qui est le cas par exemple pour $\mathfrak{G}$ pris comme l'enveloppe convexe de quelques matrices de coût.

Comparaison à d'autres stratégies d'optimisation utilisées en minimax OT
Nous rappelons que nous souhaitons calculer le point selle de notre problème, et pas seulement le plan de transport qui minimise $\max_{\mathbf{C} \in \mathfrak{G}} \langle ., \mathbf{C} \rangle$, ou la matrice des coût qui maximise $\min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \cdot \rangle$. Cette dernière est concave, mais elle n'est pas différentiable pour le problème de Kantorovitch non régularisé, car le polytope de transport Π a une frontière non lisse. Ce manque de différentiabilité est mis en évidence dans Paty and Cuturi (2019, Section 5.2), et est motive les auteurs pour considérer le transport optimal régularisé dans leur Algorithme 2, ce qui leur permet d'obtenir le point selle.

Quant à StrOT (Alvarez-Melis et al., 2018), les auteurs trouvent le point selle dans le cas d'un transport optimal non régularisé en appliquant l'algorithme *Saddle Point Mirror Prox* (SP-MP, voir par exemple Bubeck (2015, Section 5.2.3)). Cela étant dit, SP-MP exige que la fonction de coût du problème min-max (un produit intérieur dans notre cas et dans le cas de Alvarez-Melis et al. (2018)) soit lisse, alors que l'algorithme de Mutapcic and Boyd (2009) que nous utilisons, après l'avoir adapté pour trouver le point selle, fonctionne même lorsque le coût considéré est non différentiable. Nous pouvons imaginer imposer certaines contraintes à la matrice de coût $\mathbf{C}$ en résolvant le problème suivant :

$$\min_{\mathbf{P} \in \Pi} \max_{\mathbf{C} \in \mathfrak{G}} \langle \mathbf{P}, \mathbf{C} \rangle + \Psi(\mathbf{C})$$

où Ψ est une fonction Lipschitzienne concave non différentiable. C'est le cas de $\Psi : \mathbf{C} \to -\|\mathbf{C} - \mathbf{C}_0\|_1$ par exemple, où $\mathbf{C}_0$ est une matrice de coût de référence. Dans ce cas, l'algorithme de Mutapcic and Boyd (2009) est théoriquement garanti de trouver la solution, alors que SP-MP ne l'est pas.

F.5.2.4 Variations pour différents choix de $\mathfrak{G}$

Nous exprimons le Problème (F.53) sur $\mathfrak{P}_t \times \mathfrak{G}$ à une étape $t \geq 0$ de Algorithm 1, pour un nombre fini de matrices de coût et pour la famille de coûts de Mahalanobis, deux cas que nous avons présentés précédemment. Le résultat suivant concerne le premier cas.

Proposition F.5.4. *Soit* $\mathfrak{G} = \mathrm{Conv}(\{\mathbf{C}_1, ..., \mathbf{C}_K\})$. *Alors, pour* $t \geq 0$, *résoudre le problème* (F.53) *sur* $\mathfrak{P}_t \times \mathfrak{G}$ *est équivalent au problème d'optimisation linéaire suivant:*

$$\begin{aligned} \min_{\mathbf{p} \in \mathbb{R}_+^K} \quad & \mathbf{1}_K^T \mathbf{p} \\ s.t. \quad & \mathbf{G}\mathbf{p} \geq \mathbf{1}_{|\mathfrak{P}_t|}, \end{aligned} \tag{F.55}$$

où $\mathbf{G} \in \mathbb{R}^{|\mathfrak{P}_t| \times K}$ *avec* $\mathbf{G}_{kl} = \langle \mathbf{P}_k, \mathbf{C}_l \rangle$. *De plus, le point selle* $(\mathbf{P}^*, \mathbf{C}^*)$ *est donné par*

$$\mathbf{C}^* = \frac{\sum_{k=1}^K p_k^* \mathbf{C}_k}{\sum_{k=1}^K p_k^*}, \qquad \mathbf{P}^* = \frac{\sum_{l=1}^{|\mathfrak{P}_t|} q_l^* \mathbf{P}_l}{\sum_{l=1}^{|\mathfrak{P}_t|} q_l^*},$$

où $\mathbf{p}^* = (p_1, \cdots, p_K)$ *et* $\mathbf{q}^* = (q_1, \cdots, q_{|\mathfrak{P}_t|})$ *sont les solutions de* (F.55) *et son dual.*

Pour le cas d'une famille de distance de Mahalanobis, nous proposons un résultat plus générale pour l'ensemble suivant:

$$\mathfrak{G}_{\mathbf{C}} = \{\mathbf{C} + \mathbf{E}^{\mathbf{M}} \in \mathbb{R}^{m \times m'}; (\mathbf{E}^{\mathbf{M}})_{ij} = (\mathbf{x}_i - \mathbf{x}'_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}'_j); \mathbf{M} \in \mathbb{S}_+^{n \times n}; \|\mathbf{M}\|_p \leq r\}, \tag{F.56}$$

où $r > 0$ est un rayon arbitraire.

Proposition F.5.5. *propNoncenteredMahalanobis Pour une matrice* $\mathbf{C}$ *fixée, soit* $\mathfrak{G}_{\mathbf{C}}$ *défini comme dans* (F.56). *Alors, pour* $t \geq 0$, *résoudre* (5.9) *sur* $\mathfrak{P}_t \times \mathfrak{G}_{\mathbf{C}}$, *est équivalent au problème d'optimisation convexe suivant*

$$\min_{\mathbf{P} \in \mathrm{Conv}(\mathfrak{P}_t)} r \|\mathbf{V}_{\mathbf{P}}\|_q + \langle \mathbf{P}, \mathbf{C} \rangle . \tag{F.57}$$

D'ailleurs, si $\mathbf{P}^*$ *est une solution optimale du Problème* (5.14), *alors* $\mathbf{M}^*$ *est exprimée comme dans Proposition F.5.1 où l'on remplace le couplage* $\mathcal{P}$ *par la matrice* $\mathbf{P}^*$.

F.5.2.5 Vers une notion de stabilité pour les matrices de coût

Nous définissons une nouvelle notion de stabilité d'une matrice de coût pour le transport optimal en utilisant l'ensemble $\mathfrak{G}$ de la Proposition F.5.5.

Définition F.5.1. *Pour une matrice de coût* $\mathbf{C}$ *et son ensemble ce matrice de coûts* $\mathfrak{G}_{\mathbf{C}}$ *introduit dans* (F.56), *et pour* $r > 0$, *on définit l'instabilité* $\mathrm{WS}_{\mathbf{C},r}$ *comme suit:*

$$\mathrm{WS}_{\mathbf{C},r} := W_{\mathfrak{G}_{\mathbf{C}}}(\mathcal{D}, \mathcal{D}') - W_{\mathbf{C}}(\mathcal{D}, \mathcal{D}') = \min_{\mathbf{P} \in \Pi} \max_{\|\mathbf{M}\| \leq r} \left\langle \mathbf{P}, \mathbf{C} + \mathbf{E}^{\mathbf{M}} \right\rangle - \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C} \rangle .$$

Cette définition traduit le fait que la distance de Wasserstein entre $\mathcal{D}$ et $\mathcal{D}'$ associée à une matrice de coût stable doit ne pas beaucoup différer de la distance de Wasserstein calculée en se basant sur la pire matrice de coût dans le voisinage de $\mathbf{C}$. Pour comparer les différentes valeurs d'instabilité pour une famille de fonctions de coût $\{\mathbf{C}_i\}_{i=1}^K$, nous normalisons chaque $\mathbf{C}_i$ soit en divisant par sa norme de Frobenius, soit par le coût de transport optimal qui lui est associé.

F.5.3 Expériences

F.5.3.1 Convergence et temps d'exécution

Nous considérons le cas où $\mathfrak{G}$ est l'enveloppe convexe d'un ensemble fini de matrices de coût. La convergence de Algorithm 1 est illustrée sur la Figure 5.3a (à gauche). La convergence est visiblement de plus en plus lente lorsque $|\mathfrak{G}|$ augmente, un comportement prédit par la garantie sur le nombre d'itérations. Ce dernier est néanmoins significativement plus faible que celui prédit par la borne théorique. Sur la même figure, à gauche, nous testons l'algorithme pour le transport régularisé, où l'on constate que plus la régularisation est importante, *i.e.* plus le coefficient de régularisation augmente, et plus la convergence s'accélère.

Sur la Figure 5.4a, nous comparons le temps d'exécution de notre algorithme à celui d'une approche directe qui résout le problème $\mathrm{RKP}(\Pi, \mathfrak{G})$ en considérant le problème d'optimisation linéaire:

$$\begin{aligned} \min_{\substack{\mathbf{P}\in\Pi \\ \eta\geq 0}} \quad & \eta, \\ \text{s.t.} \quad & \langle \mathbf{P}, \mathbf{C}_l \rangle \leq \eta \qquad \forall 1 \leq l \leq n. \end{aligned}$$

Notre algorithme devient de plus en plus rapide avec l'augmentation du nombre de matrices de coût considérées, contre une perte d'optimalité de l'ordre de 10^{-7}.

F.5.3.2 Comparaison à SRW

Nous considérons la distribution uniforme sur l'hypercube $[-1, 1]^n$ en dimension n pour la distribution $\mathcal{D}$, et la distribution $\mathcal{D}' := T\#\mathcal{D}$, où $T : \mathbf{x} \mapsto \mathbf{x} + 2\,\mathrm{sgn}\,(\mathbf{x}) \odot (\sum_{i=1}^{k} \mathbf{e}_i)$ et $\otimes$ dénote le produit matriciel de Hadamard et $1 \leq k \leq n$. Par construction, transporter les exemples tirés de $\mathcal{D}$ vers $\mathcal{D}$ se fait dans un sous-espace de dimension k. 3 cas sont testés:

1. Distance euclidienne carrée après une projection sur les vecteurs de la base canonique:
$$\mathfrak{G} = \{\mathbf{C}_s \in \mathbb{R}^{m\times m'} | (\mathbf{C}_s)_{ij} = ((\mathbf{x}_i - \mathbf{x}'_j)^T \mathbf{e}_s)^2; 1 \leq s \leq n\}$$

2. Distance euclidienne carrée après une projection sur toutes les combinaisons de deux vecteurs de la base canonique:
$$\mathfrak{G} = \{\mathbf{C}_{sl} \in \mathbb{R}^{m\times m'} | (\mathbf{C}_{sl})_{ij} = ((\mathbf{x}_i - \mathbf{x}'_j)^T (\mathbf{e}_s + \mathbf{e}_l))^2; 1 \leq s < l \leq n\}$$

3. La boule Mahalanobis avec la norme euclidienne centrée en 0 (Section 5.2.2.1).

La Figure 5.5 montre la comparaison de ces trois cas avec le couplage obtenu par SRW pour $k = 2$ (Paty and Cuturi, 2019) et le couplage avec le transport optimal classique W_2 ayant pour fonction coût la distance euclidienne carrée. SRW et nos 3 cas considérés arrivent à retrouver les direction de transport mieux qu'un problème OT classique avec la distance euclidienne. En particulier, notre approche avec la boule Mahalanobis est résistante au fléau de la dimensionnalité sans nécessiter d'hyperparamètre k fixant la dimension du sous-espace considéré en amont, ce qui est illustré dans sur la Figure 5.6.

F.5.3.3 Stabilité et sensibilité au bruit

Nous illustrons la corrélation entre la stabilité d'une matrice de coût et la sensibilité de la distance de Wasserstein à la présence de bruit, en utilisant deux jeux de données simulées et réelles. Après avoir aléatoirement généré une famille de matrices de coût $\{\mathbf{C}_i\}_{i=1}^{5} 0$ à partir de matrices PSD aléatoires définissant la distance de Mahalanobis, nous calculons leurs stabilités respectives $\mathrm{WS}_{\mathbf{C}_i, r=0.01}$. De plus, nous introduisons du bruit pour chaque

$\mathbf{C}_i$ avec une matrice de coût $\mathbf{E}^{\mathbf{N}}$ construite à partir d'une matrice PSD de norme de Frobenius égale à r, et nous calculons la sensibilité au bruit comme suit:

$$\mathrm{NS}_{\mathbf{C}_i} = \left| \min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C}_i \rangle - \min_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C}_i + \mathbf{E}^{\mathbf{N}} \rangle \right|.$$

Nous comparons cette dernière quantité à l'instabilité sur la Figure 5.7, sur laquelle on déduit une corrélation remarquable entre les deux quantités.

F.5.3.4 Transfert de couleurs

Notre but ici est transférer les couleurs d'une image de crépuscule rougeâtre (Figure 5.8, gauche) à une image de l'océan en pleine journée bleuâtre. Nous faisons le transfert sur un sous ensemble de pixels dans chaque image et étendons la transformation induite par le transport optimal à toute l'image en utilisant la procédure décrite dans Ferradans et al. (2014). Comme dans l'expérience précédente, nous générons une famille $\{\mathbf{C}_i\}_{i=1}^{5}0$ avec une matrice PSD de norme 1 comme matrices de distance "principales", et nous générons aussi 50 matrices aléatoires non liées à la tâche question. Les résultats sont observables sur la Figure 5.9, et montrent un écart significatif en terme de stabilité entre les matrices de coût Mahalanobis et celles aléatoires. De plus, nous visualisons le résultat de transfert de couleur conséquent pour les matrices de Mahalanobis, et on remarque que même si la matrice la plus stable et celle de la distance euclidienne conduisent à des résultats visuellement similaires, une évaluation plus fine montre que l'on a moins de discontinuité de couleurs sur l'image en utilisant la matrice la plus stable.

F.6 Conclusion

Tout au long de cette dissertation, nous avons traité le problème difficile d'adaptation de domaine, et nous y avons fourni des contributions de trois points de vu, à savoir l'apprentissage avec des fonctions de similarité, la résolution d'une tâche de classification à vaste marge, et l'utilisation du transport optimal et de ses variations pour dériver des algorithmes efficaces.

Motivés par le manque d'analyse théorique de l'apprentissage basé sur des fonctions de similarité quand les distributions produisant les données d'entraînement et de test changent, notre première contribution concerne l'extension de la théorie de fonctions de similarité $(\epsilon, \gamma, \tau)-$bonnes à l'adaptation de domaine. Nous avons répondu à une question concernant la convenabilité d'une fonction de similarité sur un domaine cible en terme de sa performance sur la source et de la divergence entre les deux domaines et entre leurs distributions de points de référence potentiellement différentes. Cependant, avec du recul, nous avons prouvé le manque d'intérêt à être contraint par le cadre (ϵ, γ, τ), duquel nous avons gardé seulement l'aspect vaste marge de classification puisque ce dernier est crucial dans l'évaluation de la confiance d'un classifieur en ses prédictions.

Dans notre seconde contribution, nous avons prouvé de nouvelles bornes sur le risque de violation de marge sur le domaine cible. En plus d'avoir généralisé des résultats antérieurs de la litérature, ces garanties nous ont mené à définir une nouvelle variation adversariale du problème de transport optimal qui est plus attachée à l'espace d'hypothèses considéré, à l'opposé d'autres méthodes d'adaptation de domaine basées sur sur l'OT. Puis, nous avons dérivé un algorithme d'adaptation de domaine prenant la forme d'un problème d'optimisation convexe. Finalement, nous avons testé cet algorithme sur des données simulées et réelles après les avoir projetées dans des espaces induits par des fonctions de similarité.

Finalement, avec la possibilité de généraliser le terme de transport optimal introduit à des situations plus diverses, nous avons défini une nouvelle variation du problème de

Monge-Kantorovitch appelée RKP, où le plan de transport et la fonction de coût sont appris conjointement sous la forme d'un problème minimax. Nous avons exploré plusieurs cas particuliers de cette formulation, conditionnés par l'ensemble de coûts terrains et nous avons décrit comment résoudre leurs problèmes d'optimisation associés. Par le biais d'évaluations empiriques, nous avons montré que notre formulation réussit à capter les directions de déplacement dans des sous-espaces de petite dimension et nous permet de définir une nouvelle notion de stabilité de matrice de coûts.

Bibliography

M. E. Abbasnejad, D. Ramachandram, and R. Mandava. A survey of the state of the art in learning the kernels. *Knowledge and Information Systems*, 31(2):193–221, 2012.

R. Aljundi, R. Emonet, D. Muselet, and M. Sebban. Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

D. Alvarez-Melis, T. Jaakkola, and S. Jegelka. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1771–1780, 2018.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, volume 70, pages 214–223, 2017.

S. Arlot, A. Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

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

F. R. Bach and M. I. Jordan. Learning spectral clustering. In *NIPS*, pages 305–312, 2004.

M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Domain adaptation on the statistical manifold. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2481–2488, 2014.

M. Baktashmotlagh, M. Har, i, and M. Salzmann. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research*, 17(108):1–30, 2016.

M. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 287–298, 2008a.

M.-F. Balcan, A. Blum, and N. Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1):89–112, 2008b.

P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel methods—support vector learning*, pages 43–54, 1999.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

A. Bellet, A. Habrard, and M. Sebban. Similarity Learning for Provably Accurate Sparse Linear Classification. In *International Conference on Machine Learning*, 2012.

A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

S. Ben-David and R. Urner. On the Hardness of Domain Adaptation and the Utility of Unlabeled Target Samples. In *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 139–153, 2012.

S. Ben-David and R. Urner. Domain adaptation–can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence*, 70(3):185–202, 2014.

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

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010.

S. Ben-David, T. Lu, T. Luu, and D. Pal. Impossibility Theorems for Domain Adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

D. P. Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty.* PhD thesis, Massachusetts Institute of Technology, 1971.

B. Bhushan Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

C. M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, 2006.

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

J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, page 120, 2006.

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

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

N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71:1–71:10, 2016.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

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

K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain Separation Networks. *arXiv:1608.06019 [cs]*, 2016b.

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

D. W. Boyd. The power method for lp norms. *Linear Algebra and its Applications*, 9: 95–101, 1974.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

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

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3):231–357, 2015.

R. Caseiro, J. F. Henriques, P. Martins, and J. Batista. Beyond the shortest path: Unsupervised domain adaptation by sampling subspaces along the spline flow. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3846–3854, 2015.

O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics lecture notes-monograph series. 2007.

G. Chechik, U. Shalit, V. Sharma, and S. Bengio. An online algorithm for large scale image similarity learning. In *NIPS*, pages 306–314, 2009.

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

M. Chen, K. Q. Weinberger, and Y. Chen. Automatic feature decomposition for single view co-training. In *ICML*, 2011b.

Y. Chen, G. Wang, and S. Dong. Learning with progressive transductive support vector machine. *Pattern Recognition Letters*, 24(12):1845–1855, 2003.

S. Chopra, S. Balakrishnan, and R. Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, volume 2, 2013.

W. Chu, F. D. L. Torre, and J. F. Cohn. Selective Transfer Machine for Personalized Facial Action Unit Detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

C. Cortes and M. Mohri. Domain adaptation in regression. In *ALT*, 2011.

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

C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

C. Cortes, Y. Mansour, and M. Mohri. Learning Bounds for Importance Weighting. In *Advances in Neural Information Processing Systems 23*, pages 442–450. 2010.

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

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865, 2016.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

G. Csurka. Domain Adaptation for Visual Applications: A Comprehensive Survey. *arXiv:1702.05374 [cs]*, 2017.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300. 2013.

M. Cuturi and D. Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

J. M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

H. Daumé and D. Marcu. Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1):101–126, 2006.

H. Daumé III. Frustratingly Easy Domain Adaptation. *arXiv:0907.1815 [cs]*, 2009.

J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic Metric Learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 209–216, 2007.

S. Dhouib and I. Redko. Analyse théorique de l'apprentissage avec des fonctions de similarités pour l'adaptation de domaine. In *Conférence sur l'Apprentissage Automatique 2018*, Rouen, France, 2018a.

S. Dhouib and I. Redko. Revisiting (\epsilon, \gamma, \tau)-similarity learning for domain adaptation. In *Advances in Neural Information Processing Systems 31*, pages 7397–7406. 2018b.

S. Dhouib, I. Redko, and C. Lartizien. On learning a large margin classifier for domain adaptation based on similarity functions. In *21 eme Conférence sur l'Apprentissage Automatique (CAp)*, Toulouse, France, 2019.

S. Dhouib, I. Redko, T. Kerdoncuff, R. Emonet, and M. Sebban. A swiss army knife for minimax optimal transport. In *Thirty-seventh International Conference on Machine Learning*, 2020a.

S. Dhouib, I. Redko, and C. Lartizien. Margin-aware adversarial domain adaptation with optimal transport. In *Thirty-seventh International Conference on Machine Learning*, 2020b.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

J. L. Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3):455–486, 1940.

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised Visual Domain Adaptation Using Subspace Alignment. In *ICCV 2013*, pages 2960–2967, Sydney, Australia, 2013.

S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

R. Flamary and N. Courty. Pot python optimal transport library, 2017.

A. Forrow, J.-C. Hütter, M. Nitzan, P. Rigollet, G. Schiebinger, and J. Weed. Statistical Optimal Transport via Factored Couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465, 2019.

Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

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

A. Genevay, G. Peyre, and M. Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

B. Geng, D. Tao, and C. Xu. DAML: Domain Adaptation Metric Learning. *IEEE Transactions on Image Processing*, 20(10):2980–2989, 2011.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 353–360, New York, NY, USA, 2009.

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers. In *International Conference on Machine Learning*, pages 738–746, 2013.

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

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. PAC-Bayesian Theorems for Domain Adaptation with Specialization to Linear Classifiers. Research Report, 2016b.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002.

M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

B. Gong, K. Grauman, and F. Sha. Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation. In *International Conference on Machine Learning*, pages 222–230, 2013.

B. Gong, K. Grauman, and F. Sha. Learning kernels for unsupervised domain adaptation with applications to visual object recognition. *International Journal of Computer Vision*, 109(1-2):3–27, 2014.

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

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

R. Gopalan, Ruonan Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 International Conference on Computer Vision*, pages 999–1006, 2011.

R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2288–2302, 2013.

L. Gottlieb, L. Kontorovich, and R. Krauthgamer. Efficient classification for metric data. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 433–440, 2010.

A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *IPMI*, pages 261–272, 2015.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22*, pages 673–681. 2009a.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009b.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

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

R. Grone. Certain isometries of rectangular complex matrices. *Linear Algebra and its Applications*, 29:161–171, 1980.

Z.-C. Guo and Y. Ying. Guaranteed classification via regularized similarity learning. *Neural Computation*, 26(3):497–522, 2014.

A. Habrard, J.-P. Peyrache, and M. Sebban. Boosting for Unsupervised Domain Adaptation. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 433–448, 2013a.

A. Habrard, J.-P. Peyrache, and M. Sebban. Iterative Self-labeling Domain Adaptation for Linear Structured Image Classification. *International Journal on Artificial Intelligence Tools*, 2013b.

M. Harel and S. Mannor. Learning from multiple outlooks. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 401–408, 2011.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval. *Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich*, 2001.

N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

E. Hille and R. S. Phillips. *Functional analysis and semi-groups*, volume 31. 1996.

J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, page 218–227, San Francisco, CA, USA, 1998.

P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman. Online metric learning and fast similarity search. In *NIPS*, pages 761–768, 2009.

T. Joachims. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209, 1999.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800. 2009.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009.

G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. Contrastive Adaptation Network for Unsupervised Domain Adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4888–4897, 2019.

L. Kantorovich. On mass transportation. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 37, pages 199–201, 1942.

P. Kar and P. Jain. Similarity-based Learning via Data Driven Embeddings. In *Advances in Neural Information Processing Systems 24*, pages 1998–2006. 2011.

W. Karush. Minima of functions of several variables with inequalities as side conditions. Master's thesis, Department of Mathematics, University of Chicago, Chicago, IL, USA, 1939.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–457, Boston, MA, 2000.

V. Koltchinskii and D. Panchenko. Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

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

D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *CoRR*, abs/1908.08729, 2019.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, Calif., 1951.

B. Kulis. Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, 5 (4):287–364, 2013.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *ICML*, volume 37, pages 957–966, 2015.

C. Laclau, I. Redko, B. Matei, Y. Bennani, and V. Brault. Co-clustering through optimal transport. In *ICML*, pages 1955–1964, 2017.

R. Lajugie, F. R. Bach, and S. Arlot. Large-margin metric learning for constrained partitioning problems. In *ICML*, pages 297–305, 2014.

M. T. Law, Y. Yu, M. Cord, and E. P. Xing. Closed-form training of mahalanobis distance for supervised clustering. In *CVPR*, pages 3909–3917, 2016.

T.-N. Le, A. Habrard, and M. Sebban. Deep multi-Wasserstein unsupervised domain adaptation. *Pattern Recognition Letters*, 125:249–255, 2019.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

C.-P. Lee and C.-J. Lin. A study on l2-loss (squared hinge-loss) multiclass svm. *Neural computation*, 25(5):1302–1323, 2013.

R. Li, X. Ye, H. Zhou, and H. Zha. Learning to match via inverse optimal transport. *J. Mach. Learn. Res.*, 20:80:1–80:37, 2019.

W. Li, Y. Gao, L. Wang, L. Zhou, J. Huo, and Y. Shi. OPML: A one-pass closed-form solution for online metric learning. *Pattern Recognition*, 75:302–314, 2018.

S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2014.

M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105, 2015.

M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2208–2217, 2017.

J. R. Magnus. A representation theorem for (trap) 1/p. *Linear Algebra and its Applications*, 95:127–134, 1987.

P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

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

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 367–374, Arlington, Virginia, USA, 2009a.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009b.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain Adaptation with Multiple Sources. In *Advances in Neural Information Processing Systems 21*, pages 1041–1048. 2009c.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the rÉnyi divergence. In *UAI*, pages 367–374, 2009d.

A. Margolis. A literature review of domain adaptation with unlabeled data. *Tec. Report*, pages 1–42, 2011.

D. McClosky, E. Charniak, and M. Johnson. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 337–344. Association for Computational Linguistics, 2006.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

F. Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

J. Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

G. Monge. *Mémoire sur la théorie des déblais et des remblais*. 1781.

J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

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

A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods and Software*, 24(3):381–406, 2009.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. 1983.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

M.-I. Nicolae, É. Gaussier, A. Habrard, and M. Sebban. Joint semi-supervised similarity learning for linear classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 594–609. Springer, 2015.

O. Nikodym. Sur une généralisation des intégrales de m. j. radon. *Fundamenta Mathematicae*, 15(1):131–179, 1930.

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

S. J. Pan, J. T. Kwok, and Q. Yang. Transfer Learning via Dimensionality Reduction. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 677–682, 2008.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

P. Panareda Busto and J. Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017.

F. Paty and M. Cuturi. Subspace robust wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5072–5081, 2019.

K. Pearson. LIII. On lines and planes of closest fit to systems of points in space, 1901.

M. Perrot and A. Habrard. Regressive Virtual Metric Learning. In *Advances in Neural Information Processing Systems 28*, pages 1810–1818. 2015a.

M. Perrot and A. Habrard. A Theoretical Analysis of Metric Hypothesis Transfer Learning. In *ICML*, 2015b.

P. O. Pinheiro. Unsupervised Domain Adaptation with Similarity Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018.

J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods-Support Vector Learning*, 208, 1998.

B. Quanz and J. Huan. Large margin transductive transfer learning. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1327–1336, 2009.

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

I. Redko, A. Habrard, and M. Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 737–753. Springer, 2017.

I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. A survey on domain adaptation theory. *CoRR*, abs/2004.11829, 2020.

F. Riesz. Démonstration nouvelle d'un théorème concernant les opérations fonctionnelles linéaires. *Annales scientifiques de l'École Normale Supérieure*, 3e série, 31:9–14, 1914.

B. Roark and M. Bacchiani. Supervised and unsupervised pcfg adaptation to novel domains. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 126–133. Association for Computational Linguistics, 2003.

L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *J. Mach. Learn. Res.*, 11:905–934, 2010.

A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018.

S. Saitoh. *Integral transforms, reproducing kernels and their applications*, volume 369. 1997.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55 (58-63):94, 2015.

K. Schacke. On the kronecker product. Master's thesis, University of Waterloo, 2003.

B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *Artificial Neural Networks — ICANN'97*, Lecture Notes in Computer Science, pages 583–588, 1997.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

S. Shalev-shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *ICML*, pages 743–750, 2004.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, Stability and Uniform Convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

J. Shen, Y. Qu, W. Zhang, and Y. Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

R. Shu, H. H. Bui, H. Narui, and S. Ermon. A DIRT-T Approach to Unsupervised Domain Adaptation. *arXiv:1802.08735 [cs, stat]*, 2018.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

M. Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

D. Steinberg. Computation of matrix norms with applications to robust optimization. *Research thesis, Technion-Israel University of Technology*, 2, 2005.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.

M. Sugiyama and K.-R. Müller. Model selection under covariate shift. In *International Conference on Artificial Neural Networks*, pages 235–240. Springer, 2005.

M. Sugiyama, K.-R. Müller, et al. Input-dependent estimation of generalization error under covariate shift. *Statistics and Decisions-International Journal Stochastic Methods and Models*, 23(4):249–280, 2005.

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

B. Sun and K. Saenko. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, volume 4, pages 24–1, 2015.

B. Sun and K. Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. *arXiv:1607.01719 [cs]*, 2016.

B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2058–2065, 2016.

S. Sun, H. Shi, and Y. Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92, 2015.

Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17: 138–155, 2009.

E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

R. Urner, S. Shalev-Shwartz, and S. Ben-David. Access to unlabeled data can speed up prediction time. In *ICML*, 2011.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems 4*, pages 831–838. 1992.

V. Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

V. N. Vapnik and A. Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280, 1971.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold,

R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, pages 589–615, 2006.

M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *Advances in Neural Information Processing Systems 18*, pages 1473–1480. 2006.

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

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

E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance Metric Learning with Application to Clustering with Side-Information. In *Advances in Neural Information Processing Systems 15*, pages 521–528. 2003.

H. Xu and S. Mannor. Robustness and generalization. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 503–515, 2010.

H. Xu and S. Mannor. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012.

H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the Class Weight Bias: Weighted Maximum Mean Discrepancy for Unsupervised Domain Adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 945–954, 2017.

K. You, X. Wang, M. Long, and M. Jordan. Towards Accurate Model Selection in Deep Unsupervised Domain Adaptation. In *International Conference on Machine Learning*, pages 7124–7133, 2019.

T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain Adaptation under Target and Conditional Shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

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

L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.

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

P. Zhao and Z. Zhou. Label distribution learning by optimal transport. In *AAAI*, pages 4506–4513, 2018.

S. Zhao, B. Li, C. Reed, P. Xu, and K. Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *ArXiv*, abs/2002.12169, 2020.

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

J. Zhu, S. Rosset, R. Tibshirani, and T. J. Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, pages 49–56, 2004.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *arXiv preprint arXiv:1911.02685*, 2019.

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

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : DHOUIB
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 23/11/2020

Prénoms : Sofiane

TITRE : Contributions to unsupervised domain adaptation: similarity functions, optimal transport and theoretical guarantees

NATURE : Doctorat

Muméro d'ordre : 2020LYMEI117

Ecole doctorale : Electronique, Electrotechnique et Automatique

Spécialité : Traitement du Signal et de l'Image

RESUME : Nous nous intéressons à l'adaptation de domaine, une branche de l'apprentissage statistique où l'on considère un jeu de données d'entraînement ayant des étiquettes observables, appelée source, et un jeu de données de test aux étiquettes non accessibles, appelé cible. Contrairement au cadre classique de l'apprentissage supervisé, les deux jeux de données proviennent potentiellement de deux distributions de probabilité différentes, nommées elles aussi distributions source et cible.
Nos contributions portent essentiellement sur l'aspect méthode dans l'objectif de développer un algorithme d'adaptation de domaine pour les classifieurs binaires à vaste marge. Au début de la thèse, nous nous sommes intéressés à l'apprentissage avec des fonctions de similarité dites (ϵ,τ)-bonnes dans le cadre d'adaptation de domaine, vu que ces fonctions ont été introduites dans la littérature dans le cadre classique de l'apprentissage supervisé. C'est le sujet de notre première contribution dans laquelle nous étudions théoriquement la performance d'une fonction de similarité sur une distribution cible, étant donné qu'elle est convenable pour la distribution source. Ensuite, nous avons abandonné ce cadre pour nous orienter plus généralement vers la classification à vaste marge dans le cadre de l'adaptation domaine et en partant suppositions plus faibles que celles prises dans la première contribution. Dans ce contexte, nous avons proposé une nouvelle étude théorique et un algorithme d'adaptation de domaine, ce qui constitue notre deuxième contribution. Cette dernière contient un aspect que nous avons généralisé pour travailler sur une variation adversariale du problème du transport optimal, ce qui est le sujet de notre dernière contribution.

MOTS-CLÉS : Apprentissage supervisé, adaptation de domaine, fonctions de similarité, transport optimal

Laboratoire (s) de recherche : Centre de recherche en acqusition et traitement de l'image pour la santé (CREATIS)

Directeurs de thèse: Carole LARTIZIEN et Ievgen REDKO

Président de jury : Michèle Sebag