Structural optimal transport for domain Adaptation with theoretical guarantees

Mourad El Hamri

UNIVERSITÉ SORBONNE PARIS NORD

NNT: 2022PA13XXXX

Thèse de doctorat

Structural Optimal Transport for Domain Adaptation with Theoretical Guarantees

Thèse de doctorat de l'Université Sorbonne Paris Nord
préparée au sein du Laboratoire d'Informatique de Paris Nord
et La Maison des Sciences Numériques

École doctorale n° 146 - École doctorale Galilée
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à LaMSN, le 15/12/2022, par

Mourad El Hamri

Composition du Jury :

Mathilde Mougeot Professeur, École Normale Supérieure Paris-Saclay	Présidente
Nicolas Courty Professeur, Université Bretagne Sud	Rapporteur
Marc Sebban Professeur, Université Jean Monnet	Rapporteur
Yann Guermeur Directeur de recherche, Loria	Examinateur
Basarab Matei Maître de conférences HDR, Université Sorbonne Paris Nord	Examinateur
Gabriel Peyré Directeur de recherche, École Normale Supérieure de Paris	Examinateur
Nicoleta Rogovschi Maître de conférences HDR, Université Paris Cité	Examinatrice
Abdelfattah Touzani Professeur, Université Sidi Mohamed Ben Abdellah	Examinateur
Younès Bennani Professeur, Université Sorbonne Paris Nord	Directeur de thèse
Issam Falih Maître de conférences, Université Clermont Auvergne	Co-encadrant

Dedicated to my parents
.. in love and gratitude

Remerciements

Voici, paraît-il, les pages les plus agréables et les plus délicates à écrire, celles que le lecteur découvre en premier et que l'auteur rédige en dernier. Certainement, j'aurais pu filer à la française, on n'a jamais connu personne se voir refuser un diplôme pour avoir manqué de politesse en omettant quelques remerciements. En revanche, ces deux pages découlent de ma gratitude envers un grand nombre de personnes qui m'ont témoigné une générosité scientifique ou personnelle et dont je suis redevable. Indubitablement, il est toujours périlleux de se prémunir contre la banalité lors d'une telle entreprise. Je nourris toutefois l'espoir que l'intense sentiment de reconnaissance que j'éprouve à l'égard de ceux qui m'ont épaulé trouvera résonance dans ces modestes propos.

Pour une thèse débutée il y a trois ans de cela, je pourrais remercier la terre entière et ce serait bien mon genre. Donnons-nous donc un peu de contenance en essayant de ne pas trop décevoir ceux qui viennent simplement vérifier ici si je les remercie à leur juste valeur, voire espèrent trouver quelques traits d'humour assez courants dans cet exercice. Quant à celles et ceux qui auraient ouvert cette thèse pour son contenu scientifique, je leur souhaite une agréable lecture.

C'est sans hésitation que mes remerciements se tournent en premier lieu vers mon directeur de thèse, Younès Bennani, qui m'a guidé tout au long de ces trois années faites de hauts et de bas, sans jamais se départir de son enthousiasme. Sa virtuosité scientifique et son implication en tant qu'encadrant m'ont accompagné dans mes pérégrinations, de l'adaptation de domaine au transport optimal en passant par les bornes de généralisation, sujets théoriques mais très riches et passionnants. Au-delà de l'encadrement scientifique, l'aspect humain est tout aussi important pour réussir à mener à bien une thèse, inévitablement parsemée de moments de doute. Merci donc d'avoir cru en moi, de m'avoir poussé à aller de l'avant et d'avoir réussi à atteindre l'équilibre délicat d'être un guide indéfectible tout en me proposant une immense liberté.

Je me trouve également reconnaissant envers mon co-encadrant, Issam Falih, qui m'a beaucoup apporté sur le plan scientifique tout en étant présent pour m'aider sur le plan humain. Je n'oublierai jamais son aide. Je le remercie pour la passion qu'il m'a transmise, pour la patience qu'il a toujours eue avec moi et pour ses encouragements.

Je suis très reconnaissant envers les deux rapporteurs de cette thèse, Nicolas Courty et Marc Sebban, auteurs de certains des articles qui ont le plus influencé mes travaux sur le transport optimal et l'adaptation de domaine. Je suis donc particulièrement heureux du fait qu'ils aient pris de leur temps précieux pour lire le contenu de ce manuscrit et m'ont fait part de leurs commentaires enrichissants et encourageants. Je remercie aussi vivement les autres membres du jury dont j'admire le travail et qui m'ont fait l'honneur de s'intéresser au mien. Merci à Yann Guermeur, Basarab Matei, Mathilde Meugeot, Nicoleta Rogovschi, Abdelfattah Touzani et spécialement à Gabriel Peyré, qui m'a également permis de suivre son cours sur le transport optimal à l'ENS de la rue d'Ulm au début de ma thèse.

Je voudrais remercier toutes celles et tous ceux qui font que la jungle fourmillante que constitue LaMSN prend parfois des airs de jardin familier. J'ai une pensée particulière pour Ahmed, qui a été bien plus qu'un ami, il a été un frère. Merci à Yohan pour sa bienveillance et pour les agréables discussions littéraires que nous avons eues. Merci à Marc pour son humour et sa jovialité amusante. Sa passion pour le tsikoudiá sur les plages de Chersónissos m'a toujours été impressionnante. Merci également à Chloé, Yassir, Khalil, Abir, Nosseiba, Amer, Léo et Pépita. Merci à Hamid pour son dévouement et à Omar pour ses mélodies inachevées et son thé à la menthe. Merci aux anciens membres de LaMSN, Fatima-Ezzahrae, Sarah et Kaoutar. Merci à Faouzi pour son immense gentillesse. Merci à Nistor pour son humour inégalé. Sa présence et ses conseils m'ont souvent permis de me détendre et de me remonter le moral. Je n'oublierai jamais le temps merveilleux que nous avons passé à Marseille et en Crète. Merci à Bihes pour tous les moments agréables que nous avons pu partager chez lui, il a toujours été si généreux et attentionné, offrant à tout le monde une atmosphère chaleureuse et accueillante. Merci Clémence pour les instants très plaisants que nous avons partagés durant nos trajets en voiture et pour avoir égayé nos voyages avec tes goûts musicaux sensationnels.

Je tiens également à remercier tous les membres du LIPN, en particulier la directrice Frédérique Bassino pour sa gentillesse et sa disponibilité.

Ces trois années n'auraient bien sûr pas été les mêmes sans mes amis qui ont toujours été à mes côtés. Merci à Hamza, Omar, Yassine, Oussama, Anouar et Salah-Eddine. Quinze ans, c'est bien plus qu'un doctorat, c'est une vraie relation d'amitié à travers le temps. Je remercie également Laurence, Meryem, Othmane et Alejandra. Je vous dois une grosse excuse pour toutes les soirées que j'ai ratées à cause de mes deadlines de conférences qui m'emprisonnaient ! Je remercie également les doctorants qui, croisés en conférence à un coin ou un autre du monde, m'ont souvent permis de tisser ou resserrer des liens d'amitié. Mes pensées vont tout particulièrement à Irina et Gabrielle, avec une émotion sincère.

Je n'aurais sûrement pas eu le courage d'aller au bout de cette aventure sans ma famille, qui a toujours été présente même de l'autre côté de la Méditerranée et n'a jamais cessé de croire en moi. J'ai une pensée toute particulière et émue pour mes parents pour leur perpétuel soutien au cours de cette longue aventure et la confiance qu'ils m'ont infailliblement témoignée. Je sais bien qu'ils ne comprennent rien à ce que je fais, mais leur amour me réchauffe le cœur. Si ce texte leur est dédié, c'est aussi parce qu'ils en ont été, peut-être inconsciemment, les grands artisans. Merci également à Hicham, Jalal et Douae, la plus belle fratrie qui soit. Enfin, merci Pauline, pour ta patience et ta présence durant les périodes difficiles de cette thèse.

J'avais dit un peu de contenance, alors trêve de remerciements, passons aux choses sérieuses.

Mourad El Hamri
Neuilly-sur-Seine
Décembre, 2022

Avant-propos

L'être humain, étant l'apprenant le plus puissant de la planète, a accumulé depuis la nuit des temps un grand nombre de compétences d'apprentissage. L'une d'entre elles est la flexibilité cognitive face à un environnement en constant changement (Uddin, 2021). À titre d'exemple, un enfant regardant La Jeune Fille à la perle de Vermeer, La Tête de Léda de Léonard de Vinci ou une photo de Natalie Portman, ne devrait avoir aucun problème à reconnaître qu'il s'agit d'une femme malgré le caractère très différent de chaque portrait. Une tâche aussi simple pour le cerveau humain peut s'avérer compliquée pour les modèles du machine learning. En effet, un modèle d'apprentissage supervisé entraîné sur des images photographiques peut se révéler incapable de généraliser à des dessins ou des tableaux (Venkateswara et al., 2017).

Cette incapacité est due au fait que les modèles d'apprentissage supervisé reposent sur une hypothèse fondamentale, à savoir que les données d'apprentissage et de test sont tirées de la même distribution de probabilité (Mohri et al., 2018). Bien que ce postulat soit tout à fait légitime dans certaines applications, il devient irraisonnable pour de nombreux problèmes du monde réel (Pan and Yang, 2009). Pour revenir à notre exemple de classification d'images, la variation des dispositifs d'acquisition, la présence ou l'absence d'arrière-plan, ou le changement des conditions d'éclairage sont des manifestations d'écart de distribution entre les données d'apprentissage et de test, et peuvent affecter négativement les performances des modèles d'apprentissage supervisé (Saenko et al., 2010).

Par ailleurs, si l'étiquetage manuel peut sembler une solution réalisable, une telle approche n'est pas raisonnable en pratique, car il est souvent excessivement coûteux de collecter à partir de zéro un nouvel ensemble de données étiquetées de haute qualité avec la même distribution que les données de test, en raison du manque de temps, de ressources ou d'autres facteurs, et ce serait un immense gâchis de rejeter totalement les connaissances disponibles sur un ensemble d'apprentissage étiqueté différent, mais connexe.

Cette situation délicate a favorisé l'émergence de l'adaptation de domaine (Redko et al., 2019b), une branche du machine learning qui tient compte du changement de distributions entre les données d'apprentissage et les données de test, et dans laquelle ces distributions sont respectivement appelées domaines source et cible. Depuis, la recherche en adaptation de domaine s'est développée selon deux voies complémentaires. La première, purement théorique, vise à identifier les conditions qui reflètent la parenté entre les deux domaines et qui permettent d'apprendre malgré le changement de distribution. Parmi ces conditions, une faible divergence entre les deux domaines est commune à la quasi-totalité de la littérature sur l'adaptation de domaine, avec des variations en fonction du choix de la mesure de divergence. La seconde, plus algorithmique, cherche souvent à réduire la divergence entre les deux

domaines pour les rendre indiscernables à travers une procédure d'alignement, réduisant ainsi le problème à une tâche classique d'apprentissage supervisé.

Ce double objectif explique le succès sans égal de la théorie du transport optimal en adaptation de domaine. En effet, le transport optimal (Villani, 2009) induit une distance bien définie entre les distributions de probabilité, permettant ainsi de quantifier la divergence entre les domaines source et cible, et comme son nom l'indique, offre une possibilité géométrique de transporter un ensemble de points vers un autre selon le principe du moindre effort, conduisant ainsi à l'alignement des deux domaines. Il n'est donc pas surprenant de constater la production scientifique prolifique qui a suivi le travail fondateur de (Courty et al., 2016), où de nombreux chercheurs en adaptation de domaine se sont appuyés sur le transport optimal pour modéliser des tâches, calculer des solutions et fournir des garanties théoriques (Redko et al., 2017; Courty et al., 2017; Shen et al., 2018; Redko et al., 2019a; Rakotomamonjy et al., 2022).

Cependant, en exploitant uniquement la capacité naturelle du transport optimal à capturer la géométrie sous-jacente des données, on laisse de côté d'autres informations structurelles importantes qui ne sont pas capturées directement à partir des distances par paires entre les données dans l'espace d'entrée. La définition de structure dans cette thèse doit être comprise dans le sens implicite du mot, c'est-à-dire lorsque des labels ou d'autres métadonnées peuvent conférer une structure latente aux échantillons ou lorsqu'il pourrait y avoir une prior ou un biais structurel sur la représentation des données (Battaglia et al., 2018). Par opposition au sens traditionnel qui fait référence à une structure explicite dans les données d'intérêt, par exemple lorsque celles-ci consistent en des séquences, des arbres ou des graphes.

L'incorporation d'une telle information structurelle peut susciter certaines propriétés désirables en adaptation de domaine, comme la préservation compacte des classes pendant le transport. C'est d'ailleurs ce qui a conduit (Courty et al., 2016) à proposer l'inclusion de cette information en ajoutant un régularisateur favorisant la parcimonie structurelle dans le plan de transport optimal, de telle sorte qu'une donnée cible ne reçoive des masses que d'échantillons sources appartenant à la même classe. De leur côté, (Alvarez-Melis et al., 2018) ont tenté d'intégrer l'information structurelle dans le problème du transport optimal en développant une généralisation non linéaire basée sur les fonctions sous-modulaires. Cependant, l'application de cette méthode en adaptation de domaine ne prend en compte que les structures disponibles dans le domaine source étiqueté, en partitionnant les échantillons en fonction de leurs labels de classe, alors que chaque échantillon cible forme son propre cluster. Néanmoins, des structures cachées plus riches dans le domaine cible peuvent être exploitées.

À ce stade, de nombreuses questions naturelles se présentent à l'esprit : Comment apprendre les structures cachées dans le domaine cible non étiqueté ? Le transport optimal peut-il assurer cette tâche ? Est-il possible d'incorporer ces structures dans le problème du transport optimal ? Si c'est le cas, dans quelle mesure cette incorporation est bénéfique pour l'adaptation de domaine ? A-t-elle des garanties théoriques, et que pourraient apporter ces dernières par rapport aux résultats théoriques de l'état de l'art ? C'est dans cette optique que s'inscrit cette thèse, qui se propose

d'apporter des éléments de réponse à ces questions, en élaborant des approches incorporant dans le transport optimal des structures découvertes par le transport optimal lui-même. L'objectif ici est double, d'abord de trouver les structures cachées par des outils du transport optimal, puis de les induire dans le processus de transport pour résoudre efficacement la tâche d'adaptation de domaine.

Plan

Cette thèse couvre la majeure partie des travaux réalisés par l'auteur dans le cadre du doctorat, et se concentre principalement sur une seule ligne de recherche qui est le transport optimal structurel pour l'adaptation de domaine. Des travaux supplémentaires basés sur le transport optimal pour le clustering multi-vues et le clustering collaboratif (Ben Bouazza et al., 2019, 2022) ne sont pas inclus dans cette thèse. Le reste du manuscrit est organisé de la manière suivante :

- **Le chapitre 2** établit le contexte mathématique pour le reste de la thèse. Il est présenté comme une revue des concepts fondamentaux de la théorie du transport optimal, mettant en évidence les propriétés et les résultats clés qui forment la base des concepts présentés dans tous les chapitres suivants. Bien que ce chapitre contienne des notions cruciales auxquelles il sera fait référence tout au long de la thèse, un lecteur familier avec le transport optimal peut sans risque le passer.

- **Le chapitre 3** aborde la problématique principale de cette thèse, à savoir l'adaptation de domaine. Après avoir introduit de manière formelle les principaux concepts de la théorie de l'apprentissage statistique, nous définissons l'adaptation de domaine ainsi que ses différents cas de figure. Ensuite, nous mettons l'accent sur les résultats théoriques de l'état de l'art, consistant principalement en des bornes de généralisation et nous examinons plusieurs mesures de divergence menant à ces bornes. Nous concluons en soulignant les avancées algorithmiques dans le sujet. Un lecteur chevronné en adaptation de domaine peut passer cette partie bien qu'elle contienne des concepts cruciaux qui seront discutés tout au long de la thèse.

- **Le chapitre 4** présente la première contribution de cette thèse, consacrée au transport optimal hiérarchique pour l'adaptation de domaine. Il est basé sur notre publication (El Hamri et al., 2022b) et donne quelques réponses à la question de l'apprentissage des structures cibles cachées en utilisant le transport optimal, et leurs incorporation ensuite dans le processus du transport pour résoudre le problème d'adaptation de domaine. Il s'agit d'une formulation structurelle du transport optimal qui exploite, au-delà des informations géométriques capturées par la métrique de base, des informations structurelles plus riches dans les domaines source et cible. L'information structurelle dans le domaine source étiqueté est formée instinctivement en regroupant les échantillons dans des structures en fonction de leurs étiquettes de classe. L'exploitation des structures cachées dans le domaine cible non étiqueté est réduite au problème d'apprentissage des mesures de probabilité à

travers le barycentre de Wasserstein, que nous prouvons théoriquement être équivalent au clustering spectral.

- **Le chapitre 5** présente la deuxième contribution de cette thèse (El Hamri et al., 2022d), où nous étudions un nouveau cadre théorique d'adaptation de domaine à travers le transport optimal hiérarchique. Ce paradigme fournit des garanties théoriques sous la forme de bornes de généralisation en permettant de considérer l'organisation structurelle implicite des échantillons dans les deux domaines en classes ou clusters. De plus, nous fournissons une nouvelle mesure de divergence entre les domaines source et cible, appelée la distance de Wasserstein Hiérarchique, qui indique, sous des hypothèses modérées, quelles structures doivent être alignées pour mener à une adaptation réussie.

- **Le chapitre 6** porte sur une troisième contribution (El Hamri et al., 2021a,b,c), où nous développons cette fois-ci une approche semi-supervisée en raison de la nécessité d'autres techniques pour détecter les structures cachées dans le domaine cible, en dehors du clustering. En effet, ce travail porte sur l'élaboration d'une approche de propagation de labels basée sur le transport optimal. L'intérêt du transport optimal dans ce contexte est de capturer la géométrie de l'espace d'entrée dans son intégralité et les relations entre les échantillons étiquetés et non étiquetés avec une vision globale. Cela évitera de devoir utiliser, comme dans les approches traditionnelles, des relations locales ou par paires entre les données, et les inconvénients qui en découlent. Cette approche effectue une propagation de labels incrémentale, contrôlée par un score qui surveille la certitude des prédictions.

- **Le chapitre 7** présente la quatrième contribution basée sur (El Hamri et al., 2022a,c). Il concerne l'utilisation de la technique semi-supervisée développée dans le chapitre précédent pour apprendre des structures cachées dans le domaine cible et les utiliser afin de créer de manière incrémentale des structures sources augmentées, permettant l'apprentissage d'une suite de sous-espaces latents domaine-invariants et discriminants, au sein desquels il devient facile d'étiqueter progressivement les données du domaine cible.

- **Le chapitre 8** résume les principaux résultats présentés dans cette thèse. Nous discutons également les perspectives d'avenir potentielles pour chacune des contributions proposées, incluant des versions fédératives et multi-sources des méthodes d'adaptation de domaine proposées et des extensions des contributions théoriques.

"Nitens lux, horrenda procella, tenebris aeternis involuta."

- Évariste Galois

Contents

List of Figures

List of Tables

Notations

Optimal Transport

c	Cost function
C	Cost matrix
T	Transport map
γ	Transport plan
$\Pi(\mu, \nu)$	Transport plans set
$U(a, b)$	Transportation polytope
$\langle ., . \rangle_F$	Frobenius inner product
$\mathcal{W}_p$	Wasserstein distance of order p
H	Relative entropy
ε	Entropic regularization strength

Statistical Learning

$\mathcal{X}$	Input space
$\mathcal{Y}$	Label space [1]
$\mathcal{H}$	Hypothesis space
h	Hypothesis function
$\mathcal{D}$	An arbitrary distribution
$\mu_{\mathcal{D}}$	Marginal distribution of $\mathcal{D}$ over $\mathcal{X}$
$\mathcal{S}$	Source domain
$\mathcal{T}$	Target domain
$\mu_{\mathcal{S}}$	Marginal source distribution over $\mathcal{X}$
$\mu_{\mathcal{T}}$	Marginal target distribution over $\mathcal{X}$
l	Loss function
$\epsilon_{\mathcal{D}}$	True risk of h over $\mathcal{D}$
$\epsilon_{\hat{\mathcal{D}}}$	Empirical risk of h over $\mathcal{D}$
$\epsilon_{\mathcal{S}}$	Source risk
$\epsilon_{\mathcal{T}}$	Target risk
VC	Vapnik-Chervonenkis dimension
$\mathcal{R}_m$	Rademacher complexity

Linear Algebra and Measure theory

$trace(A)$	Trace of a square matrix A
A^T	Transpose of a square matrix A
$\|B\|$	Cardinal of a set B
$\mathcal{P}(\mathcal{X})$	Set of probability measures on a space $\mathcal{X}$
δ_x	Dirac measure at x
$\mathbb{E}$	Expectation of a random variable
d	An arbitrary distance
d	Dimension of $\mathbb{R}^{\mathrm{d}}$

[1] Except in Chapter 2, where it denotes another input space different from $\mathcal{X}$.

CHAPTER 1

INTRODUCTION

> Structures don't take the streets.
>
> Lucien Goldmann

Humans, as the most powerful learners on the planet, have accumulated since dawn of time a lot of learning skills. One of these is cognitive flexibility to constantly changing environment (Uddin, 2021). By way of example, a child looking at Vermeer's Girl with a Pearl Earring, Da Vinci's Head of Leda, or Natalie Portman's photos, should have no problem recognizing that it is a woman despite the very different character of each portrait. Such an ostensibly simple task for human brain can be complicated for machine learning models. Indeed, a supervised learning model trained on photographic images may be incapable to generalize well to sketches or paintings (Venkateswara et al., 2017).

This incapacity is attributable to the reliance of supervised learning models on a fundamental assumption, namely that training and test data are drawn from the same probability distribution (Mohri et al., 2018).. While this postulate is quite legitimate in some applications, it becomes unreasonable for many real-world problems (Pan and Yang, 2009). Going back to our example of image classification, the variation of acquisition devices, the presence or absence of backgrounds, or the change of lighting conditions are manifestations of the shift between the training and test distributions, and can negatively affect the performance of supervised learning models (Saenko et al., 2010).

While manual labeling may appear like a feasible solution, such an approach is unreasonable in practice, since it is often prohibitively expensive to collect from scratch a new large high quality labeled dataset with the same distribution as the test data, due to lack of time, resources, or other factors, and it would be an immense waste to totally reject the available knowledge on a different, yet related labeled training set.

Such a challenging situation has promoted the emergence of domain adaptation (Redko et al., 2019b), a sub-field of machine learning, that takes into account the distributional shift between training and test data, and in which the training set and test set distributions are respectively called source and target domains. Since then, research on domain adaptation has developed along two complementary tracks. The first, being purely theoretical, aims to identify conditions that reflect the relatedness between the two domains and help to learn despite the distribution's shift. Among these conditions, a slight divergence between source and target domains is common to almost all the literature on domain adaptation, with variations depending

on the choice of how to quantify this divergence. The second, more algorithmic, often seeks to reduce the divergence between the two domains to make them indiscernible through an alignment procedure, thus reducing the problem to a classical supervised learning task.

This dual objective explains the unrivaled success of optimal transport theory in domain adaptation. Indeed, optimal transport (Villani, 2009) induces a well-defined notion of distance between probability distributions allowing to quantify the divergence between the source and target domains, and as its name suggests, offers a geometrically driven possibility of transporting a set of points to another according to the principle of least effort, thus leading to the alignment of both domains. It is therefore not surprising to see the prolific scientific production ensuing the seminal work of (Courty et al., 2016), where many domain adaptation researchers have relied on optimal transport to model tasks, compute solutions, and provide theoretical guarantees (Courty et al., 2017; Redko et al., 2017; Shen et al., 2018; Redko et al., 2019a; Rakotomamonjy et al., 2022).

However, exploiting only the natural ability of optimal transport to capture the underlying geometry of the data, leaves other information behind, since there is further important structural information that remains uncaptured directly from the pairwise distances between data in the input space. The definition of structure in this thesis must be understood in the implicit sense of the word, namely when labels or other metadata might confer a latent structure to samples or when there could be structural priors on their representation (Battaglia et al., 2018). As opposed to the traditional meaning that refers to an explicit structure in the data of interest, such as when these consist of sequences, trees, or graphs.

Inducing such structural information can elicit some desired properties in domain adaptation, like preserving compact classes during transportation. It is, moreover, what led (Courty et al., 2016) to propose the inclusion of this structural information by adding a regularizer that promotes group sparsity in the optimal transport plan, such that a given target data receives masses only from source samples belonging to the same class. On their side, (Alvarez-Melis et al., 2018) attempted to incorporate structural information into the optimal transport problem by developing a nonlinear generalization based on submodular functions. However, the application of this method in domain adaptation only takes into account the available structures in the labeled source domain, by partitioning samples according to their class labels, while every target sample forms its own cluster. Nonetheless, richer hidden structures in the target domain can eventually be exploited.

At this stage, many natural questions arise in mind: How to learn hidden structures in the unlabeled target domain? Can optimal transport handle this task? Is it possible to incorporate these structures into the optimal transport problem? If this is the case, to what extent this incorporation is rewarding for domain adaptation? Does it have any theoretical guarantees, and what could these latters bring compared to state-of-the-art theoretical results? It is in this light that this thesis takes place, intending to provide some answers to these questions, by elaborating approaches incorporating into optimal transport, structures learned by optimal transport itself. The goal here is twofold, first finding the hidden structures through optimal transport theory and then inducing them into the transportation process to efficiently solve the domain adaptation task.

Outline

This thesis covers mostly all the author's work conducted as part of the Ph.D. requirements, and mainly focuses on a single line of research that is structural optimal transport for domain adaptation. Additional work based on optimal transport for multi-view clustering and collaborative clustering (Ben Bouazza et al., 2019, 2022) are not included in this thesis. The rest of the manuscript is organized in the following way:

- **Chapter 2** sets up the mathematical background for the rest of the thesis. It is presented as a review of fundamental concepts of optimal transport theory, highlighting key properties and results that form the foundation of the concepts presented in all subsequent chapters. While this chapter contains crucial notions that will be referred to throughout the thesis, a reader familiar with optimal transport can safely skip it.

- **Chapter 3** addresses the main problematic of this thesis, namely domain adaptation. After introducing the major concepts of statistical learning theory in a formal way, the definition of domain adaptation is provided as well as its different settings that may occur. After that, we place great emphasis on the state-of-the-art theoretical results, consisting mostly of generalization bounds and we cover several measures of divergence leading to these bounds. We conclude by pointing out the algorithmic advances in the field. A seasoned domain adaptation reader may skip this part although it contains crucial concepts that will be discussed throughout the thesis.

- **Chapter 4** is dedicated to hierarchical optimal transport for domain adaptation. It is based on our publication (El Hamri et al., 2022b) and gives some answers to the question of learning hidden target structures using optimal transport and subsequently inducing them into the transportation process to resolve the problem of domain adaptation. This is a structural formulation of optimal transport that leverages beyond the geometrical information captured by the ground metric, richer structural information in the source and target domains. The additional information in the labeled source domain is formed instinctively by grouping samples into structures according to their class labels. While exploring hidden structures in the unlabeled target domain is reduced to the problem of learning probability measures through Wasserstein barycenter, which we prove theoretically to be equivalent to spectral clustering.

- **Chapter 5** corresponds to the contribution (El Hamri et al., 2022d), where we study a new theoretical framework of domain adaptation through hierarchical optimal transport. This paradigm provides theoretical guarantees in the form of generalization bounds and allows to consider the implicit structural organization of samples in both domains into classes or clusters. Additionally, we provide a new divergence measure between the source and target domains called Hierarchical Wasserstein distance that indicates under mild assumptions, which structures have to be aligned to lead to a successful adaptation.

- **Chapter 6** is based on our publications (El Hamri et al., 2021a,b,c), where we develop this time a semi-supervised approach because of the necessity of other techniques to detect hidden structures in the target domain, outside clustering. Indeed, this work is concerned with elaborating a label propagation approach based on optimal transport. The appeal of optimal transport in this setting is to capture the geometry of the entire input space and the relationship between labeled and unlabeled samples from a global level. This will prevent the necessity of using, as in traditional approaches, local or pairwise relationships between data, and the inconvenience that comes with it. This approach performs incremental label propagation, controlled by a score that watches over the certainty of predictions.

- **Chapter 7** represents the fourth contribution based on (El Hamri et al., 2022a,c). It deals with using the semi-supervised technique developed in the previous chapter to learn hidden structures in the target domain and use them to incrementally create augmented source structures, allowing learning a sequence of domain-invariant and discriminative latent subspaces, within which it becomes easy to progressively label the target samples.

- **Chapter 8** summarizes the main results presented in this thesis. We also discuss future perspectives of each contribution, including federated and multi-source versions of the proposed approaches and extensions of the theoretical contributions.

CHAPTER 2

OPTIMAL TRANSPORT

Contents

In this chapter, we present the key concepts of optimal transport theory on which this thesis will rely. This presentation focuses on the theoretical and computational aspects of optimal transport, with the purpose of using them in domain adaptation and more generally in machine learning problems. We begin by introducing Monge's first formulation, followed by the relaxation of Kantorovich and its dual problem. The connections between the two formulations are outlined. We highlight the case where the ground cost is a distance to a power, which defines the Wasserstein distance. The computational challenges associated with optimal transport will lead us to study special cases of real line and Gaussian distributions that can be solved in closed form. Then we present entropy-regularized optimal transport that can be easily solved using Sinkhorn's algorithm, followed by its sample complexity. The problem of Wasserstein barycenter is carefully discussed through special cases where it has a closed form, and then a numerical approximation scheme. Thereafter, we address briefly three extensions of optimal transport, namely Sliced Wasserstein distance, Gromov-Wasserstein distance, and unbalanced optimal transport, before concluding by highlighting some notable optimal transport toolboxes.

There is nothing more practical than a good theory.

Kurt Lewin

2.1 Optimal transport: a new twist on an old problem

Well known for several centuries for its logistical and economic applications, the problem of optimal transport has undergone a spectacular revival because of its unsuspected links with fluid mechanics, partial differential equations, and other fields of mathematics. Currently, thanks to a series of theoretical and algorithmic advances, it is considered the new mathematics of machine learning. In this section, we briefly trace this two-century journey.

The Founding Fathers: Optimal transport is a long-standing problem that has matured over time to give birth to a rich mathematical theory and numerous applications. Roots of optimal transport can be traced back to 1781, when the French mathematician Gaspard Monge (Monge, 1781), originally motivated by his observation of workers moving soil from the ground to build fortifications, raised the problem of optimally mapping two measures μ and ν of equal mass onto each other, according to a cost that is equal to the distance traveled by the workers per unit of mass. Owing to its mathematical difficulty, a long period of sleep followed Monge's problem until the relaxation of the Soviet mathematician Leonid Kantorovich in the thick of World War II (Kantorovich, 1942), who instead of optimizing on one-to-one maps that push forward μ to ν, his sights turned to couplings between μ and ν. This new formulation has allowed optimal transport theory to thrive since it has been inserted into an appropriate framework that gave the possibility to find that solutions actually exist and to study them. Notably, the formulation of Kantorovich embraces the case of discrete distributions, which can be interpreted as a problem of resource allocation as addressed in (Tolstoi, 1930; Hitchcock, 1941). This discrete version of Kantorovich's problem was numerically solved by George Dantzig (Dantzig, 1949), with further algorithmic refinements starting from the 1950s with the development of the linear programming literature (Dantzig, 1951) and min-cost flow problems (Ford and Fulkerson, 1962; Goldberg and Tarjan, 1989), marking thus the end of a fruitful chapter in which optimal transport became one of the fundamental problems of mathematical programming.

A Phoenix Rising from the Ashes: The mathematical aspects of optimal transport including the difficult Monge problem, were increasingly better understood in the late 1980s. In his pioneering paper, (Brenier, 1987) demonstrated the existence of an optimal Monge map between measures that admit a density in the case of a quadratic ground cost and characterized this map as the unique transportation map that is the gradient of a convex function. This groundbreaking result served as a basis for further theoretical research on Monge maps. Specifically, it allows weaving a link with the Monge-Ampère partial differential equation, which (Caffarelli, 1991) used to show regularity properties of the optimal map in the quadratic case. (McCann, 1997) then provided measure interpolants that now take his name and that represent the optimal geodesic transport between two measures according to the Wasserstein distance, defined by optimal transport when the ground cost is a distance to a power $p \geq 1$. Realizing that the space of measures equipped with the

Wasserstein distance shares certain central properties with manifolds has opened the door to the fundamental work of (Jordan et al., 1998), who demonstrated that the Fokker-Plank equation can be recast as a Wasserstein proximal minimization scheme, known as the JKO scheme, of functional taking measures as arguments. This construction was perfected in (Ambrosio et al., 2005), where a gradient flow theory generalizing that of Euclidean spaces was constructed on the Wasserstein space. Additional links with partial differential equations and fluid mechanics were elaborated in (Benamou and Brenier, 2000), given the so-called dynamic formulation of optimal transport. These works paved the way for decisive contributions by both (Villani, 2009) and (Figalli et al., 2010) whose respective works on the Ricci curvature and isoperimetric inequalities, among others, were recognized with Fields medals.

A Swiss Army Knife for Machine Learning: Simultaneously, in the early 2000s, optimal transport theory began to emerge in more applied domains. In fact, discrete optimal transport has experienced a spectacular revival in (Rubner et al., 2000) for image retrieval tasks under the name of the earth mover's distance. From then, it was applied in image processing and computer graphics (Rabin et al., 2011; Bonneel et al., 2011). However, the impact of optimal transport in machine learning community has long been limited because of its computational complexity that reaches $\mathcal{O}(n^3 log(n))$ despite specialized solvers (Pele and Werman, 2009). This issue was mitigated by the addition of an entropic regularization term to Kantorovich's problem by (Cuturi, 2013). Entropic regularization not only guarantees the uniqueness of the solution by strict convexity but also allows to solve the corresponding problem in $\mathcal{O}(n^2)$ using Sinkhorn algorithm (Sinkhorn, 1964). As a consequence, this regularization has paved the way for a pervasive use in machine learning, for instance in supervised and semi-supervised learning (Frogner et al., 2015; Solomon et al., 2014), unsupervised learning (Arjovsky et al., 2017; Genevay et al., 2018), natural language processing (Kusner et al., 2015) and domain adaptation (Courty et al., 2016, 2017; Redko et al., 2019a), to name a few. However, applications of optimal transport in machine learning are still hampered by several problems. For example, the disadvantageous statistical properties of optimal transport due to its high sample complexity. (Weed and Bach, 2019) have proved that the estimation of Wasserstein distances necessitates an exponential number of samples with respect to the intrinsic dimension of the set on which measures are supported. Entropic regularization has been shown to not only mitigate computational challenges but also to allow for better sampling rates (Genevay et al., 2019). The statistical and computational burden of optimal transport is one of the aspects that the optimal transport community is currently tackling, yet other interesting lines of research on applications of optimal transport are also being explored. For example, it turned out in various works that the marginal constraints of optimal transport might be too constraining for some applications (Schiebinger et al., 2019; Frogner et al., 2015), which has prompted the elaboration of unbalanced optimal transport (Chizat, 2017), where the constraints are replaced by penalties. Another concern of the optimal transport community is that its applicability is usually restricted to the situation where samples are in a common ground metric space, which is mostly Euclidean. This restriction inhibits its application to a range of machine learning problems where there is additional explicit structural information about the data that cannot usually be described in the Euclidean framework, e.g. when samples are represented by graphs, trees, or time series, or when samples are in different metric spaces, which has led to the development of optimal transport on incomparable spaces (Vayer, 2020).

2.2 The problem of Monge

The problem of optimal transport, originally proposed by Gaspard Monge in 1781 (Monge, 1781) was motivated by military applications. The goal was to find how to transport a certain amount of soil from a quarry (déblai) to a construction site (remblai) in the most economical way. A formal contemporary formulation of this problem is given as follows:

Definition 2.1 (The problem of Monge) *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ \cup \{+\infty\}$ a positive cost function over $\mathcal{X} \times \mathcal{Y}$, which represents the work needed to move a unit of mass from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. The problem of Monge asks to find a measurable transport map $\mathrm{T} : \mathcal{X} \to \mathcal{Y}$ that transports the mass represented by the probability measure μ to the mass represented by the probability measure ν while minimizing the total cost of this transportation:*

$$(\mathcal{M}) \quad \inf_{\mathrm{T}} \left\{ \int_{\mathcal{X}} c(x, \mathrm{T}(x)) d\mu(x) \mid \mathrm{T}\#\mu = \nu \right\}, \tag{2.1}$$

where $\mathrm{T}\#\mu$ stands for the image measure of μ by T, defined by: for all measurable subset $\mathcal{B} \subset \mathcal{Y}$, $\mathrm{T}\#\mu(\mathcal{B}) = \mu(\mathrm{T}^{-1}(\mathcal{B}))$.

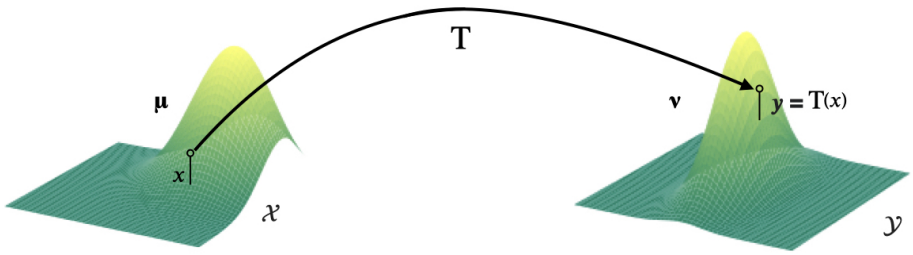

FIGURE 2.1: Illustration of Monge's problem: T is a transport map from $\mathcal{X}$ to $\mathcal{Y}$.

In general, finding such an optimal map T of Monge's problem is quite difficult since the solution may not exists, it is the case for instance when μ is a Dirac measure and ν is not. This also highlights the intrinsic asymmetry of this problem, as conversely, it is always possible to find a Monge map going to a Dirac measure. Moreover, the problem of Monge is highly nonlinear on T, and the constraint $\mathrm{T}\#\mu = \nu$ is not closed under weak convergence which is one of the major difficulties preventing from an easy analysis of Monge's problem. Thus, the problem of Monge has stayed an open question for many years despite some half-hearted attempts (Appell, 1887) and results on the existence of the optimal Monge map and how to characterize it have not even been addressed.

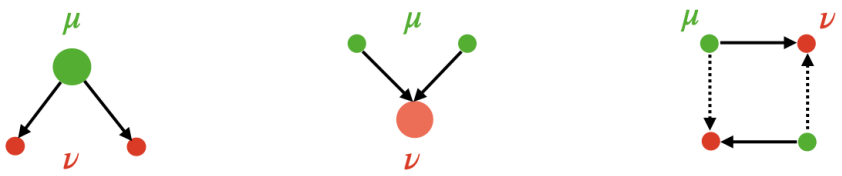

FIGURE 2.2: Monge's problem in the discrete case: (left) In this situation there is no Monge's map of μ onto ν because no function can satisfy $\mathrm{T}(x_1) = y_1$ and $\mathrm{T}(x_1) = y_2$ when $y_1 \neq y_2$. (center) The only possible Monge's map T is $\mathrm{T}(x_1) = y_1$ and $\mathrm{T}(x_2) = y_1$. (right) All points are equidistant from each other, then the solution of Monge's problem is not unique.

2.3 The problem of Monge-Kantorovich

A long period of sleep followed Monge's formulation until the convex relaxation of the Soviet mathematician Leonid Kantorovitch in the thick of World War II (Kantorovich, 1942). The main underlying idea behind the formulation of Kantorovich is to consider a probabilistic coupling γ instead of a deterministic map T to describe the displacement of the mass of μ: instead of specifying for each x, which is the destination $\mathrm{T}(x)$ of the mass originally located at x, we specify for each pair (x, y) the amount of mass going from x to y. A rigorous formulation of this problem is given in the following way:

Definition 2.2 (The problem of Monge-Kantorovich) *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ \cup \{+\infty\}$ a positive cost function over $\mathcal{X} \times \mathcal{Y}$. The problem of Monge-Kantorovich asks to find a joint probability measure $\gamma \in \Pi(\mu, \nu)$ that minimizes:*

$$(\mathcal{MK}) \quad \inf_{\gamma} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\gamma(x, y) \mid \gamma \in \Pi(\mu, \nu) \right\}, \tag{2.2}$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid proj_{\mathcal{X}}\#\gamma = \mu,\ proj_{\mathcal{Y}}\#\gamma = \nu\}$ is the transport plans set, constituted of all joint probability measures γ on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.

The constraints $proj_{\mathcal{X}}\#\gamma = \mu$, $proj_{\mathcal{Y}}\#\gamma = \nu$ mean that we restrict our attention to the movements that really take mass distributed according to μ and move it onto ν.

Example 2.3 (The problem of Monge-Kantorovich between discrete measures) *In the discrete setting, when measures μ and ν are only available through discrete samples $X = \{x_1, ..., x_n\} \subset \mathcal{X}$ and $Y = \{y_1, ..., y_m\} \subset \mathcal{Y}$, their empirical distributions can be expressed as $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, where $a = (a_1, ..., a_n)$ and $b = (b_1, ..., b_m)$ are vectors in the probability simplex $\sum_n = \{a \in \mathbb{R}_+^n \mid \sum_{i=1}^{n} a_i = 1\}$ and $\sum_m$ respectively. The cost function only needs to be specified for every pair $(x_i, y_j)_{\substack{1\le i\le n \\ 1\le j\le m}} \in X \times Y$ yielding a cost matrix $C \in \mathcal{M}_{n\times m}(\mathbb{R}^+)$. Then, the problem of Monge-Kantorovich becomes a linear program parametrized by the transportation polytope $U(a, b) = \{\gamma \in \mathcal{M}_{n\times m}(\mathbb{R}^+) \mid \gamma \mathbb{1}_m = a \text{ and } \gamma^{\mathrm{T}} \mathbb{1}_n = b\}$, which acts as a feasible set, and the matrix C which acts as a cost parameter:*

$$(\mathcal{D}_{\mathcal{MK}}) \quad \inf_{\gamma \in U(a,b)} \langle \gamma, C \rangle_F, \tag{2.3}$$

where $\langle ., . \rangle_F$ is the Frobenius inner product, defined by $\langle \gamma, C \rangle_F = trace(\gamma^{\mathrm{T}} C)$.

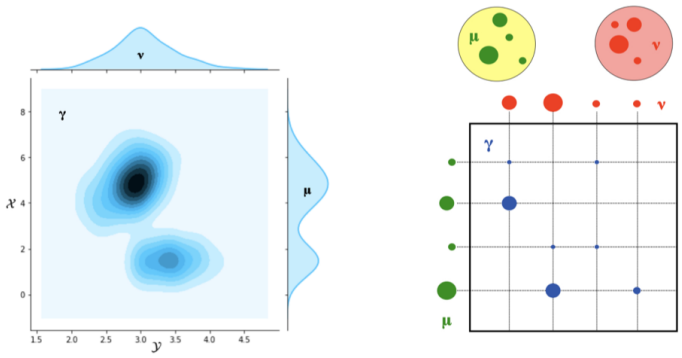

FIGURE 2.3: Continuous setting: The joint probability distribution γ is a transport plan between μ and ν (left). Discrete setting: The positive entries of the discrete transport plan γ are displayed as blue disks with a radius proportional to the entry values (right).

The problem of Monge-Kantorovich $(\mathcal{MK})$ is much easier to handle than the original one proposed by Monge $(\mathcal{M})$ for many reasons. For example, it is clear that if the mass splitting really occurs, then this movement cannot be described by a map T, whereas Kantorovich's formulation allows it since mass in x can a priori move to different destinations y. Moreover, there always exists a transport plan γ in $\Pi(\mu,\nu)$, i.e. $\gamma = \mu \otimes \nu$, unlike Monge's formulation where no transport map exists for instance when μ is a Dirac measure and ν is not. Furthermore, transport plans include transport maps, since $\mathrm{T}\#\mu = \nu$ implies that $\gamma = (Id \times \mathrm{T})\#\mu$ belongs to $\Pi(\mu,\nu)$. And lastly, in contrast to $(\mathcal{M})$, the formulation of Kantorovich $(\mathcal{MK})$ guarantees the existence of a solution under very general assumptions as shown by the following theorem:

Theorem 2.4 (Existence of an optimal transport plan) *Let $(\mathcal{X},\mu)$ and $(\mathcal{Y},\nu)$ be two Polish probability spaces and $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+ \cup \{+\infty\}$ a positive lower semi-continuous cost function. Then, the problem of Monge-Kantorovich $(\mathcal{MK})$ admits a solution.*

This existence theorem does not imply that the optimal cost is finite. It might be that all transport plans lead to an infinite total cost, i.e. $\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,d\gamma(x,y) = +\infty$ $\forall\gamma \in \Pi(\mu,\nu)$.

2.4 Kantorovich duality

The problem of Monge-Kantorovich $(\mathcal{MK})$ is a constrained convex minimization problem, and as such, it can be naturally paired with a dual problem called Kantorovich dual $(\mathcal{KD})$, which is a constrained concave maximization problem, defined as follows:

Proposition 2.5 (Kantorovich dual problem) *The Kantorovich dual problem $(\mathcal{KD})$ is the following:*

$$(\mathcal{KD}) \quad \sup_{\psi,\phi}\left\{\int_{\mathcal{X}}\psi(x)\,d\mu(x) + \int_{\mathcal{Y}}\phi(y)\,d\nu(y) \mid (\psi,\phi)\in\mathcal{R}(c)\right\}, \tag{2.4}$$

where $\mathcal{R}(c) = \left\{(\psi,\phi)\in L^1(\mu)\times L^1(\nu) : \forall(x,y), \psi(x)+\phi(y)\le c(x,y)\right\}$ is the set of admissible Kantorovich potentials.

Example 2.6 (Kantorovich dual problem between discrete measures) *The discrete Kantorovich dual problem $(\mathcal{D}_{\mathcal{KD}})$ is the following:*

$$(\mathcal{D}_{\mathcal{KD}}) \quad \max_{\mathrm{f},\mathrm{g}\in\mathcal{R}(c)} \langle \mathrm{f}, a\rangle + \langle \mathrm{g}, b\rangle, \tag{2.5}$$

where $\mathcal{R}(c) = \{(\mathrm{f},\mathrm{g})\in\mathbb{R}^n\times\mathbb{R}^m : \forall(i,j)\in[\![1,n]\!]\times[\![1,m]\!],\ \mathrm{f}_i+\mathrm{g}_j \le C_{i,j}\}$ is the set of admissible Kantorovich potentials.

One can naturally wonder if the Kantorovich dual problem $(\mathcal{KD})$ leads to the same optimum as the primal Monge-Kantorovich problem $(\mathcal{MK})$. To answer this question, we need to become familiar with a fundamental concept of optimal transport theory called cyclical monotonicity and the notion of c-concavity.

Definition 2.7 (Cyclical monotonicity) *Let $\mathcal{X},\mathcal{Y}$ be arbitrary sets and $c : \mathcal{X}\times\mathcal{Y} \to]-\infty,+\infty]$ be a function. A subset $\Gamma \subset \mathcal{X}\times\mathcal{Y}$ is said to be c-cyclically monotone if, for*

any $N \in \mathbb{N}$, and any family $(x_1, y_1), ..., (x_N, y_N)$ of points in Γ, holds the inequality:

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}), \tag{2.6}$$

with the convention $y_{N+1} = y_1$. A transport plan is said to be c-cyclically monotone if it is concentrated on a c-cyclically monotone set.

The c-cyclical monotonicity suggests that an optimal transport plan can not be improved. Then, it is obvious that an optimal transport plan should be c-cyclically monotone. The converse property is less obvious, but we will see that it holds true under mild conditions.

Definition 2.8 (c-transform) *Let $\mathcal{X}, \mathcal{Y}$ be sets, $c : \mathcal{X} \times \mathcal{Y} \to]-\infty, +\infty]$, and a function $\psi : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$. Its c-transform is the function $\psi^c : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ defined by:*

$$\forall y \in \mathcal{Y} \qquad \psi^c(x) = \inf_{x \in \mathcal{X}} (c(x, y) - \psi(x)). \tag{2.7}$$

Remark 2.9 *If $c = -x.y$ on $\mathbb{R}^n \times \mathbb{R}^n$, then the c-transform coincides with the usual Legendre transform.*

Definition 2.10 (c-concavity) *Let $\mathcal{X}, \mathcal{Y}$ be sets, and $c : \mathcal{X} \times \mathcal{Y} \to]-\infty, +\infty]$. A function $\phi : \mathcal{U} \to \mathbb{R} \cup \{-\infty\}$ is said to be c-concave if it is not identically $-\infty$, and there exists $\psi : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ such that $\phi = \psi^c$. Then its c-transform is the function ϕ^c defined by:*

$$\forall x \in \mathcal{X} \qquad \phi^c(x) = \sup_{y \in \mathcal{Y}} (\phi(y) - c(x, y)). \tag{2.8}$$

Remark 2.11 *If $c = d$ is a distance on some metric space $\mathcal{X}$, then a function ϕ is c-concave if and only if it is Lipschitz continuous with Lipschitz constant less than 1. Moreover, we have $\phi^c = -\phi$.*

We are now ready to provide the following characterization of the optimal transport plans:

Theorem 2.12 (Fundamental theorem of optimal transport) *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ a lower-semi continuous cost function, such that $\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\gamma(x, y) < \infty$. Then for any $\gamma \in \Pi(\mu, \nu)$ the following statements are equivalent:*

1. *γ is optimal*
2. *γ is c-cyclically monotone*
3. *There is a c-concave ψ such that, γ-almost surely, $\psi(x) + \psi^c(y) = c(x, y)$.*

A direct implication of the fundamental theorem of optimal transport is related to Kantorovich's duality. In fact, the equivalence between (i) and (iii) is very useful to prove the main theorem of this section:

Theorem 2.13 (Kantorovich duality) *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ \cup \{+\infty\}$ a positive lower semi-continuous cost function. Then, strong duality holds. More precisely the dual Kantorovich problem $(\mathcal{KD})$ leads to the same optimum as the primal Monge-Kantorovich problem $(\mathcal{MK})$. More formally:*

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\gamma(x, y) = \sup_{(\psi,\phi) \in \mathcal{R}(c)} \int_{\mathcal{X}} \psi(x)\, d\mu(x) + \int_{\mathcal{Y}} \phi(y)\, d\nu(y). \tag{2.9}$$

Remark 2.14 *The discrete Kantorovich dual problem $(\mathcal{D}_{\mathcal{KD}})$ problem leads to the same optimum as the discrete primal Monge-Kantorovich problem $(\mathcal{D}_{\mathcal{MK}})$ thanks to the strong duality for linear programs (Bertsimas and Tsitsiklis, 1997).*

2.5 Bridging Monge and Kantorovich

In light of the previous considerations, it is natural to ask under which conditions a Monge map might exist, and what links exist between Monge and Kantorovich formulations. In fact, in some cases with additional assumptions on the cost function c, it is possible to prove that the optimal transport plan γ does not allow mass splitting. The mass located at x is only sent to a unique destination $\mathrm{T}(x)$, thus providing a solution to the original problem of Monge $(\mathcal{M})$. That is what is done by Brenier in (Brenier, 1987). This result can be easily adapted to other costs such as strictly convex functions of the difference $x - y$ (Santambrogio, 2015).

Theorem 2.15 *Let μ and ν be two probability measures on a compact $\Omega \subset \mathbb{R}^d$, such that μ is absolutely continuous. Consider a cost function $c(x, y) = h(x - y)$ where h is a strictly convex function. Then, there exists a unique optimal transport map T and a unique optimal transport plan γ, and T and γ are related by $\gamma = (Id \times \mathrm{T})\#\mu$.*

Hence, under the conditions of Theorem 2.15, an optimal Monge map exists and can equivalently be described as an optimal transportation plan supported on its graph. In particular, Theorem 2.15 holds when $c(x, y) = ||x - y||^p$ with $p > 1$. The $p = 2$ case holds a particular place in the optimal transport theory, as shown by Brenier in his seminal paper (Brenier, 1987). The major contribution of Theorem 2.16 is the unique characterization of the transport map as the gradient of a convex function.

Theorem 2.16 *Let μ and ν be two probability measures on $\mathbb{R}^d$ with finite moment of order 2, such that μ is absolutely continuous with respect to the Lebesgue measure, and $c(x, y) = ||x - y||^2$. Then the problem $(\mathcal{M})$ admits a unique solution, which is characterized among all transport maps as being the gradient of a convex function φ: $\forall x \in \mathbb{R}^d \quad \mathrm{T}^*(x) = \nabla\varphi(x)$.*

In the previous theorem, we showed the uniqueness of the optimal transport plan by giving an explicit expression for the optimal map. Yet, it is possible to use a more general argument: every time that we know that any optimal transport plan γ must be induced by a map T, then we have uniqueness of T. The proof is easy and follows from the convexity of the optimal transport plans set.

2.6 Wasserstein distance

When $\mathcal{X} = \mathcal{Y}$ is a polish metric space endowed with a distance d, a natural choice is to use it as a cost function, e.g. $c(x, y) = d(x, y)^p$ for $p \in [1 + \infty[$. Then, the problem $(\mathcal{MK})$ induces a metric between probability measures over $\mathcal{X}$, called the p-Wasserstein distance.

Definition 2.17 (Wasserstein space) *Let $(\mathcal{X}, d)$ be a Polish metric space and let $p \in [1, \infty[$. The Wasserstein space of order p is defined as:*

$$\mathcal{P}_p(\mathcal{X}) = \left\{ \mu \in \mathcal{P}(\mathcal{X}) \mid \int_{\mathcal{X}} d(x_0, x)^p d\mu(x) < +\infty \right\}, \tag{2.10}$$

where $x_0 \in \mathcal{X}$ is arbitrary.

The space $\mathcal{P}_p(\mathcal{X})$ does not depend on the choice of the point x_0. In other words, $\mathcal{P}_p(\mathcal{X})$ is the space of probability measures that have a finite moment of order p.

Definition 2.18 (Wasserstein distance) *Let $(\mathcal{X}, d)$ be a Polish metric space and let $p \in [1, \infty[$. For any two probability measures μ, ν in $\mathcal{P}_p(\mathcal{X})$, the Wasserstein distance of order p between μ and ν is defined by:*

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} d(x, y)^p \, d\gamma(x, y) \right)^{1/p}, \tag{2.11}$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X}^2) \mid proj_{\mathcal{X}}\#\gamma = \mu, \, proj_{\mathcal{Y}}\#\gamma = \nu\}$ is the transport plans set, constituted of all joint probability measures γ on $\mathcal{X}^2$ that have marginals μ and ν.

Example 2.19 (Wasserstein distance between discrete measures) *Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ be two discrete probability distributions on $\mathcal{X}$, the Wasserstein distance of order p between μ and ν is defined by:*

$$\mathcal{W}_p(\mu, \nu) = \left(\min_{\gamma \in U(a,b)} \langle \gamma, C \rangle_F \right)^{1/p}, \tag{2.12}$$

where $U(a, b) = \{\gamma \in \mathcal{M}_{n \times m}(\mathbb{R}^+) \mid \gamma \mathbb{1}_m = a \text{ and } \gamma^{\mathbf{T}} \mathbb{1}_n = b\}$ is the transportation polytope and $C \in \mathcal{M}_{n \times m}(\mathbb{R}^+)$ is the matrix of pairwise distances between elements of X and Y raised to the power p.

Remark 2.20 *The distance $\mathcal{W}_1$ is commonly called the Kantorovich–Rubinstein distance. Theorem 2.13 and Remark 2.11 together lead to the useful duality formula: For any $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$:*

$$\mathcal{W}_1(\mu, \nu) = \sup_{||\psi||_{Lip} \leq 1} \left\{ \int_{\mathcal{X}} \psi \, d\mu - \int_{\mathcal{X}} \psi \, d\nu \right\}, \tag{2.13}$$

where $\{\psi : || \psi ||_{Lip} \leq 1\}$ is the set of all Lipschitz functions on $(\mathcal{X}, d)$ with Lipschitz constant at most 1.

Example 2.21 *$\mathcal{W}_p(\delta_x, \delta_y) = d(x, y)$. In this example, the distance does not depend on p.*

Example 2.22 *In the case where $c(x, y) = \mathbb{1}_{x \neq y}$, the Wasserstein distance between two probability distributions is equal to their total variation distance.*

The Wasserstein distance $\mathcal{W}_p$ enjoys a very interesting property, which is the metrizability of weak convergence in $\mathcal{P}_p(\mathcal{X})$:

Theorem 2.23 (Wasserstein distance $\mathcal{W}_p$ metrizes $\mathcal{P}_p(\mathcal{X})$) *Let $(\mathcal{X}, d)$ be a Polish metric space and let $p \in [1, \infty[$, then the Wasserstein $\mathcal{W}_p$ metrizes the weak convergence in $\mathcal{P}_p(\mathcal{X})$. In other words, a sequence of measure $(\mu_k)_{k \in \mathbb{N}}$ converges weakly in $\mathcal{P}_p(\mathcal{X})$ to another measure μ if and only if $\mathcal{W}_p(\mu_k, \mu) \to 0$.*

2.7 Special cases

Two special cases will be addressed in this section, notably, the case where μ and ν are probability distributions on the real line $\mathbb{R}$ and the case when they are Gaussian distributions in $\mathbb{R}^d$. These special cases are well known to have closed-form solutions as stated by the next theorems from (Santambrogio, 2015) and (Peyré and Cuturi, 2019) respectively.

The statement of the closed form solution of probability measures on $\mathbb{R}$ requires the following definition of cumulative distribution function and its pseudo inverse:

Definition 2.24 (Cumulative distribution function and its pseudo inverse) *Let μ be a probability measure on $\mathbb{R}$, i.e. $\mu \in \mathcal{P}(\mathbb{R})$. The cumulative distribution function $F_\mu : \mathbb{R} \to [0,1]$ is defined by:*

$$\forall x \in \mathbb{R} \quad F_\mu(x) = \mu(]-\infty, x]). \tag{2.14}$$

Its pseudo inverse $F_\mu^{-1} : [0,1] \to \mathbb{R}$, (also called the generalized quantile function) is given by:

$$\forall t \in [0,1] \quad F_\mu^{-1}(t) = \inf_t \{x \in \mathbb{R} \mid F_\mu(x) \geq t\}. \tag{2.15}$$

Theorem 2.25 (Closed-form expression on the real-line) *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ be two probability measures on $\mathbb{R}$. Consider the cost $c(x,y) = h(y-x)$ where $h : \mathbb{R} \to \mathbb{R}^+$ is a strictly convex function. Then, the problem of Monge-Kantorovich has a unique solution given by $\gamma = (F_\mu^{-1} \times F_\nu^{-1})\#\mathcal{L}_{[0,1]}$, where $\mathcal{L}_{[0,1]}$ is the Lebesgue measure restricted to $[0,1]$. In the case where μ is atomless, then γ is supported on the map $\mathrm{T}(x) = F_\nu^{-1}(F_\mu(x))$, i.e. $\gamma = (Id \times \mathrm{T})\#\mu$. If h is only convex then the optimal transport plan γ is still optimal but not necessarily unique.*

For discrete measures, when $c(x,y) = |x-y|^p$, this theorem stipulates that it is sufficient to sort the support of the distributions in order to find the optimal coupling. In the special case of discrete probability distributions with $m = n$ and $a = b = \mathbb{1}_n/n$, this corresponds to sort $x_1 < x_2 < \ldots < x_n$ and $y_1 < y_2 < \ldots < y_n$ and to associate x_1 with y_1, x_2 with y_2 and so on, in this case the p-Wasserstein distance has the simple formula: $\mathcal{W}_p^p(\mu,\nu) = \frac{1}{n}\sum_{i=1}^n |x_i - y_i|^p$. In the generic case of discrete probability distributions with $m \neq n$ or a and b are arbitrary vectors in the probability simplex Σ_n and Σ_m respectively, the previous theorem states that, after sorting the points, the optimal plan is obtained by putting as much mass as possible from x_1 to y_1 and to add the remaining mass to y_2. This procedure is repeated until there is no more mass left.

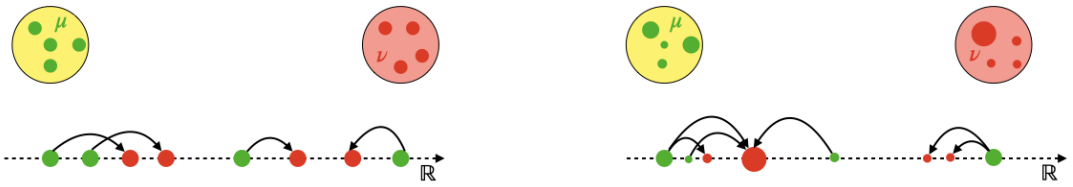

FIGURE 2.4: Real line discrete transport: uniform weights (left), non-uniform weights (right).

The other special case of closed-form solutions that arise when probability measures are Gaussian on $\mathbb{R}^d$ is provided by the following statement:

Theorem 2.26 (Closed-form expression for Gaussians) *Let $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$ be two Gaussians in $\mathbb{R}^d$. Consider the cost $c(x,y) = h(y-x)$ where h is a strictly convex function. Let*

$$\mathrm{T} : x \mapsto m_\nu + A(x - m_\mu), \tag{2.16}$$

where

$$A = \Sigma_\mu^{-\frac{1}{2}}(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})\Sigma_\mu^{-\frac{1}{2}}. \tag{2.17}$$

Then T *is the unique optimal solution of* $(\mathcal{M})$ *and* $\gamma = (Id \times \mathrm{T})\#\mu$ *is the unique optimal solution of* $(\mathcal{MK})$.

In particular, when $c(x, y) = ||x - y||_2$ *is the Euclidean distance on* $\mathbb{R}^d$*, the 2-Wasserstein distance is given by:*

$$\mathcal{W}_2^2(\mu, \nu) = ||m_\mu - m_\nu||_2^2 + \mathcal{B}(\Sigma_\mu, \Sigma_\nu)^2, \tag{2.18}$$

where $\mathcal{B}(\Sigma_\mu, \Sigma_\nu) = tr(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}})$ *is Bures metric.*

Remark 2.27 *A similar result of the previous theorem exists for elliptically contoured distributions that can be seen as a generalization of Gaussians (Gelbrich, 1990). In this case, the* $\mathcal{W}_2$ *admits also a closed form.*

2.8 Entropic regularization of optimal transport

Besides the special cases in the previous section that have closed-form solutions, solving optimal transport can be expensive to compute, even in the relatively simple discrete setting. Indeed, as stated before, the problem of optimal transport between two discrete probability measures $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ is a linear program, and thus can be solved with the simplex algorithm or interior point methods exactly in $\mathcal{O}(r^3 \log(r))$ where $r = \max(n, m)$ (Pele and Werman, 2009), which is a heavy computational price tag specially for large-scale machine learning. A further limitation of optimal transport is its curse of dimensionality. In fact, considering a probability measure μ over $\mathbb{R}^d$ and its empirical estimation $\hat{\mu}_n$, the sample complexity of the estimation of the Wasserstein distance is exponential in the dimension of the ambient space. More precisely $\mathbb{E}[\mathcal{W}_p(\mu, \hat{\mu}_n)] = \mathcal{O}(n^{\frac{-1}{d}})$ (Weed and Bach, 2019), thus the empirical distribution $\hat{\mu}_n$ becomes less and less representative as the dimension d of the ambient space $\mathbb{R}^d$ becomes large. In this section, we discuss how entropic regularization can help to overcome these two limitations.

First introduced in (Schrödinger, 1931) in statistical physics, the entropic regularization has received renewed attention in machine learning following (Cuturi, 2013), who showed that Sinkhorn's algorithm provides an efficient and scalable approximation to optimal transport. Let's start by presenting the regularized optimal transport problem and its dual form:

Definition 2.28 (The entropy-regularized optimal transport problem) *Let* $\varepsilon > 0$ *be the regularization strength. The entropy-regularized optimal transport problem is defined as:*

$$(\mathcal{MK}_\varepsilon) \quad \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\gamma(x, y) + \varepsilon H(\gamma|\mu \otimes \nu), \tag{2.19}$$

where $H(\gamma|\mu \otimes \nu) = \int_{\mathcal{X}\times\mathcal{Y}} \log(\frac{d\gamma(x,y)}{d\mu(x)d\nu(y)}) d\gamma(x, y) - \int_{\mathcal{X}\times\mathcal{Y}} d\gamma(x, y) + \int_{\mathcal{X}\times\mathcal{Y}} d\mu(x)d\nu(y)$ *is the relative entropy of the transport plan* γ *with respect to the product measure* $\mu \otimes \nu$.

As the relative entropy H is strictly convex in its first argument, this regularization term turns the convex problem $(\mathcal{MK})$ into a strictly convex problem $(\mathcal{MK}_\varepsilon)$, and as such, $(\mathcal{MK}_\varepsilon)$ has a unique solution.

Definition 2.29 (The dual entropy-regularized optimal transport problem) *The dual of entropy-regularized optimal transport problem reads:*

$$(\mathcal{KD}_\varepsilon) \sup_{(\psi,\phi)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \psi(x)\, d\mu(x) + \int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{\psi(x)+\phi(y)-c(x,y)}{\varepsilon}} d\mu(x) d\nu(y). \tag{2.20}$$

The dual entropy-regularized optimal transport problem can be rewritten as the maximization of an expectation with respect to the product measure $\mu \otimes \nu$ (Genevay et al., 2016):

Proposition 2.30 *The dual of entropy-regularized optimal transport $(\mathcal{KD}_\varepsilon)$ has the following equivalent formulation:*

$$\sup_{(\psi,\phi)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \mathbb{E}_{\mu\otimes\nu}\left[f_\varepsilon^{XY}(\psi,\phi)\right], \tag{2.21}$$

where $f_\varepsilon^{xy}(\psi,\phi) = \psi(x) + \phi(y) - \varepsilon e^{\frac{\psi(x)+\phi(y)-c(x,y)}{\varepsilon}}$.

The following result from (Genevay, 2019) further shows that strong duality holds:

Proposition 2.31 (Strong duality of entropy-regularized optimal transport problem) *The dual of entropy-regularized optimal transport problem $(\mathcal{KD}_\varepsilon)$ leads to the same optimum as the primal entropy-regularized optimal transport problem $(\mathcal{MK}_\varepsilon)$.*

Besides, the primal-dual relationship is given by:

$$d\gamma(x,y) = \exp\left(\frac{\psi(x)+\phi(y)-c(x,y)}{\varepsilon}\right) d\mu(x) d\nu(y) \tag{2.22}$$

and ψ,ϕ satisfy

$$\begin{aligned} \psi(x) &= -\varepsilon \log\left(\int_{\mathcal{Y}} e^{\frac{\phi(y)-c(x,y)}{\varepsilon}} d\nu(y)\right) \qquad \mu - a.s. \\ \phi(y) &= -\varepsilon \log\left(\int_{\mathcal{X}} e^{\frac{\psi(x)-c(x,y)}{\varepsilon}} d\mu(x)\right) \qquad \nu - a.s. \end{aligned} \tag{2.23}$$

Example 2.32 (Discrete entropy-regularized optimal transport problem) *In the discrete setting, the entropy-regularized optimal transport is defined as:*

$$(\mathcal{D}_{\mathcal{MK}_\varepsilon}) \quad \min_{\gamma\in U(a,b)} \langle \gamma, C\rangle_F + \varepsilon H(\gamma|a\otimes b), \tag{2.24}$$

where H[1]$(\gamma|a\otimes b) = \sum_{i,j} \gamma_{i,j} \log(\frac{\gamma_{i,j}}{a_i b_j}) - \gamma_{i,j} + a_i b_j$.

The problem $(\mathcal{D}_{\mathcal{MK}_\varepsilon})$ admits the following dual:

$$(\mathcal{D}_{\mathcal{KD}_\varepsilon}) \quad \max_{\mathrm{f}\in\mathbb{R}^n, \mathrm{g}\in\mathbb{R}^m} \langle \mathrm{f}, a\rangle + \langle \mathrm{g}, b\rangle - \varepsilon \langle e^{\mathrm{f}/\varepsilon}, \mathrm{K} e^{\mathrm{g}/\varepsilon}\rangle, \tag{2.25}$$

where $\mathrm{K} = e^{-C/\varepsilon}$.

Strong duality can be proven by the general duality theorem (Karush–Kuhn–Tucker conditions and Lagrangian formulation).

[1]Entropic regularization of optimal transport was first introduced with the following formulation of entropy: $H(\gamma) = \sum_{i,j} \gamma_{i,j} \log((\gamma_{i,j}) - 1)$, (Cuturi, 2013).

2.8.1 Sinkhorn's algorithm

This section is devoted to the resolution of the discrete entropy-regularized optimal transport problem $(\mathcal{D}_{\mathcal{MK}_\varepsilon})$ using Sinkhorn's algorithm.

Proposition 2.33 *The solution of the problem $(\mathcal{D}_{\mathcal{MK}_\varepsilon})$ is unique and has the form*

$$\gamma = \mathrm{diag}(\mathrm{u})\mathrm{K}\mathrm{diag}(\mathrm{v}), \tag{2.26}$$

for two unknown scaling variable $(\mathrm{u}, \mathrm{v}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$. The variables u, v in (2.26) are linked to f, g in (2.25) through the relations: $\mathrm{u} = e^{\frac{\mathrm{f}}{\varepsilon}}, \mathrm{v} = e^{\frac{\mathrm{g}}{\varepsilon}}$.

According to the previous proposition we have $\gamma = \mathrm{diag}(\mathrm{u})\mathrm{K}\mathrm{diag}(\mathrm{v})$. The variables (u, v) must therefore satisfy the following nonlinear equations which correspond to the mass conservation constraints $\gamma \in U(a, b)$:

$$\mathrm{diag}(\mathrm{u})\mathrm{K}\mathrm{diag}(\mathrm{v})\mathbb{1}_m = a \qquad \text{and} \qquad \mathrm{diag}(\mathrm{v})\mathrm{K}^{\mathbf{T}}\mathrm{diag}(\mathrm{u})\mathbb{1}_n = b, \tag{2.27}$$

These two equations can be further simplified, since $\mathrm{diag}(\mathrm{v})\mathbb{1}_m$ is simply v and $\mathrm{diag}(\mathrm{u})\mathbb{1}_n$ is simply u, thus:

$$\mathrm{u} \odot (\mathrm{Kv}) = a \qquad \text{and} \qquad \mathrm{v} \odot (\mathrm{K}^{\mathbf{T}}\mathrm{u}) = b, \tag{2.28}$$

where $\odot$ corresponds to entrywise multiplication of vectors.

This problem is known in the numerical analysis community as the matrix scaling problem. An intuitive way to handle these equations is to solve them iteratively, by modifying first u so that it satisfies the left-hand side of Equation (2.28) and then v to satisfy its right-hand side:

$$\mathrm{u}^{(l+1)} = \frac{a}{\mathrm{Kv}^{(l)}} \qquad \text{and} \qquad \mathrm{v}^{(l+1)} = \frac{b}{\mathrm{K}^{\mathbf{T}}\mathrm{u}^{(l+1)}}, \tag{2.29}$$

These two updates define Sinkhorn's algorithm (Cuturi, 2013) initialized with an arbitrary positive vector $\mathrm{v}^{(0)} = \mathbb{1}_m$. The name of the algorithm is due to Richard Sinkhorn who first proved the convergence of updates in (2.29) (Sinkhorn, 1964).

Algorithm 2.1 Sinkhorn's algorithm

Parameters: ε
Input : C
Compute $\mathrm{K} = e^{-C/\varepsilon}$
Initialize $\mathrm{v}^{(0)} = \mathbb{1}_m$
while *not converged* **do**
  $\mathrm{u}^{(l+1)} \leftarrow \frac{a}{\mathrm{Kv}^{(l)}}$
  $\mathrm{v}^{(l+1)} \leftarrow \frac{b}{\mathrm{K}^{\mathbf{T}}\mathrm{u}^{(l+1)}}$
end
return $\gamma_\varepsilon^* = \mathrm{diag}(\mathrm{u})\mathrm{K}\mathrm{diag}(\mathrm{v})$

A practical feature of Sinkhorn's algorithm is its ease of implementation and the fact that it can be efficiently parallelized using graphics processing units (GPUs) as noted in (Cuturi, 2013). Regarding its complexity, (Altschuler et al., 2017) showed that for $n = m$, Sinkhorn's algorithm computes a τ-approximate solution of the original discrete optimal transport problem in $\mathcal{O}(n^2 \log(n)\tau^{-3})$.

Proposition 2.34 (Impact of ε) *In the limit $\varepsilon \to 0$, the unique solution γ_ε^* of $(\mathcal{D}_{\mathcal{MK}_\varepsilon})$ converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem $(\mathcal{D}_{\mathcal{MK}})$. In the limit $\varepsilon \to +\infty$, γ_ε^* converges to $\mu \otimes \nu$.*

The previous proposition states that for a small regularization ε, the solution converges to the maximum entropy unregularized optimal transport plan, and for a large regularization ε, the solution converges to the transport plan with maximal entropy, namely the joint probability $\mu \otimes \nu$. Figures 2.5 and 2.6 show visually the effect of these two convergences.

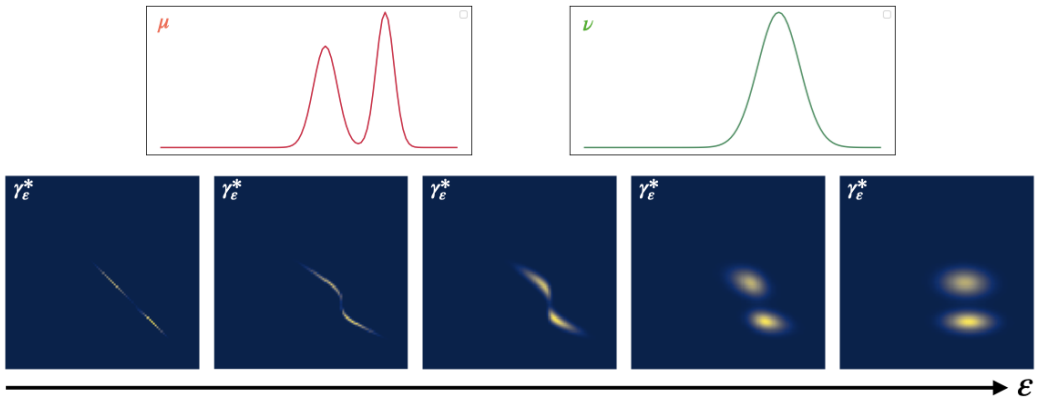

FIGURE 2.5: Impact of ε on the optimal transport plan γ_ε^* between two one-dimensional probability distributions. As ε increases the transport plan tends to blur and converges to the product $\mu \otimes \nu$.

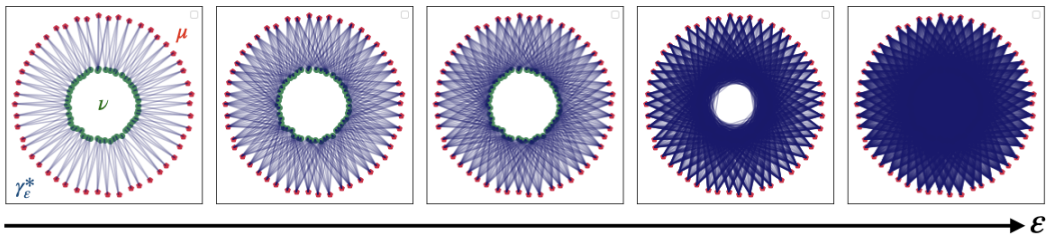

FIGURE 2.6: Impact of ε on the optimal transport plan γ_ε^* between two discrete probability distributions. As ε increases the transport plan becomes more and more dense.

From a practical point of view Sinkhorn's algorithm suffers from numerical stability issues when $\varepsilon \to 0$. In fact, when the regularization parameter ε is small compared to the entries of the cost matrix C, the kernel $\mathrm{K} = e^{-C/\varepsilon}$ becomes too negligible to be stored in memory as positive numbers and becomes instead null. This can then result in a matrix product $\mathrm{K}v$ or $\mathrm{K}^T u$ with ever smaller entries that become null and result in a division by 0 in the Sinkhorn update of Equation (2.29). Such issues can be partly resolved by carrying out computations on the multipliers u and v in the log domain. To this end (Schmitzer, 2019) suggest a log-sum-exp stabilization trick whose iterations turn out to be mathematically equivalent to the original iterations while being stable for arbitrary $\varepsilon > 0$. The downside is that it requires nm computations of exp at each step.

Furthermore, Sinkhorn's algorithm only copes with discrete measures. Thus, taking advantage of the formulation of the dual entropy-regularized optimal transport problem as the maximization of an expectation, outlined in Proposition 2.30, (Genevay et al., 2016) proposed to use stochastic optimization tools to cope with large-scale optimal transport problems and to handle discrete or continuous distributions. In fact, Stochastic Averaged Gradient (SAG) can be used to compute a solution in the discrete case, each iteration of this algorithm costs $\mathcal{O}(r)$ operations where $r = \max(n, m)$, which makes it scale better in large-scale problems than Sinkhorn's algorithm, while still enjoying a convergence rate of $\mathcal{O}(1/k)$, k being the number of iterations. The semi-discrete case (when μ is an arbitrary measure and ν is a discrete measure) can be solved using Averaged SGD with the convergence rate $\mathcal{O}(1/\sqrt{k})$. In the continuous setting, the problem is infinite-dimensional so it can not be solved using SGD anymore. (Genevay et al., 2016) proposed then to use a kernel expansion of the dual variables in a reproducing kernel Hilbert space (RKHS) and solve the problem using a kernel SGD with a quadratic complexity $\mathcal{O}(k^2)$.

2.8.2 Sample complexity

Let consider the problem of Monge-Kantorovich $(\mathcal{MK})$, and let $\mathcal{W}_c$ be the total cost:

$$\mathcal{W}_c(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\gamma(x, y) \tag{2.30}$$

and let consider also the entropy-regularized optimal transport problem $(\mathcal{MK}_\varepsilon)$, and let $\mathcal{W}_{c,\varepsilon}$ be the regularized total cost:

$$\mathcal{W}_{c,\varepsilon}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\gamma(x, y) + \varepsilon H(\gamma|\mu \otimes \nu). \tag{2.31}$$

In order to cancel the bias $\mathcal{W}_{c,\varepsilon}(\mu, \mu) \neq 0$ introduced by the entropic regularization, (Genevay et al., 2018) introduced the following corrected regularized divergence, called Sinkhorn divergence $\mathcal{SD}_{c,\varepsilon}$, defined as follows:

Definition 2.35 (Sinkhorn divergence) *The Sinkhorn divergence $\mathcal{SD}_{c,\varepsilon}$ between two probability measures μ, ν is defined as:*

$$\mathcal{SD}_{c,\varepsilon}(\mu, \nu) = \mathcal{W}_{c,\varepsilon}(\mu, \nu) - \frac{1}{2}\mathcal{W}_{c,\varepsilon}(\mu, \mu) - \frac{1}{2}\mathcal{W}_{c,\varepsilon}(\nu, \nu). \tag{2.32}$$

Far from simply correcting the bias of $\mathcal{W}_{c,\varepsilon}(\mu, \mu)$, the Sinkhorn divergence also appears as an interpolating discrepancy between Wasserstein distance and Maximum Mean Discrepancy (MMD) (Genevay et al., 2018):

Theorem 2.36 (Asymptotics of Sinkhorn Divergence with respect to ε) *The Sinkhorn Divergence $\mathcal{SD}_{c,\varepsilon}$ has the following asymptotic behavior in ε:*

1. *as $\varepsilon \to 0$, $\quad \mathcal{SD}_{c,\varepsilon}(\mu, \nu) \to \mathcal{W}_c(\mu, \nu)$*
2. *as $\varepsilon \to +\infty$, $\quad \mathcal{SD}_{c,\varepsilon}(\mu, \nu) \to \frac{1}{2} MMD^2_{-c}(\mu, \nu)$*

When $-c$ is a positive definite kernel, MMD_{-c} is the MMD with the kernel that is minus the cost used in the optimal transport problem.

(Feydy et al., 2019) then proved that the Sinkhorn divergence $\mathcal{SD}_{c,\varepsilon}$ defines a suitable loss function:

Theorem 2.37 *Let $c(x,y) = ||x-y||^p, p \geq 1$. Then for all compactly supported $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\mathcal{SD}_{c,\varepsilon}(\mu,\nu)$ defines a symmetric positive definite divergence, which is convex in μ or ν (but not jointly), and metrizes weak convergence.*

In statistical contexts, direct access to a distribution of interest μ is generally not available, instead, we have only access to i.i.d. samples from μ, or, equivalently, to an empirical distribution $\hat{\mu}_n$. For $\hat{\mu}_n$ to serve as a reasonable proxy to μ, we should insist that $\hat{\mu}_n$ and μ are close in the Wasserstein sense. In the large-n limit, this is indeed the case: if $(\mathcal{X}, d)$ is a polish metric space, then for any $p \in [1, +\infty[$, we have $\mathcal{W}_p(\mu, \hat{\mu}_n) \to 0$ μ - a.s. thanks to the weak convergence of $\hat{\mu}_n$ to μ almost surely (Varadarajan, 1958) and the metrization of weak convergence by the Wasserstein distance (Villani, 2009). This result raises the question of quantifying the speed or rate of convergence of $\hat{\mu}_n$ to μ in $\mathcal{W}_p$ distance. This rate is often called the sample complexity of $\mathcal{W}_p$. Unfortunately, when $\mathcal{X} = \mathbb{R}^d$, the convergence of $\hat{\mu}_n$ to μ exhibits the so-called curse of dimensionality (Bellman, 1961), since the convergence rate is $\mathbb{E}[\mathcal{W}_p(\mu, \hat{\mu}_n)] = \mathcal{O}(n^{-\frac{1}{d}})$ (Dudley, 1969). Thus, in the high-dimensional regime, the empirical distribution $\hat{\mu}_n$ becomes less and less representative as d becomes large, so that the convergence of $\hat{\mu}_n$ to μ in Wasserstein distance is slow. This notion of sample complexity is crucial in machine learning, as bad sample complexity implies overfitting and high gradient variance for parameter estimation. Sample complexity of optimal transport appears to be another major bottleneck for the use of optimal transport in high-dimensional machine learning problems.

A remedy to this problem lies, again, in entropic regularization. Indeed, (Genevay et al., 2019) showed that $\mathcal{SD}_{c,\varepsilon}$ benefits from the same sample complexity as MMD, scaling in $1/\sqrt{n}$ but with a constant that depends on the inverse of the regularization parameter:

$$\mathbb{E}[\mathcal{SD}_{c,\varepsilon}(\mu, \hat{\mu}_n)] \leq \mathcal{F}(\varepsilon)\mathcal{O}(n^{-\frac{1}{2}}). \tag{2.33}$$

2.9 Wasserstein barycenter

Wasserstein barycenter has become popular due to its ability to provide a natural extension of the notion of averaging points to the notion of averaging point clouds. Importantly, it naturally inherits the ability of optimal transport to capture the geometric properties of the data. This section is dedicated to the presentation of this interesting concept.

Definition 2.38 (Wasserstein Barycenter) *Given a set $(\nu_i)_{i\in[\![1,n]\!]}$ of probability measures defined on some space $\mathcal{X}$ and a stochastic vector $\lambda \in \Sigma_n$, a Wasserstein barycenter $\hat{\mu}$ of $(\nu_i)_{i\in[\![1,n]\!]}$ is a minimizer of the following variational problem:*

$$\inf_{\mu\in\mathcal{P}_p(\mathcal{X})} \sum_{i=1}^{n} \lambda_i \mathcal{W}_p^p(\mu, \nu_i). \tag{2.34}$$

Wasserstein barycenter is a special case of the so-called Fréchet mean or Karcher mean (Karcher, 2014) in the metric space $(\mathcal{P}_p(\mathcal{X}), \mathcal{W}_p)$. In a general metric space, finding Fréchet mean is usually a difficult nonconvex optimization problem. Fortunately, in the case of Wasserstein distance, the problem can be formulated as a convex program for which existence can be proved as we show in the following special cases:

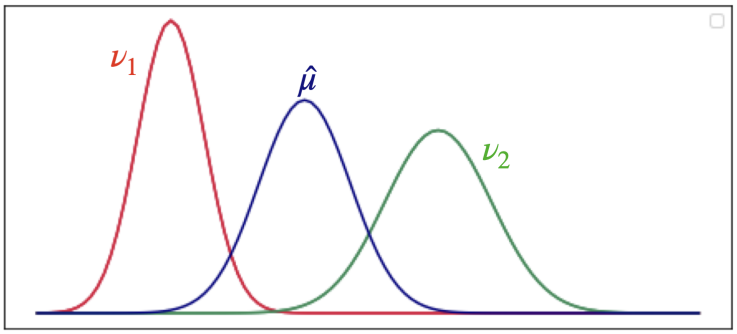

FIGURE 2.7: The Wasserstein barycenter $\hat{\mu}$ of two one-dimensional probability distributions ν_1 and ν_2.

2.9.1 Special cases

In the case where $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = ||x - y||^2$, (Agueh and Carlier, 2011) show that the problem of Wasserstein barycenter is convex and if one of the input measures has a density, then this barycenter is unique.

Also in the specific case, where $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = ||x - y||^2$, the barycenter of Gaussian distributions $\nu_i = \mathcal{N}(m_i, \Sigma_i)$, is itself a Gaussian $\mathcal{N}(m^*, \Sigma^*)$, where $m^* = \sum_{i=1}^n \lambda_i m_i$ and Σ^* is the minimizer of $\Sigma \mapsto \sum_{i=1}^n \lambda_i \mathcal{B}(\Sigma, \Sigma_i)^2$, where $\mathcal{B}$ is the Bure metric.

For one-dimensional distributions, the Wasserstein barycenter can be computed almost in closed form. The simplest case is for empirical measures with m points, i.e. $\mu_i = \frac{1}{m}\sum_{j=1}^m \delta_{y_{i,j}}$, where the points are assumed to be sorted $y_{i,1} \leq y_{i,2} \leq ... \leq y_{i,m}$. Then, the barycenter μ_λ is also an empirical measure on m points: $\mu_\lambda = \frac{1}{m}\sum_{j=1}^m \delta_{x_{\lambda,j}}$, where $x_{\lambda,j} = A_\lambda(x_{i,j})_j$ with A_λ is the barycentric map defined by $A_\lambda(x_i)_i = \underset{x\in\mathbb{R}}{\text{argmin}} \sum_{i=1}^n \lambda_i |x - x_i|^p$.

The last special case we present is when $n = 2$ where $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = ||x-y||^2$. This setting corresponds to the so-called McCann interpolant (McCann, 1997), where one wants to find:

$$\inf_{\mu\in\mathcal{P}_p(\mathbb{R}^d)} (1-t)\mathcal{W}_2^2(\mu, \nu_1) + t\mathcal{W}_2^2(\mu, \nu_2), \tag{2.35}$$

with $t \in [0, 1]$ and ν_1 is regular with respect to Lebesgue measure. Using Brenier's theorem we know that there exists a unique push-forward such that $\mathrm{T}\#\nu_1 = \nu_2$, in this case the barycenter is unique and obtained with $\mu_t = ((1-t)Id + t\mathrm{T})\#\nu_1$. In practice, when the probability measures ν_1, ν_2 are discrete with respectively n and m atoms this interpolant can be computed by $\mu_t = \sum_{i=1}^n \sum_{j=1}^m \gamma^*_{i,j}\delta_{(1-t)x_i+ty_j}$ where γ^* is an optimal transport plan between ν_1, ν_2.

2.9.2 Numerical scheme of Wasserstein barycenter

Despite the special cases above, it is difficult in practice to find a solution of the Wasserstein barycenter problem in the general framework. In what follows, we detail a solution for the scenario where the input measures are discrete. More formally let $(\nu_i)_{i\in[\![1,n]\!]}$ be discrete probability measures with weights $b_i \in \Sigma_{n_i}$ and that are supported on $Y_i = (y_q^i)_{q\in[\![1,n_i]\!]} \in \mathcal{M}_{n_i\times d}(\mathbb{R})$, for each $i \in [\![1, n]\!]$. Instead of looking at

all possible discrete probability measures, we can search a k atoms probability measure i.e. of the form $\hat{\mu} = \sum_{j=1}^{k} a_j \delta_{x_j}$ where $X = (x_j)_{j \in [\![1,k]\!]} \in \mathcal{M}_{k \times d}(\mathbb{R})$ and $a \in \Sigma_k$. Overall the resulting problem is:

$$\min_{\substack{a \in \Sigma_k, X \in \mathcal{M}_{k \times d}(\mathbb{R}) \\ \forall i \in [\![1,n]\!], \gamma_i \in U(a, b_i)}} \sum_{i=1}^{n} \lambda_i \langle \gamma_i, C_{XY_i} \rangle_F, \tag{2.36}$$

where $C_{XY_i} \in \mathcal{M}_{k \times n_i}(\mathbb{R}^+)$ is the matrix defined by all pair to pair costs between the points of the barycenter $\hat{\mu}$ and ν_i, i.e. $C_{XY_i} = (c(x_j, y_q^i))_{j,q \in [\![1,k]\!] \times [\![1,n_i]\!]}$.

In (Cuturi and Doucet, 2014) authors proposed to solve the problem (2.36) using Block Coordinate Descent (BCD) that alternates between minimizing with respect to a, X and γ_i while keeping others fixed:

1. The minimization with respect to all γ_i with a, X fixed involves solving n discrete optimal transport problems.

2. The minimization with respect to X with a, γ_i fixed can be performed in closed-form in the case $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = ||x - y||_2^2$:

$$X = Diag\left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} \lambda_i \gamma_i Y_i\right) \tag{2.37}$$

3. The minimization with respect to the weight a with X, γ_i fixed relies on the optimal dual variables of all optimal transport sub-problems of step (1) and applies a projected subgradient minimization with respect to a.

These three steps are repeated until convergence of X and a. The major drawback of this approach is its computational complexity which is driven by the calculation of many optimal transport problems. When the support X is fixed, the problem reduces to:

$$\min_{\substack{a \in \Sigma_k \\ \forall i \in [\![1,n]\!], \gamma_i \in U(a, b_i)}} \sum_{i=1}^{n} \lambda_i \langle \gamma_i, C_{XY_i} \rangle_F \tag{2.38}$$

We can use regularized optimal transport to obtain fast and smooth approximations of the original barycenter problem as given by:

$$\sum_{i=1}^{n} \lambda_i \langle \gamma_i, C_{XY_i} \rangle_F + \varepsilon H(\gamma | a \otimes b_i) \tag{2.39}$$

The resulting problem is a smooth convex minimization problem, which can be tackled using gradient descent (Cuturi and Doucet, 2014).

2.10 Optimal transport extensions

In this section, we discuss briefly other variants of optimal transport. The first one is called Sliced Wasserstein distance, which is an alternative optimal transport distance obtained by computing infinitely many linear projections of the high-dimensional distribution to one-dimensional distributions and then computing the average of the Wasserstein distance between these one-dimensional representations. The second

variant is the Gromov-Wasserstein distance, which allows comparing measures in different metric spaces. The third variant is the unbalanced optimal transport that allows the comparison of probability distributions that do not share the same mass.

2.10.1 Sliced Wasserstein distance

Besides entropic regularization of optimal transport, there are many other methods of approximating the optimal transport plan. One of them is based on the closed-form expression of optimal transport for probability distributions over the real line resulting on the so-called Sliced Wasserstein distance ($\mathcal{SW}$) (Rabin et al., 2011). Considering $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the main insight is to average the Wasserstein distance between projections on sampled one-dimensional directions. Specifically:

Definition 2.39 (Sliced Wasserstein Distance) *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and let $p \in [1, \infty[$. The p-Sliced Wasserstein $\mathcal{SW}_p$ is given by:*

$$\mathcal{SW}_p^p(\mu, \nu) = \int_{\mathrm{S}^d} \mathcal{W}_p^p(P_\theta \# \mu, P_\theta \# \nu) d\theta \tag{2.40}$$

where $\mathrm{S}^d = \{\theta \in \mathbb{R}^d : ||\theta|| = 1\}$ is the d-dimensional sphere, and $P_\theta : x \in \mathbb{R}^d \to \mathbb{R}$ is the projection.

$\mathcal{SW}$ enjoys several interesting properties. First $\mathcal{SW}_2$ induces a similar topology than $\mathcal{W}_2$: it defines a distance on $\mathcal{P}_p(\mathbb{R}^d)$ (Bonnotte, 2013) that metrizes the weak convergence (Nadjahi et al., 2019) and which is equivalent to the Wasserstein distance for measures with compact supports (Nadjahi et al., 2020). Secondly, from a practical side, the overall complexity of computing $\mathcal{SW}$ is $\mathcal{O}(n \log(n))$.

2.10.2 Gromov–Wasserstein distance

The Wasserstein distance provides an efficient way to compare probability measures when a distance is defined between their supports. Unfortunately, in the case where the measures are supported on samples living in different metric spaces, the definition of a meaningful ground distance is not straightforward, thus, the Wasserstein distance can no longer be defined. In this section, we present the Gromov-Wasserstein ($\mathcal{GW}$) distance (Mémoli, 2011), which can compare measures lying in incomparable metric spaces by comparing intra-domain distances.

Before defining Gromov-Wasserstein distance, we need the following definitions from (Sturm, 2012):

Definition 2.40 (Metric measure space) *A metric measure space (mm-space) is a triple $(\mathcal{X}, d_\mathcal{X}, \mu)$ where $(\mathcal{X}, d_\mathcal{X})$ is a polish metric space and μ is a Borel probability measure in $\mathcal{X}$.*

Definition 2.41 (Isometric metric measure spaces) *Two metric measure spaces $(\mathcal{X}, d_\mathcal{X}, \mu)$ and $(\mathcal{Y}, d_\mathcal{Y}, \nu)$ are called isometric if there exists a bijection $\varphi : \mathcal{X} \to \mathcal{Y}$ such that $\varphi \# \mu = \nu$ and $d_\mathcal{Y}(\varphi(x), \varphi(x')) = d_\mathcal{X}(x, x')$.*

Definition 2.42 (Gromov-Wasserstein distance) *Let $(\mathcal{X}, d_\mathcal{X}, \mu)$, $(\mathcal{Y}, d_\mathcal{Y}, \nu)$ be two measure metric spaces and let $p \in [1, \infty[$, one defines:*

$$\mathcal{GW}_p((\mu, d_\mathcal{X}), (\nu, d_\mathcal{Y})) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_\mathcal{X}(x, x') - d_\mathcal{Y}(y, y')|^p \, d\gamma(x, y) d\gamma(x', y') \right)^{1/p}. \tag{2.41}$$

$\mathcal{GW}_p$ defines a distance between metric measure spaces up to isometries, called the Gromov-Wasserstein distance of order p.

Example 2.43 (Gromov-Wasserstein distance between discrete measures) *Let consider $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ two discrete probability measures over $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ respectively. Let $p \in [1, \infty[$ and let C_1, C_2 be the matrices of pair-to-pair distances inside each space, i.e. $\forall i, i' \in [\![1, n]\!], C_1(i, i') = d_\mathcal{X}(x_i, x_{i'})$ and $\forall j, j' \in [\![1, m]\!], C_2(j, j') = d_\mathcal{Y}(y_j, y_{j'})$. Then the p-Gromov-Wasserstein distance is given by:*

$$\mathcal{GW}_p((\mu, C_1), (\nu, C_2)) = \left(\inf_{\gamma \in U(a,b)} \sum_{i,j,i',j'} |C_1(i, i') - C_2(j, j')|^p \gamma_{i,j} \gamma_{i',j'} \right)^{1/p}. \quad (2.42)$$

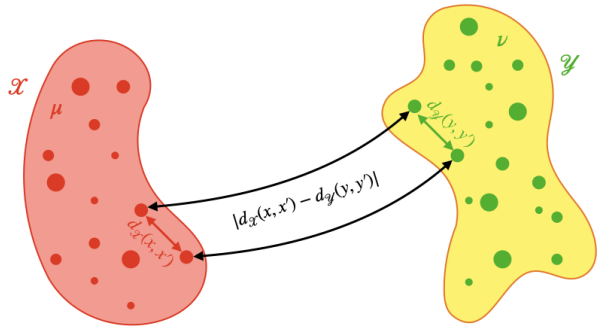

FIGURE 2.8: Gromov-Wasserstein distance allows to compare two different metric measure spaces. The resulting coupling tends to associate pairs of points with similar distances within each pair: the more similar $d_\mathcal{X}(x_i, x_{i'})$ is to $d_\mathcal{Y}(y_j, y_{j'})$, the stronger the transport coefficients $\gamma_{i,j}$ and $\gamma_{i',j'}$ are.

The Gromov-Wasserstein distance in (2.42) is challenging to compute, as it requires solving a nonconvex quadratic program that is NP-hard (Peyré et al., 2016). An entropic regularization has been proposed to reduce this computational burden of Gromov-Wasserstein (Peyré et al., 2016), and a Sinkhorn-like algorithm can also be adapted to the entropy-regularized Gromov-Wasserstein problem. More recently, a sliced variant has been introduced in (Titouan et al., 2019) in the case of the specific squared Euclidean ground cost.

2.10.3 Unbalanced optimal transport

Due to the mass conservation constraint, optimal transport can not compare measures with different total masses. Unbalanced optimal transport is a generalization that relaxes the conservation of mass constraints by replacing them with penalties.

First, let us give a preliminary definition (Csiszár, 1975) necessary for the statement of unbalanced optimal transport formulation:

Definition 2.44 (φ-divergence) *Let φ be a convex, lower semi-continuous function such that $\varphi(1) = 0$. The φ-divergence between probability measures μ and ν is defined by:*

$$\mathcal{D}_\varphi(\mu|\nu) = \int_\mathcal{X} \varphi\left(\frac{d\mu}{d\nu}(x)\right) d\nu(x) + \varphi_\infty \mu^\perp(\mathcal{X}), \quad (2.43)$$

where $\varphi_\infty = \lim_{x \to +\infty} \frac{\varphi(x)}{x}$ *and* $\mu^\perp(\mathcal{X})$ *denotes the mass of the part of* μ *that is not absolutely continuous with respect to* ν *in the Lebesgue's decomposition, i.e.* $\mu = \frac{d\mu}{d\nu}(x)\nu + \mu^\perp$.

Example 2.45 (Examples of φ-divergence) *Kullback-Leibler* KL*, Jensen-Shannon* JS*, Total Variation* TV*, and Hellinger* H^2 *are examples of* φ*-divergence.*

Definition 2.46 (Unbalanced optimal transport) *Let* $\mu \in \mathcal{M}_+(\mathcal{X})$ *and* $\nu \in \mathcal{M}_+(\mathcal{Y})$ *be two positive measures. Unbalanced Optimal Transport is defined as the following minimization problem*

$$(\mathcal{UOT}) \quad \inf_{\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) + \tau\mathcal{D}_\varphi(proj_{\mathcal{X}}\#\gamma|\mu) + \tau\mathcal{D}_\varphi(proj_{\mathcal{Y}}\#\gamma|\nu) \quad (2.44)$$

where τ *is the marginal penalization that controls how much mass variations are penalized as opposed to transportation of the mass.*

Note that there is no constraint on the transport plan besides positivity: it is not required to have marginals equal to μ and ν nor to have mass 1.

Regarding the metric properties of unbalanced optimal transport, we have the following theorem:

Theorem 2.47 *Consider the square Euclidean ground cost* $c(x,y) = ||x-y||^2$ *and Kullback-Leibler divergence* KL*, then the unbalanced optimal transport cost is the Gaussian-Hellinger distance, which is a distance on* $\mathcal{P}(\mathbb{R}^d)$.

Finally, an unbalanced variant of the Gromov-Wasserstein distance was proposed and studied in (Séjourné et al., 2021), and it is possible to define an entropic extension of unbalanced optimal transport that is computable via a generalized Sinkhorn's algorithm (Chizat et al., 2018).

2.11 Optimal transport toolboxes

It is a custom for optimal transport papers to be accompanied by their open-source solvers. However, most of them are either outdated or not currently maintained. Thankfully, there are many good open-source optimal transport toolboxes, such as `OTT`, `OTJulia`, `GeomLoss` and `OTT-JAX`. But, the first reference remains undoubtedly `POT` (Flamary et al., 2021), which is an optimal transport toolbox written in Python, that has contributed considerably to the democratization of optimal transport in the machine learning community.

Bibliographical notes

The impressive book of the Fields Medalist Villani is definitely the Bible of optimal transport theory. For a general introduction to this theory with a particular focus on connections with different fields of applied mathematics, Santambrogio's book remains the cornerstone. While the most complete reference about computational aspects of optimal transport is the excellent book by Peyré and Cuturi. The recent book by the Fields Medalist Figalli and Glaudo provides on its part a gentle introduction to this theory and can serve as a starting point for exploring the beautiful world of optimal transport.

CHAPTER 3

DOMAIN ADAPTATION

Contents

We have on this earth what makes life worth living:
April's recurrence,
the aroma of bread at dawn,
a woman's opinion of men,
the writings of Aeschylus,
love's beginnings,
grass on a stone,
mothers standing on a flute's thread
and the invader's fear of memories.

Mahmoud Darwich

3.1 Domain adaptation: simulating human brain flexibility to environments change

If an image classifier was trained on photo images, would it work on sketch images? Can an epileptic seizure detector trained using one patient's electroencephalogram data diagnose another patient's brain activity? What if a vehicle detector trained using daytime images is tested in nighttime images? Is it possible to deploy a semantic segmentation model trained using postmortem imaging during intra-operative surgery? Answers to these questions lie on the ability of machine learning models to deal with the common problem of distributional shift between training and test data.

Most supervised learning algorithms strongly rely on an over-simplified assumption, that is, the training and test data are independent and identically distributed (i.i.d.), while distributional shift scenarios are commonly encountered in practice. As a consequence, a traditional learning model will typically suffer significant performance drops on a test sample drawn from a different distribution than the training sample. This performance drop does not even spare deep learning models that have been seriously hampered in front of this shift as revealed by the studies conducted in (Hendrycks and Dietterich, 2018; Recht et al., 2019).

Statistical learning theory, for its part, guarantees that a model's empirical risk is close to its true risk under the standard assumption that the training and test data are drawn independently from the same probability distribution, but this prolific theoretical machinery stands on the sidelines when this assumption is violated, and manifestly fails to establish generalization bounds of a learning model built in a setting that does not preserve the i.i.d. assumption.

This dual challenge, both theoretical and algorithmic, has promoted the rise of domain adaptation, a new sub-field of statistical learning theory that takes into account the shift between training and test data distributions.

This chapter is dedicated to the presentation of theoretical and algorithmic advances in domain adaptation. For the sake of clarity, we start by defining notions that are needed to conceptualize the supervised learning problem and then we investigate it theoretically through statistical learning theory. Subsequently, we formally define the problem of domain adaptation and we further cover in an exhaustive manner its theoretical guarantees. Thereafter, we exhibit the most relevant domain adaptation algorithms that were proposed in the literature.

3.2 Statistical learning theory

Supervised learning is arguably the most widespread task of machine learning and has enjoyed much success on a broad spectrum of application domains (Kotsiantis et al., 2007). In this section, we introduce the usual supervised learning setting and the different theoretical guarantees based on the concepts of Vapnik-Chervonenkis dimension (Vapnik, 2006; Vapnik and Chervonenkis, 2015) and Rademacher complexity (Koltchinskii and Panchenko, 2000), those from the more recent algorithmic stability (Bousquet and Elisseeff, 2002) and algorithmic robustness (Xu and Mannor, 2012) frameworks and finally the generalization bounds related to the PAC-Bayesian theory (McAllester, 1998).

3.2.1 Preliminary definitions

We denote by $\mathcal{X}$ the input space, which is the set of all possible samples, and by $\mathcal{Y}$ the label space composed of all possible output values. In the remainder of the manuscript, we will focus on the most popular setting $\mathcal{X} \subset \mathbb{R}^{\mathrm{d}}$, where d is the number of features describing each sample. As for the output values, we will consider the classification task [1], where $\mathcal{Y}$ is a discrete finite set $\mathcal{Y} = \{C_1, ..., C_k\}$ and $k \geq 2$ is the number of candidate classes.

We assume that samples are independently and identically distributed (i.i.d.) according to some fixed but unknown joint probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. In practice, the joint distribution $\mathcal{D}$ is observed only through a finite training set $S = \{(x_i, y_i)\}_{i=1}^{m} \sim (\mathcal{D})^m$ composed of m samples drawn i.i.d. from $\mathcal{D}$.

We further use $\mathcal{H} = \{h \mid h : \mathcal{X} \to \mathcal{Y}\}$ to denote a predetermined possible infinite set called hypothesis space that consists of functions h representing a possible deterministic rule of how the output values are generated from the input observations. These functions h are usually called hypothesis, or more specifically classifiers or regressors, depending on the nature of $\mathcal{Y}$.

To evaluate the performance of a given hypothesis h, the conventional approach is to use a function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ quantifying the disagreement between the value of $h(x)$ and the observed output y. In other words, this function models the loss sustained by h when predicting the value of y as $h(x)$, hence the name loss function for l.

The most natural loss function is the $0-1$ loss, $l_{0-1} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$, which is defined for a training example (x, y) as:

$$l_{0-1}(h(x), y) = \begin{cases} 1, \text{ if } h(x) \neq y\,, \\ 0, \text{ otherwise.} \end{cases} \tag{3.1}$$

However, the minimization of the $0-1$ loss function is an NP-hard problem (Arora et al., 1997) and thus other surrogate loss functions should be considered to approximate it. The best proxy to this discontinuous nonconvex function is the hinge loss (Ben-David et al., 2012) defined for a given pair (x, y) by:

$$l_{\text{hinge}}(h(x), y) = \max\left(0, 1 - yh(x)\right). \tag{3.2}$$

[1] For regression, $\mathcal{Y}$ is a continuously infinite subset of $\mathbb{R}^k$, where $k \in \mathbb{N}^*$ is the dimension of the output space.

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

$$l_{\text{lin}}(h(x), y) = \frac{1}{2}\left(1 - yh(x)\right). \tag{3.3}$$

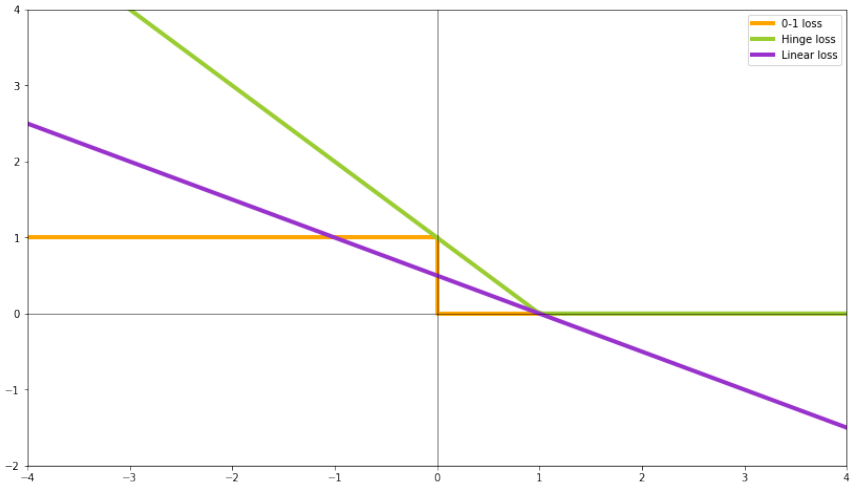

FIGURE 3.1: Illustration of the $0-1$ loss, the hinge loss and the linear loss.

The aggregation of the losses of single samples to the entire dataset is generally undertaken by its expectation over the distribution $\mathcal{D}$, known as the true risk.

Definition 3.1 (True risk) *Given a hypothesis $h \in \mathcal{H}$, a joint probability distribution $\mathcal{D}$, and a loss function l, the true risk of h over $\mathcal{D}$ is:*

$$\epsilon_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} l(h(x), y). \tag{3.4}$$

The calculation of the true risk of a hypothesis $h \in \mathcal{H}$ is not possible since the generating distribution $\mathcal{D}$ is unknown. However, we can measure the empirical risk of a hypothesis on the training set S.

Definition 3.2 (Empirical risk) *Given a loss function l and a training set $S = \{(x_i, y_i)\}_{i=1}^m$, where each example is drawn i.i.d. from the joint distribution $\mathcal{D}$, the empirical risk of a given hypothesis $h \in \mathcal{H}$ is defined as:*

$$\epsilon_{\hat{\mathcal{D}}}(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i). \tag{3.5}$$

where $\hat{\mathcal{D}} = \frac{1}{m}\sum_{i=1}^{m} \delta_{(x_i,y_i)}$ is the empirical distribution associated to the training set S.

3.2.2 No-free lunch theorem

The i.i.d. assumption involves that each sample brings new information that is independent from other previously seen samples. This may lead to the belief in the existence of a universal learner returning a hypothesis that approaches the perfect hypothesis more and more as the sample size increases. However, this belief is wrong, since any learner can exhibit an arbitrary bad behavior on a set of finite size. This assertion is usually formalized by the "No-free lunch" theorem (Shalev-Shwartz and Ben-David, 2014) that can be stated as follows.

Theorem 3.3 (No-free lunch) *Let $\mathcal{A}$ be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a space $\mathcal{X}$. Let m be any number smaller than $\frac{|\mathcal{X}|}{2}$ representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *there exists a function $h : \mathcal{X} \to \{0,1\}$ with $\epsilon_{\mathcal{D}}(h) = 0$,*
2. *with probability of at least $\frac{1}{7}$ over the random choice of $S \sim (\mathcal{D})^m$, we have that $\epsilon_{\mathcal{D}}(\mathcal{A}(S)) \geq \frac{1}{8}$.*

As stated earlier, this theorem indicates that for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner. In other words, there is no universal learner that succeeds on all tasks. Therefore, each learner must be designed for a specific task using prior knowledge about that task in order to succeed, and such prior knowledge can be expressed by restricting the hypothesis space. We present here several strategies that can be employed to get a low risk hypothesis, by selecting a good hypothesis space and restricting this latter to have a reasonable complexity.

3.2.3 Risk minimizing strategies

A number of well-known strategies allowing to find a hypothesis with good generalization capacities and low risk, namely:

3.2.3.1 Empirical risk minimization

Following the law of large numbers, the empirical risk converges to the true risk when the number of available samples tends to infinity and thus provides a good proxy for the latter. Therefore, this proposes to look for a hypothesis minimizing the empirical risk:

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \epsilon_{\hat{\mathcal{D}}}(h). \tag{3.6}$$

and hope that $\mathcal{H}$ was constrained beforehand to a reasonable class of functions using some a priori knowledge about the learning problem.

Nevertheless, when this knowledge is not available, this strategy may be prone to a serious shortcoming related to overfitting, where the resulting hypothesis h can perfectly fit the observed set S while performing poorly on the underlying distribution $\mathcal{D}$. This phenomenon is reflected by a high empirical risk of h on a set $S' \neq S$ generated from $\mathcal{D}$ that was not employed for learning. The overfitting can be attributed to one of the two most usual factors:

- The set S may not be sufficiently representative of the unknown distribution $\mathcal{D}$, this may occur when S is not large enough or when the labels are noisy, and thus collecting a larger amount of samples helps to overcome this nuisance.
- Although S is large, overfitting may arise due to an excessive richness of $\mathcal{H}$, imploding that a slight change in the dataset due, for example, to altering a few learning samples, may change significantly the learned hypothesis. Consequently, the performance of h differs considerably for different samples drawn from $\mathcal{D}$, which suggests that the performance on the set S cannot be trusted as a gauge of the performance on the entire distribution. For this reason, overfitting is also known as a high variance or high complexity problem.

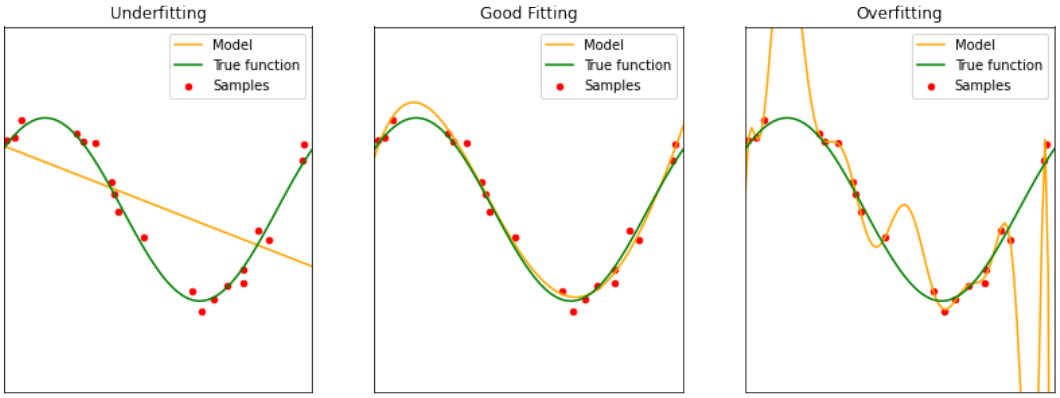

FIGURE 3.2: Illustration of overfitting, underfitting and good fitting. (left) the model is underfitted and does not properly capture the true behavior of the true distribution function. (middle) a good model that follows the true distribution of the samples. (right) the model is overfitted and tries to follow perfectly the points of the available samples.

To prevent overfitting, one needs to collect sufficient samples and somehow prevent the hypothesis under consideration from being excessively flexible. While the first point concerns the data gathering process itself, the second can be attained by enforcing some restrictions on the flexibility of the considered hypothesis space $\mathcal{H}$. This is the motivation behind the following two learning strategies.

3.2.3.2 Structural risk minimization

Originally introduced in (Vapnik, 1991), the structural risk minimization aims at minimizing the empirical risk while penalizing the structure of the considered hypothesis space based on its complexity.

Actually, as defining the hypothesis space beforehand using a priori knowledge is generally quite hard, one may rather consider a possibly infinite set of embedded hypothesis spaces

$$\{\mathcal{H}_i \mid \forall i \in I, \mathcal{H}_i \subset \mathcal{H}_{i+1}\}, \tag{3.7}$$

and a penalization $\text{pen}(.)$ applied to the hypothesis space $\mathcal{H}_i$. $\text{pen}(.)$ is an increasing function for set inclusion, which implies that $\text{pen}(\mathcal{H}_i) \leq \text{pen}(\mathcal{H}_{i+1})$.
More formally, the structural risk minimization consists in finding:

$$h^* = \underset{h \in \mathcal{H}_i, i \in I}{\text{argmin}}\, \epsilon_{\hat{\mathcal{D}}}(h) + \text{pen}(\mathcal{H}_i). \tag{3.8}$$

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 $\mathcal{H}_i$ and the empirical risk value.

3.2.3.3 Regularized risk minimization

This strategy introduces a trade-off between searching for a hypothesis with a low empirical risk and high generalization capacities by penalizing the hypothesis function itself via a regularizer $\text{reg}(.)$ which penalizes excessively flexible hypothesis:

$$h^* = \underset{h \in \mathcal{H}}{\text{argmin}}\, \epsilon_{\hat{\mathcal{D}}}(h) + \lambda.\text{reg}(h). \tag{3.9}$$

where λ is a positive parameter controlling the trade-off between the empirical risk minimization and the regularization strength. This approach is by far the most used risk minimizing strategy, at least when it comes to algorithmic implementations.

3.2.4 Generalization bounds

Statistical learning theory (Vapnik, 1999) yields conditions ensuring the convergence of the empirical risk to the true risk for a given hypothesis space. These results are referred to as generalization bounds, and they are commonly stated in the form of probably approximately correct (PAC) bounds (Valiant, 1984) that have the following form:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left\{|\epsilon_{\hat{\mathcal{D}}}(h) - \epsilon_{\mathcal{D}}(h)| \leq \varepsilon\right\} \geq 1 - \delta, \tag{3.10}$$

where $\varepsilon > 0$ and $\delta \in (0, 1]$. This statement basically indicates that we aim to upper-bound the gap between the true risk and its estimated counterpart by the least possible value of ε and with a high probability over the random choice of the learning set S. The major concern is then to understand whether $\epsilon_{\hat{\mathcal{D}}}(h)$ converges to $\epsilon^l_{\mathcal{D}}(h)$ with an increasing size of the training set, and what is the rate of this convergence. We now present a number of theoretical frameworks that have been advanced in the literature to demonstrate the various ingredients on which this speed may depend.

3.2.4.1 Vapnik-Chervonenkis bounds

Vapnik-Charvonenkis (VC) bounds (Vapnik, 1971) are based on the original quantification of the complexity of a given hypothesis space. This concept of complexity is captured by the famous VC dimension.

The VC dimension is a purely combinatorial notion based on the concepts of dichotomy and of shattering.

Definition 3.4 (Dichotomy) *Given a binary hypothesis space $\mathcal{H}$, a dichotomy of a subset $\mathcal{X}' \subset \mathcal{X}$ is one of the possible ways of labeling the samples of $\mathcal{X}'$ using a hypothesis in $\mathcal{H}$.*

Definition 3.5 (Growth function) *The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis space $\mathcal{H}$ is defined by:*

$$\forall m \in \mathbb{N},\ \Pi_{\mathcal{H}}(m) = \max_{\{x_1,\dots,x_m\}\subseteq\mathcal{X}} |\{(h(x_1), ..., h(x_m)) : h \in \mathcal{H}\}|. \tag{3.11}$$

Thus, $\Pi_{\mathcal{H}}(m)$ is the maximum number of dichotomies for m samples using hypothesis in $\mathcal{H}$.

Definition 3.6 (Shattering) *A set $\mathcal{X}'$ of $m \geq 1$ samples is said to be shattered by a binary hypothesis space $\mathcal{H}$ when $\mathcal{H}$ realizes all possible dichotomies of $\mathcal{X}'$, that is when $\Pi_{\mathcal{H}}(m) = 2^m$.*

Definition 3.7 (VC dimension) *The Vapnik-Chervonenkis (VC) dimension of a binary hypothesis space $\mathcal{H}$ is the cardinality of the largest subset $\mathcal{X}' \subset \mathcal{X}$ that can be labeled in all of the possible ways by hypothesis from $\mathcal{H}$. More formally, we have:*

$$\mathrm{VC}(\mathcal{H}) = \max_{\mathcal{X}'\subseteq\mathcal{X}} \{|\mathcal{X}'| : \Pi_{\mathcal{H}}(|\mathcal{X}'|) = 2^{|\mathcal{X}'|}\}. \tag{3.12}$$

The VC dimension is a measure of the richness of the hypothesis space $\mathcal{H}$ and captures from which cardinality $\mathcal{H}$ stops behaving like functions from $\mathcal{Y}^{\mathcal{X}}$, as these latter can label any finite set $\mathcal{X}' \subseteq \mathcal{X}$ in all of the possible $2^{|\mathcal{X}'|}$ ways.

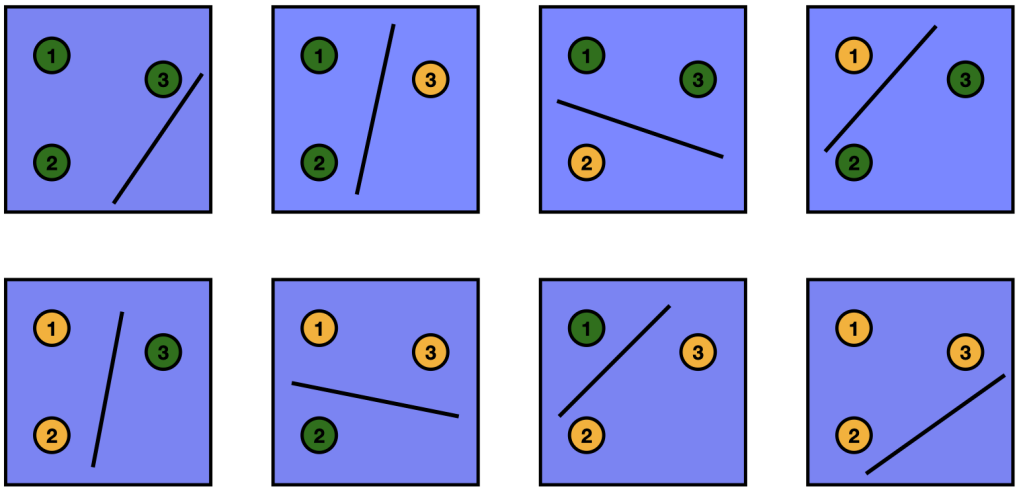

FIGURE 3.3: Illustration of the idea behind VC dimension. Here, half-planes in $\mathbb{R}^{\mathrm{d}}$ with $\mathrm{d} = 2$ can correctly classify at most three points for all possible labelings. The VC dimension is then 2 + 1.

The following theorem uses the VC dimension of a hypothesis space to upper-bound the gap between the true and the empirical risk for a given loss function.

Theorem 3.8 *Let $\mathcal{X}$ be an input space, $\mathcal{Y} = \{-1, +1\}$ the label space, and $\mathcal{D}$ their joint distribution. Let S be a finite sample of size m drawn i.i.d. from $\mathcal{D}$, and $\mathcal{H}$ be a hypothesis space of* VC *dimension* $\mathrm{VC}(\mathcal{H})$*. Then for any $\delta \in (0, 1]$ with probability of at least $1 - \delta$ over the random choice of the training sample $S \sim (\mathcal{D})^m$, the following holds:*

$$\forall h \in \mathcal{H}, \quad \epsilon_{\mathcal{D}}(h) \leq \epsilon_{\hat{\mathcal{D}}}(h) + \sqrt{\frac{4}{m}\left(\mathrm{VC}(\mathcal{H}) \ln \frac{2em}{\mathrm{VC}(\mathcal{H})} + \ln \frac{4}{\delta}\right)}. \tag{3.13}$$

When $\mathrm{VC}(\mathcal{H})$ is known, the right hand-side of this inequality can be calculated explicitly. In general, this bound indicates that for a certain confidence level given by $1 - \delta$, the empirical risk of a hypothesis approaches its real value when m increases and this convergence is even faster for $\mathcal{H}$ with low VC dimension.

3.2.4.2 Rademacher bounds

Intuitively, the Rademacher complexity quantifies the ability of a given hypothesis space to withstand noise that potentially could be present in the data. This, in turn, was shown to lead to more accurate bounds than those based on the VC dimension (Koltchinskii and Panchenko, 2000). To present the Rademacher bounds, we first provide a definition of a Rademacher variable.

Definition 3.9 (Rademacher variable) *A random variable κ is defined as:*

$$\kappa = \begin{cases} 1, \text{ with probability } \frac{1}{2}, \\ -1, \text{ otherwise.} \end{cases} \tag{3.14}$$

is called the Rademacher variable.

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

Definition 3.10 (Empirical Rademacher complexity) *For a given hypothesis space $\mathcal{H}$ and a given unlabeled set $S = \{(x_i)\}_{i=1}^m$, the empirical Rademacher complexity of $\mathcal{H}$ associated to S is defined as follows:*

$$\mathcal{R}_S(\mathcal{H}) = \mathop{\mathbb{E}}_{\kappa}\left[\sup_{h\in\mathcal{H}} \frac{2}{m}\sum_{i=1}^{m} \kappa_i h(x_i)\right], \tag{3.15}$$

where κ is a vector of m independent Rademacher variables.

For a set S, the empirical Rademacher complexity $\mathcal{R}_S(\mathcal{H})$ measures the ability of hypothesis from $\mathcal{H}$ to correlate with random noise defined by the Rademacher random variables. If the correlation is high, then the hypothesis are too flexible and may lead to overfitting.

Definition 3.11 (Rademacher complexity) *The Rademacher complexity of a give hypothesis space $\mathcal{H}$ associated to a sample size m is defined as the expected value of $\mathcal{R}_S(\mathcal{H})$:*

$$\mathcal{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_{S\sim(\mathcal{D})^m} \mathcal{R}_S(\mathcal{H}). \tag{3.16}$$

The following theorem presents the generalization bound based on Rademacher complexity (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2002).

Theorem 3.12 *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a finite set of m samples drawn i.i.d. from $\mathcal{D}$, and $\mathcal{H}$ be a binary hypothesis space. Then, for any $\delta \in (0, 1]$ with probability of at least $1 - \delta$ over the choice of the sample $S \sim (\mathcal{D})^m$, the following holds:*

$$\forall h \in \mathcal{H}, \quad \epsilon_{\mathcal{D}}(h) \leq \epsilon_{\hat{\mathcal{D}}}(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}. \tag{3.17}$$

3.2.4.3 Algorithmic stability bounds

The previous generalization bounds based on the VC dimension and Rademacher complexity ignore the specific algorithm used, that is, they hold for any algorithm using $\mathcal{H}$ as a hypothesis space [2]. Thus, one may ask if an analysis of the properties of a specific algorithm could lead to finer guarantees. (Bousquet and Elisseeff, 2002) introduced generalization bounds that provide an answer to this question based on the concept of uniform stability of a learning algorithm. We now give its definition.

Definition 3.13 (Uniform stability) *An algorithm $\mathcal{A}$ has uniform stability β with respect to the loss function l if the following holds:*

$$\forall S \in \{\mathcal{X}\times\mathcal{Y}\}^m, \forall i \in \{1,\ldots,m\}, \sup_{(x,y)\in S} |l(h_S(x), y) - l(h_{S^{\setminus i}}(x), y)| \leq \beta, \tag{3.18}$$

where $S^{\setminus i} = S \setminus \{(x_i, y_i)\}$ and h_S and $h_{S^{\setminus i}}$ are learned by $\mathcal{A}$ from S and $S^{\setminus i}$ respectively.

The insight underlying this definition is to say that an algorithm that is supposed to generalize well should be stable against small perturbations in the training sample. Therefore, stable algorithms should have an empirical risk that stays close to their true risk. This idea is confirmed by the following theorem.

[2] By algorithm, we mean any rule that takes a sample $S \sim (\mathcal{D})^m$ and a hypothesis space $\mathcal{H}$ and outputs a hypothesis h.

Theorem 3.14 *Let $\mathcal{A}$ be an algorithm with uniform stability β with respect to a loss function l, such that $0 \leq l(h_S(x), y) \leq M$, for all $(x, y) \in (\mathcal{X} \times \mathcal{Y})$ and all sets S. Then, for any $m \geq 1$, and any $\delta \in (0, 1]$, the following bound holds with probability of at least $1 - \delta$ over the random choice of the sample S:*

$$\epsilon_{\mathcal{D}}(h) \leq \epsilon_{\hat{\mathcal{D}}}(h) + 2\beta + (4m\beta + M)\sqrt{\frac{\ln \frac{1}{\delta}}{2m}}. \tag{3.19}$$

This theorem says that an algorithm with uniform stability β generalizes well when β scales as $\frac{1}{m}$. We further note that several well-known machine learning algorithms were shown to verify the uniform stability property.

3.2.4.4 Algorithmic robustness bounds

The principal insight of algorithmic robustness (Xu and Mannor, 2012) is that a robust algorithm must exhibit similar performance in terms of the classification error for the test and training samples that are close. The similarity measure used to define if two samples are close or not is based on partitioning the joint space $\mathcal{X} \times \mathcal{Y}$ in a way that locates two similar samples of the same class in the same partition. This partition is further defined using the concept of covering numbers (Kolmogorov and Tikhomirov, 1959), as introduced below.

Definition 3.15 (Covering number) *Let (Z, ϱ) denote a metric space. For $Z' \subset Z$, we say that $\hat{Z}'$ is a ξ-covering of Z', if for any element $t \in Z'$ there is an element $\hat{t} \in \hat{Z}'$ such that $\varrho(t, \hat{t}) \leq \gamma$. Then the number of ξ-covering of Z' is expressed as:*

$$N(\xi, Z', \varrho) = \min\left\{|\hat{Z}'| : \hat{Z}' \text{ is a } \xi\text{-covering of } Z'\right\}. \tag{3.20}$$

In the case where $\mathcal{X}$ is a compact space, its covering number $N(\xi, \mathcal{X}, \varrho)$ is finite. Moreover, for the product space $\mathcal{X} \times \mathcal{Y}$, the number of ξ-covering is also finite and is equal to $|\mathcal{Y}|N(\xi, \mathcal{X}, \varrho)$. As explained earlier, the above partitioning guarantees that two samples from the same subset belong to the same class and are close to each other with respect to the metric ϱ. Keeping this in mind, algorithmic robustness is defined as follows.

Definition 3.16 (Algorithmic robustness) *Let S be a training sample of size m where each example is drawn from the joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. An algorithm $\mathcal{A}$ is said to be $(M, \epsilon(\cdot))$-robust on $\mathcal{D}$ with respect to a loss function l for $M \in \mathbb{N}$ and $\epsilon(\cdot) : (\mathcal{X} \times \mathcal{Y})^m \to \mathbb{R}$ if $\mathcal{X} \times \mathcal{Y}$ can be partitioned into M disjoint subsets denoted by $\{\mathcal{Z}_k\}_{k=1}^{M}$, so that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, (x', y') drawn from $\mathcal{D}$ and $k \in \{1, \ldots, M\}$ we have:*

$$\big((x, y), (x', y')\big) \in \mathcal{Z}_k^2 \quad \Longrightarrow \quad |l(h_S(x), y) - l(h_S(x'), y')| \leq \epsilon(S)\,, \tag{3.21}$$

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

The algorithmic robustness focuses on measuring the divergence between the costs associated to two similar points, assuming that the learned hypothesis function should be locally consistent. We are now ready to present the generalization guarantees that characterize robust algorithms that verify the definition presented above.

Theorem 3.17 *Let S be a finite set of size m drawn i.i.d from $\mathcal{D}$, $\mathcal{A}$ be $(M, \epsilon(\cdot))$-robust on $\mathcal{D}$ with respect to a loss function l, such that $0 \leq l(h_S(x), y) \leq M_l$, for all $(x, y) \in (\mathcal{X} \times \mathcal{Y})$. Then, for any $\delta \in (0, 1]$, the following bound holds with probability of at least $1 - \delta$ over the random draw of the sample $S \sim (\mathcal{D})^m$:*

$$\epsilon_{\mathcal{D}}(h) \leq \epsilon_{\hat{\mathcal{D}}}(h) + \epsilon(S) + M_l \sqrt{\frac{2M \ln 2 + 2 \ln \frac{1}{\delta}}{m}}, \tag{3.22}$$

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

As $\epsilon(S)$ is dependent on the ξ-covering of $\mathcal{X} \times \mathcal{Y}$ and its size M, it naturally involves a trade-off between M and $\epsilon(S)$. Similarly to the bounds based on uniform stability, the bound of this theorem does not depend on the complexity of the hypothesis space and thus its right hand-side can be calculated even if the latter is not computable or infinite.

3.2.4.5 PAC-Bayesian bounds

The PAC-Bayesian paradigm (Shawe-Taylor and Williamson, 1997; McAllester, 1998) procures generalization bounds for a hypothesis cast as a weighted majority vote on the hypothesis space $\mathcal{H}$, as, for example, in the ensemble methods (Dietterich, 2000). In this section, we present the PAC-Bayesian generalization bounds as introduced in (Germain et al., 2015) in the binary classification setting, where $\mathcal{Y} = \{-1, 1\}$ with the $0 - 1$ loss or the linear loss. To derive such a generalization bound, we assume the existence of a prior distribution π over $\mathcal{H}$, which models an *a-priori* belief on the hypothesis of $\mathcal{H}$ before the observation of the learning sample $S \sim (\mathcal{D})^m$. Given S, the learner seeks to find a posterior distribution ρ on $\mathcal{H}$ that leads to a well-performing ρ-weighted majority vote $B_\rho(x)$ (called the Bayes classifier), defined as:

$$B_\rho(x) = \text{sign}\left[\mathop{\mathbb{E}}_{h \sim \rho} h(x)\right]. \tag{3.23}$$

Namely, instead of finding the best hypothesis from $\mathcal{H}$, we aim to learn ρ over $\mathcal{H}$, such that this minimizes the true risk $\epsilon_{\mathcal{D}}(B_\rho)$ of the ρ-weighted majority vote. Nevertheless, the PAC-Bayesian generalization bounds do not concern directly the risk of the deterministic ρ-weighted majority vote B_ρ, but give an upper bound on the expectation over ρ of all of the individual hypothesis true risks, called the Gibbs classifier, which draws a hypothesis h from $\mathcal{H}$ according to the posterior distribution ρ, and predicts the label of x given by $h(x)$. An important behavior of the Gibbs risk is that is tightly linked to the deterministic ρ-weighted majority vote. In fact, if B_ρ miss-classifies $x \in \mathcal{X}$, then at least half of the classifiers (under the ρ measure) make a prediction error on x. Consequently, we have:

$$\epsilon_{\mathcal{D}}(B_\rho) \leq 2 \mathop{\mathbb{E}}_{h \sim \rho} \epsilon_{\mathcal{D}}(h). \tag{3.24}$$

Thus, an upper bound on $\mathop{\mathbb{E}}_{h \sim \rho} \epsilon_{\mathcal{D}}(h)$ provides an upper bound on $\epsilon_{\mathcal{D}}(B_\rho)$ as well. Note that the PAC-Bayesian generalization bounds do not directly consider the complexity of the hypothesis space $\mathcal{H}$, unlike the Rademacher complexity or the VC dimension, but they measure the discrepancy between the prior distribution π and the

posterior distribution ρ on $\mathcal{H}$ through the Kullback-Leibler divergence:

$$\mathrm{KL}\left(\rho|\pi\right) = \mathop{\mathbb{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}. \tag{3.25}$$

Below, we present a generalization bound for the $0-1$ loss function due to (Catoni, 2007), involving the Kullback-Leibler divergence.

Theorem 3.18 *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis space $\mathcal{H}$, for any prior distribution π on $\mathcal{H}$, for any $\delta\in(0,1]$ and any real number $\omega>0$, with a probability of at least $1-\delta$ over the random choice of $S\sim(\mathcal{D})^m$, we have, for all posterior distribution ρ on $\mathcal{H}$:*

$$\mathop{\mathbb{E}}_{h\sim\rho}\epsilon_{\mathcal{D}}(h) \leq \frac{\omega}{1-e^{-\omega}}\left[\mathop{\mathbb{E}}_{h\sim\rho}\epsilon_{\mathcal{S}}(h) + \frac{\mathrm{KL}\left(\rho|\pi\right)+\ln\frac{1}{\delta}}{m\omega}\right]. \tag{3.26}$$

In addition to the fact that this bound concerns the ρ-expectation of the risk over $\mathcal{H}$, the parameter ω reflects a trade-off between the ρ-expected empirical risk $\mathbb{E}_{h\sim\rho}\epsilon_{\mathcal{S}}(h)$ and the divergence term $\mathrm{KL}\left(\rho|\pi\right)$.

3.3 Domain adaptation

Most supervised learning algorithms and theoretical foundations are built on the crucial assumption that training and test data are drawn from the same probability distribution (Pan and Yang, 2009), while in real-world applications, the data-generating process is often subject to change due to several application-dependent reasons, as illustrated by the following examples:

- **Visual recognition:** For brain tumor classification, the training distribution can differ from the test one, this may arise for example when tumors with different grades are likely to exhibit different characteristics due to varying degrees of tumor severity and growth patterns. In addition, in cross-center collaborations, data acquired even with the same vendor and with the same acquisition protocol can be substantially different from one another.
- **Fraud detection:** The heterogeneous nature of the fraudster behavior can lead to a change in the distributions of training and test data. This behavior may strongly differ according to the payment system (e.g. e-commerce or shop terminal), the country, and the population segment.
- **Sentiment analysis:** For product review classification, the drift observed in the distributions of training and test data is caused by the difference in product category and the change in word frequencies.

In the three scenarios above, it would not be an overstatement to consider the same distribution assumption for the training and test samples as unrealistic. Indeed, changing the measurement instruments, the data acquisition environment, or even the sampling method induce a shift between the joint distributions of training and test data. This distributional shift will be likely to degrade significantly the generalization ability of supervised learning models. While manual labeling may appear like a feasible solution, such an approach is unreasonable in practice, since it is often prohibitively expensive to collect from scratch a new large high quality labeled dataset with the same distribution as the test data, due to lack of time, resources, or

other factors, and it would be an immense waste to totally reject the available knowledge on a different, yet related labeled training set. Such a challenging situation has promoted the emergence of domain adaptation (Redko et al., 2019b), a sub-field of statistical learning theory (Vapnik, 1999), that takes into account the shift between the distributions of training and test data, respectively called source and target domains. Domain adaptation suggests that making use of the relatedness between the source and target domains in order to transfer knowledge acquired on the former to the latter, seems to be a much more intelligent solution when compared to learning each task from scratch.

3.3.1 Formal definition

Transferring the knowledge extracted from a source domain to serve in a target one lies at the heart of transfer learning, which is a broader field encompassing domain adaptation. For the sake of clarity and homogeneity, we adopt the formalization of domain adaptation given by (Kouw and Loog, 2019) [3], stated in the following way:

Definition 3.19 (Domain Adaptation) *Let $\mathcal{S}$ and $\mathcal{T}$ be two different joint probability distributions over $\mathcal{X} \times \mathcal{Y}$ called respectively the source and target domains. We have access to a set $S = \{(x_i, y_i)\}_{i=1}^{n}$ of n labeled source samples drawn i.i.d. from the joint distribution $\mathcal{S}$ and a set $T = \{x_j\}_{j=1}^{m}$ of m unlabeled target samples drawn i.i.d. from the marginal distribution $\mu_{\mathcal{T}}$, of the joint distribution $\mathcal{T}$ over $\mathcal{X}$, more formally:*

$$S = \{(x_i, y_i)\}_{i=1}^{n} \sim (\mathcal{S})^n, \quad T = \{x_j\}_{j=1}^{m} \sim (\mu_{\mathcal{T}})^m. \tag{3.27}$$

The aim of unsupervised domain adaptation is to learn from a given hypothesis space $\mathcal{H}$ a hypothesis h with a low target risk:

$$\epsilon_{\mathcal{T}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{T}} l(h(x), y), \tag{3.28}$$

under the distributional shift assumption $\mathcal{S} \neq \mathcal{T}$.

The considered setting in the definition above corresponds to **single-source homogeneous closed set unsupervised domain adaptation**, characterized as follows:

- **Single-source domain adaptation**, where the labeled data are available from only one source domain. In contrast to **multi-source domain adaptation**, where labeled data are available from several source domains.
- **Homogeneous domain adaptation**, where the source and target domains are represented in the same input space. In contrast to **heterogeneous domain adaptation**, where they are represented in different input spaces.
- **Closed set domain adaptation**, where the source and target domains share the same label space. In contrast to **partial domain adaptation** where the target label space is a proper subset of the source label space, or **open set domain adaptation**, where the target label space includes unknown classes that are not contained in the source one.

[3] There is another well-known definition of domain adaptation as a special case of transfer learning due to (Pan and Yang, 2009). This definition is based on a different conception of domains and tasks. For examples and taxonomy of the different transfer learning settings, we refer the interested reader to (Pan and Yang, 2009; Redko et al., 2019b).

- **Unsupervised domain adaptation**, where labels are only observed from the source domain. In contrast to **semi-supervised domain adaptation**, where few labeled data are also available from the target domain.

Remark 3.20 *Unless specified otherwise, we simply use the term domain adaptation to describe the restricted setting satisfying the assumptions described above.*

Remark 3.21 *In the rest, we design by the source domain interchangeably the distribution $\mathcal{S}$ and the labeled set S, and by the target domain interchangeably the distribution $\mathcal{T}$ and the unlabeled set T.*

3.3.2 Theoretical guarantees

Supervised learning is, beyond a reasonable doubt, the most studied theoretical framework of machine learning. Many of these theoretical studies are concerned with estimating the probability that a specific hypothesis can achieve a small true risk. Uniform convergence theory guarantees the expression of such a probability in the guise of generalization bounds on the true risk, under the overriding presumption that training and test samples are drawn from the same probability distribution. However, the theoretical results of supervised learning do not cover some real-world scenarios where the data-generating processes differ between the training and test samples. Such a predicament has fostered the emersion of domain adaptation (Redko et al., 2019b), a branch of statistical learning theory (Vapnik, 1999) that considers the distributional shift between training and test samples.

In domain adaptation theory, existing generalization bounds on the target risk $\epsilon_{\mathcal{T}}$ of a given hypothesis h are often stated in a generic form implying the source risk $\epsilon_{\mathcal{S}}$, a divergence measure between the marginal distributions of the source and target domains $\text{div}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}})$, and an ability term $a(\mathcal{S}, \mathcal{T})$ assessing the capability of the given hypothesis space to successfully resolve the problem of adaptation:

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \text{div}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) + a(\mathcal{S}, \mathcal{T}), \tag{3.29}$$

The source risk $\epsilon_{\mathcal{S}}$ is estimable from finite samples and can be minimized by learning the hypothesis h from the available source labeled data.

The divergence $\text{div}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}})$ is often estimable from the observed data and is intended to be slight if the two domains are nearby. This divergence is often assessed by comparing the marginals probability distributions $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ instead of the joint distribution $\mathcal{S}$ and $\mathcal{T}$ since we do not have access to labels of the target domain. The most popular frameworks that are used to compare probability distributions in the context of domain adaptation are φ-divergences (Csiszár, 1975), Integral Probability Metric (Zolotarev, 1984) and Optimal Transport (Villani, 2009).

The ability term $a(\mathcal{S}, \mathcal{T})$ is non-estimable, and is usually formulated as the combined error of the ideal joint hypothesis.

In what follows, we present several domain adaptation generalization bounds having the generic form in 3.29.

3.3.2.1 Bounds based on the total variation distance

From a theoretical point of view, the domain adaptation problem was rigorously investigated for the first time by (Ben-David et al., 2006). The authors focused on the domain adaptation problem following Vapnik–Chervonenkis theory and considered the $0-1$ loss function in the setting of binary classification with $\mathcal{Y} = \{-1, +1\}$. The authors considered the total variation distance TV to quantify the divergence between the two domains. The Total Variation is a proper distance on the space of probability measures that quantifies the largest possible difference between the probabilities that the two measures $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ can assign to the same event B. Its definition is the following:

Definition 3.22 (Total variation) *Let $\mathcal{B}$ denote the set of measurable subsets under two probability distributions $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$. The total variation distance TV or the L^1-distance between $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ is defined as:*

$$TV(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = 2 \sup_{B \in \mathcal{B}} |\mu_{\mathcal{S}}(B) - \mu_{\mathcal{T}}(B)| \,. \tag{3.30}$$

Remark 3.23 *The total variation distance is a special case of φ-divergences given in Definition 2.44 corresponding to the function $\varphi(x) = \frac{1}{2}|x-1|$.*

Based on the total variation distance TV, the authors provided the following generalization bound.

Theorem 3.24 *Let's consider the setting of binary classification with $\mathcal{Y} = \{-1, +1\}$. Given two domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathcal{X} \times \mathcal{Y}$, a hypothesis space $\mathcal{H}$ and the $0-1$ loss function, then $\forall h \in \mathcal{H}$, the following holds :*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + TV(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) + \min \left\{ \mathop{\mathbb{E}}_{x \sim \mu_{\mathcal{S}}} \left[|f_{\mathcal{S}}(x) - f_{\mathcal{T}}(x)| \right], \mathop{\mathbb{E}}_{x \sim \mu_{\mathcal{T}}} \left[|f_{\mathcal{T}}(x) - f_{\mathcal{S}}(x)| \right] \right\}, \tag{3.31}$$

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

This theorem introduces the earliest generalization bound that links the performance of a given hypothesis function with respect to two different domains. It involves that the error obtained by a hypothesis h in the source domain is an upper bound on the true error on the target domain, where the tightness of the bound depends on the divergence between their marginal distributions accessed by the total variation distance and that of the labeling functions.

3.3.2.2 Bounds based on the $\mathcal{H}\Delta\mathcal{H}$-divergence

The employment of the total variation distance as a divergence measure between the marginal distributions of the source and target domains presents two major weaknesses. First, the total variation distance is not directly related to the concerned hypothesis space, which results in loose generalization bounds, and secondly, it is not estimable from finite samples drawn from arbitrary probability distributions (Batu et al., 2000). To overcome these limitations, (Ben-David et al., 2010) introduced a classifier-induced divergence called the $\mathcal{H}\Delta\mathcal{H}$-divergence, based on the $\mathcal{A}$-divergence provided in (Kifer et al., 2004), which is a relaxation of the total variation distance.

Definition 3.25 ($\mathcal{A}$-divergence) *Let $\mathcal{A}$ be a collection of measurable sets $A \subset \mathcal{X}$. The $\mathcal{A}$-divergence between two probability distributions $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ over $\mathcal{X}$ is:*

$$d_{\mathcal{A}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = 2 \sup_{A \in \mathcal{A}} |\mu_{\mathcal{S}}(A) - \mu_{\mathcal{T}}(A)| . \tag{3.32}$$

Based on the $\mathcal{A}$-divergence, authors in (Ben-David et al., 2010) introduced the $\mathcal{H}$-divergence, which is a pseudo-metric between the probability measures $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$, defined with respect to a binary hypothesis space $\mathcal{H}$ with $\mathcal{Y} = \{0, 1\}$. Its definition is given below:

Definition 3.26 ($\mathcal{H}$-divergence) *Given a binary hypothesis space $\mathcal{H}$, the $\mathcal{H}$-divergence between two probability distributions $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ over $\mathcal{X}$ is:*

$$d_{\mathcal{H}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = 2 \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{P}}_{x \sim \mu_{\mathcal{S}}} [h(x) = 1] - \mathop{\mathbb{P}}_{x \sim \mu_{\mathcal{T}}} [h(x) = 1] \right| . \tag{3.33}$$

The $\mathcal{H}$-divergence is estimable from finite samples as long as $\mathcal{H}$ has a finite Vapnik-Chervonenkis (VC) dimension, as shown by the following theorem.

Theorem 3.27 *Let $\mathcal{H}$ be a binary hypothesis space with a finite VC dimension and let consider two unlabeled samples S_u, T_u of size m each, drawn independently from $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ respectively. Let $\delta \in (0, 1)$. Then with a probability at least $1 - \delta$ over the choice of samples, we have:*

$$|d_{\mathcal{H}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) - \hat{d}_{\mathcal{H}}(S_u, T_u)| \leq 4 \sqrt{\frac{VC(\mathcal{H}) \log(2m) + \log \frac{2}{\delta}}{m}} . \tag{3.34}$$

This theorem shows that with an increasing number of instances and for a hypothesis space of finite VC dimension, the empirical $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}}$ can be a good proxy for its true counterpart. The former can be further calculated thanks to the following result:

Lemma 3.28 *Let $\mathcal{H}$ be a symmetric binary hypothesis space, i.e. $\forall h \in \mathcal{H}, 1 - h \in \mathcal{H}$, and let consider two unlabeled samples $S_u \sim (\mu_{\mathcal{S}})^n, T_u \sim (\mu_{\mathcal{T}})^m$, then:*

$$\frac{1}{2} \hat{d}_{\mathcal{H}}(S_u, T_u) = 1 - \min_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{x : h(x) = 0} I[x \in S] + \frac{1}{m} \sum_{x : h(x) = 1} I[x \in T] \right) . \tag{3.35}$$

Note that the expression for the empirical $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}}$ provided above is effectively the error of the best classifier for the binary classification problem of discriminating between source and target samples pseudo-labeled with 0 and 1. 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.

Next, authors of (Ben-David et al., 2010) define the symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$ for a hypothesis space $\mathcal{H}$, which is very useful in reasoning about error.

Definition 3.29 (Symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$) *Given a set $\mathcal{H}$ of binary hypothesis taking their values in $\{0, 1\}$, the symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$ is defined as follows:*

$$\mathcal{H}\Delta\mathcal{H} = \{h \oplus h' \mid h, h' \in \mathcal{H}\} = \{|h - h'| \mid h, h' \in \mathcal{H}\}. \tag{3.36}$$

A hypothesis g belongs to $\mathcal{H}\Delta\mathcal{H}$ if and only if it is written as a disagreement between two hypothesis h and h' from $\mathcal{H}$. Based on the $\mathcal{H}\Delta\mathcal{H}$-divergence, (Ben-David et al., 2010) proved a bound on the target risk given in the following theorem.

Theorem 3.30 *Let $\mathcal{H}$ be a hypothesis space of finite VC dimension. Let consider the empirical estimations of size m $\hat{\mu}_{\mathcal{S}}$ and $\hat{\mu}_{\mathcal{T}}$ of $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$, respectively, then for any $\delta \in (0,1)$ with probability of at least $1-\delta$ over the random choice of the samples, we have for all $h \in \mathcal{H}$:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mu}_{\mathcal{S}}, \hat{\mu}_{\mathcal{T}}) + 4\sqrt{\frac{VC(\mathcal{H})\log(2m) + \log\frac{2}{\delta}}{m}} + \lambda, \tag{3.37}$$

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

This bound shows that a good performance on the source domain, similar marginals in terms of the $\mathcal{H}\Delta\mathcal{H}$-divergence, and the existence of a low error of the ideal joint hypothesis are sufficient for successful adaptation. Moreover, we have $\text{VC}(\mathcal{H}\Delta\mathcal{H}) \leq 2\text{VC}(\mathcal{H})$ (Ben-David et al., 2010), hence the $\mathcal{H}\Delta\mathcal{H}$-divergence is estimable from finite samples as long as $\mathcal{H}$ has a finite VC dimension.

3.3.2.3 Bounds based on the l-discrepancy

An obvious shortcoming of the $\mathcal{H}\Delta\mathcal{H}$-divergence is its reliance on the 0 - 1 loss function. Whereas, it might be desirable to have generalization bounds for a more generic domain adaptation framework, where any arbitrary loss function with some suitable properties can be considered. To address this concern, (Mansour et al., 2009) introduced the discrepancy distance $disc_l$ that expands the previous theoretical analysis of domain adaptation for any arbitrary loss function, which is symmetric, bounded, and obeys the triangle inequality. Additionally, the discrepancy distance $disc_l$ relies on the hypothesis space $\mathcal{H}$, but the complexity term is rather related to the Rademacher complexity of $\mathcal{H}$. This distinctive refinement provides data-dependent bounds that are commonly sharper than those derived from Vapnik–Chervonenkis theory.

Definition 3.31 (Mean disagreement) *Given a loss function l and a joint probability $\mathcal{D}$, the mean disagreement [4] between two hypotheses $(h, h') \in \mathcal{H}^2$ is given by $\mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{D}}} l(h(x), h'(x))$.*

Definition 3.32 *Let $\mathcal{H}$ be a hypothesis space, and let $l : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_+$ define a loss function. The discrepancy distance $disc_l$ between two probability distributions $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ over $\mathcal{X}$ is defined by:*

$$disc_l(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = \sup_{(h,h')\in\mathcal{H}^2} \left| \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{S}}} (l(h(x), h'(x)) - \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{T}}} (l(h(x), h'(x)) \right|. \tag{3.38}$$

The l-discrepancy is a pseudo-metric that takes into account the learning task at hand via the hypothesis space $\mathcal{H}$, a property that it shares with the $\mathcal{H}$-divergence. Moreover, the l-discrepancy is related to the $\mathcal{H}\Delta\mathcal{H}$-divergence and the total variation distance. First, for the $0-1$ loss, we have $disc_l(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}})$ and for bounded loss function $\forall(y, y') \in \mathcal{Y}^2$, $l(y, y') \leq M$ for some $M > 0$, we have $disc_l(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) \leq M.\text{TV}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}})$. Furthermore, the l-discrepancy is estimable from finite samples as specified by the following theorem.

[4] By abuse of notations, we can write: $\epsilon_{\mathcal{D}}(h, h') = \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{D}}} l(h(x), h'(x))$

Theorem 3.33 *Let $l_p : \mathcal{Y}^2 \to \mathbb{R}$ be a loss function defined by $l_p(y, y') = |y - y'|^p$ for some $p > 0$. Assume there exists $M > 0$ such that for all $h, h' \in \mathcal{H}$ and all $x \in \mathcal{X}$, $l(h(x), h'(x)) \leq M$. Then, for any $\delta \in (0, 1)$, we have with a probability at least $1 - \delta$ over the choice of sample $S_u \sim (\mu_{\mathcal{S}})^n$ and $T_u \sim (\mu_{\mathcal{T}})^m$:*

$$|disc_l(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) - disc_l(S_u, T_u)| \leq 4p(\mathcal{R}_{S_u}(\mathcal{H}) + \mathcal{R}_{T_u}(\mathcal{H})) + 3M\sqrt{\log\frac{4}{\delta}}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right). \quad (3.39)$$

We now present the theorem that relates the source and target errors using $disc_l$.

Theorem 3.34 *Let $\mathcal{S}$ and $\mathcal{T}$ be the source and target domains over $\mathcal{X} \times \mathcal{Y}$, respectively. Let $\mathcal{H}$ be a hypothesis space, and let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function that is symmetric, obeys the triangle inequality, and is bounded, $\forall (y, y') \in \mathcal{Y}^2, l(y, y') \leq M$ for some $M > 0$. Then, for $h^*_{\mathcal{S}} = \underset{h \in \mathcal{H}}{argmin}\, \epsilon_{\mathcal{S}}(h)$ and $h^*_{\mathcal{T}} = \underset{h \in \mathcal{H}}{argmin}\, \epsilon_{\mathcal{T}}(h)$ denoting the ideal hypothesis for the source and target domains, we have $\forall h \in \mathcal{H}$:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h, h^*_{\mathcal{S}}) + disc_l(S_u, T_u) + 4p(\mathcal{R}_{S_u}(\mathcal{H}) + \mathcal{R}_{T_u}(\mathcal{H})) + 3M\sqrt{\log\frac{4}{\delta}}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) + \lambda', \quad (3.40)$$

where $\epsilon_{\mathcal{S}}(h, h^*_{\mathcal{S}}) = \underset{x \sim \mu_{\mathcal{S}}}{\mathbb{E}} l(h(x), h^*_{\mathcal{S}}(x))$ and $\lambda' = \epsilon_{\mathcal{S}}(h^*_{\mathcal{T}}, h^*_{\mathcal{S}}) + \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}})$.

This bound has two notable distinctions from the former ones. First, the source related term is not the risk of h but rather its disagreement with the best hypothesis $h^*_{\mathcal{S}}$ in $\mathcal{H}$. Second, the ability term λ' is a sum of the disagreement between $h^*_{\mathcal{S}}$ and $h^*_{\mathcal{T}}$ and the risk of $h^*_{\mathcal{T}}$ on the target domain.

3.3.2.4 Bounds based on the MMD distance

Despite their numerous advantages, both the $\mathcal{H}\Delta\mathcal{H}$-divergence and the discrepancy distance $disc_l$ suffer from a computational burden related to their estimation. In such a circumstance, it was natural to look for other metrics with some appealing computational properties to quantify the divergence between the two domains. Following this trend, (Redko, 2015) appealed to the Maximum Mean Discrepancy (MMD) distance to infer generalization bounds analogous to that of (Ben-David et al., 2010). These bounds turned out to be remarkably meaningful since an unbiased estimator of the squared MMD distance can be computed in linear time, and the complexity term does not depend on the Vapnik–Chervonenkis dimension but on the empirical Rademacher complexities of the hypothesis space with respect to the source and target samples.

The Maximum Mean Discrepancy (MMD) is a special case of the Integral Probability Metrics (IPMs) (Zolotarev, 1984).

Definition 3.35 (Integral Probability Metrics) *Let $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ be two probability measures defined on a measurable space $\mathcal{X}$, the* IPM *is defined as follows:*

$$\text{IPM}_{\mathcal{F}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f d\mu_{\mathcal{S}} - \int_{\mathcal{X}} f d\mu_{\mathcal{T}} \right|, \quad (3.41)$$

where $\mathcal{F}$ is a class of real-valued bounded functions defined over $\mathcal{X}$.

As mentioned by (Müller, 1997), the quantity $\text{IPM}_{\mathcal{F}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}})$ is a semi-metric, and it is a metric if and only if the function class $\mathcal{F}$ separates the set of all signed measures

with $\mu(\mathcal{F}) = 0$. It then follows that for any non-trivial function class $\mathcal{F}$, the quantity $\text{IPM}_{\mathcal{F}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}})$ is zero if $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ are the same.

For $\mathcal{F} = \{f : ||f||_{\mathcal{H}_k} \leq 1\}$ where $\mathcal{H}_k$ is a Reproducing kernel Hilbert space (RKHS) with its associated universal kernel k, the IPM distance boils down to the Maximum Mean Discrepancy (MMD).

Definition 3.36 (Maximum Mean Discrepancy) *Let $\mathcal{F} = \{f : ||f||_{\mathcal{H}_k} \leq 1\}$ where $\mathcal{H}_k$ is a RKHS with its associated universal kernel k. Then the Maximum Mean Discrepancy* MMD *distance is defined as follows:*

$$\text{MMD}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = \sup_{||f||_{\mathcal{H}_k} \leq 1} \left| \int_{\mathcal{X}} f d\mu_{\mathcal{S}} - \int_{\mathcal{X}} f d\mu_{\mathcal{T}} \right|. \tag{3.42}$$

Recalling the reproducing kernel property due to Riesz's representation theorem, we have $f(x) = \langle f, k_x \rangle$, where k_x is the image of x in $\mathcal{H}_k$ by some mapping. This allows to separate f from x in the IPM's definition, and to use the autoduality 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 3.37 *Given an RKHS $\mathcal{H}_k$ induced by a universal kernel k, the squared* MMD *between $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ verifies:*

$$\text{MMD}^2(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = \mathop{\mathbb{E}}_{x,x' \sim \mu_{\mathcal{S}}}[k(x, x')] + \mathop{\mathbb{E}}_{x,x' \sim \mu_{\mathcal{T}}}[k(x, x')] - 2 \mathop{\mathbb{E}}_{\substack{x \sim \mu_{\mathcal{S}} \\ x' \sim \mu_{\mathcal{T}}}}[k(x, x')]. \tag{3.43}$$

The latter form gives the supremum defining the MMD in a closed form and allows its efficient empirical estimation with the following theorem.

Theorem 3.38 *Given an RKHS $\mathcal{H}_k$ induced by a universal kernel k, such that $\forall x, x' \in \mathcal{X}, k(x, x') \leq K$ for some $K > 0$. Then for any $\delta \in (0, 1)$, with a probability at least $1 - \delta$ over the draw of $\hat{\mu}_{\mathcal{S}} \sim (\mu_{\mathcal{S}})^n$ and $\hat{\mu}_{\mathcal{T}} \sim (\mu_{\mathcal{T}})^m$, we have:*

$$|\text{MMD}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) - \text{MMD}(\hat{\mu}_{\mathcal{S}}, \hat{\mu}_{\mathcal{T}})| \leq \sqrt{K} \left(\frac{2}{\sqrt{n}} + \frac{2}{\sqrt{m}} + \sqrt{2 \log \frac{2}{\delta} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)} \right). \tag{3.44}$$

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).

Based on the MMD distance between the marginals of the source and target domains, (Redko, 2015) proved the following generalization bound that relates the source and target risks.

Theorem 3.39 *Let $\hat{\mu}_{\mathcal{S}}$ and $\hat{\mu}_{\mathcal{T}}$ be two empirical estimations of size m drawn i.i.d. from $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$, respectively. Then, with probability of at least $1 - \delta$, $\delta \in (0, 1)$ for all $h \in \mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$, the following holds:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \text{MMD}(\hat{\mu}_{\mathcal{S}}, \hat{\mu}_{\mathcal{T}}) + \frac{2}{m} \left(\mathop{\mathbb{E}}_{x \sim \mu_{\mathcal{S}}}[\sqrt{tr(K_{\mathcal{S}})}] + \mathop{\mathbb{E}}_{x \sim \mu_{\mathcal{T}}}[\sqrt{tr(K_{\mathcal{T}})}] \right) + 2\sqrt{\frac{\log(\frac{2}{\delta})}{2m}} + \lambda, \tag{3.45}$$

where $\text{MMD}(\hat{\mu}_{\mathcal{S}}, \hat{\mu}_{\mathcal{T}})$ is an empirical counterpart of $\text{MMD}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}})$, $K_{\mathcal{S}}$ and $K_{\mathcal{T}}$ are the kernel functions calculated on samples from $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$, respectively, and λ is the combined error of the ideal hypothesis h^* that minimizes the joint error of $\epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{T}}(h)$.

We can observe that this theorem is similar in form to Theorem 3.30. The major distinction is that the complexity term does not depend on the Vapnik-Chervonenkis dimension, but on two terms that correspond to the empirical Rademacher complexities of $\mathcal{H}$ with respect to the source and target data. In both theorems, λ acts as the combined error of the ideal joint hypothesis.

3.3.2.5 Bounds based on the Wasserstein distance

Some time later, (Redko et al., 2017) presented generalization bounds in terms of the Wasserstein distance $\mathcal{W}_1$ given in Remark 2.20, as a theoretical analysis of the seminal domain adaptation algorithm based on optimal transport (Courty et al., 2016). This analysis proved to be very fruitful for several reasons. First, the Wasserstein distance is computationally attractive, particularly in virtue of the entropic regularization introduced in (Cuturi, 2013). Furthermore, the Wasserstein distance has the ability to capture the underlying geometry of the data in both domains. Moreover, the Wasserstein distance is quite strong, and according to (Villani, 2009), it is not so hard to associate the convergence information in the Wasserstein distance with certain smoothness bound to obtain convergence in stronger distances. This powerful asset of the Wasserstein distance gives tighter bounds compared to other results in state-of-the-art.

Proposition 3.40 *Given a ground metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the Wasserstein distance $\mathcal{W}_1$ is the* IPM *associated to the space $\mathcal{F}$ of functions verifying the 1-Lipchitz property, i.e:*

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} \mid \forall x, x' \in \mathcal{X}, \, |f(x) - f(x')| \leq d(x, x')\} \tag{3.46}$$

The Wasserstein distance $\mathcal{W}_1$ can be estimated empirically from finite data and this empirical estimation can be a good proxy for its true counterpart as justified by the following result from (Bolley et al., 2007).

Theorem 3.41 *Let μ be a probability measure on $\mathbb{R}^{\mathrm{d}}$ so that for some $\alpha > 0$, we have for any $x_0 \in \mathcal{X}$, $\int_{\mathcal{X}} e^{\alpha d(x_0,x)^2} d\mu(x) < +\infty$, and let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ be its associated empirical measure defined on a sample of independent variables $\{x_i\}_{i=1}^{n}$ all distributed according to μ. Then for any $\mathrm{d}' > \mathrm{d}$ and $\zeta' < \zeta$, there exists some constant N_0 depending on d', ζ' and some square exponential moment of μ, such that for any $\varepsilon > 0$ and $N \geq N_0 \max(\varepsilon^{-(\mathrm{d}'+2)}, 1)$*

$$\mathbb{P}\left[\mathcal{W}_1(\mu, \hat{\mu}) > \varepsilon\right] \leq \exp\left(\frac{-\zeta'}{2} N \varepsilon^2\right). \tag{3.47}$$

The following lemma relates the Wasserstein distance with the source and target error functions for an arbitrary pair of hypothesis.

Lemma 3.42 *Let $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{X})$ be two probability measures on $\mathbb{R}^{\mathrm{d}}$. Assume that the cost function $c(x, y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_k}$, where $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS) equipped with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ induced by $\phi : \mathcal{X} \to \mathcal{H}_k$ and $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_k}$. Assume further that the loss function $l_{h,f} : x \mapsto l(h(x), f(x))$ is convex, symmetric, bounded, obeys triangle equality, and has the parametric form $|h(x) - f(x)|^q$ for some $q > 0$. Assume also that the kernel k in the RKHS $\mathcal{H}_k$ is square-root integrable w.r.t. both $\mu_{\mathcal{S}}, \mu_{\mathcal{T}}$ for all $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{X})$ where $0 \leq k(x, y) \leq K, \forall x, y \in \Omega$. If $\|l\|_{\mathcal{H}_k} \leq 1$, then the following holds:*

$$\forall (h, h') \in \mathcal{H}_k^2, \quad \epsilon_{\mathcal{T}}(h, h') \leq \epsilon_{\mathcal{S}}(h, h') + \mathcal{W}_1(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}). \tag{3.48}$$

We can now use the combination of Lemma 3.42 with the Theorem 3.41 to give the following generalization bound based on the Wasserstein distance $\mathcal{W}_1$.

Theorem 3.43 *With the assumptions of Lemma 3.42. Let $\hat{\mu}_{\mathcal{S}}$ and $\hat{\mu}_{\mathcal{T}}$ be two empirical measures of size n and m drawn i.i.d. from $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$, respectively. Then for any $\mathrm{d}' > \mathrm{d}$ and $\zeta' < \sqrt{2}$, there exists some constant N_0 depending on d', such that for any $\delta > 0$ and $\min(n, m) \geq N_0 \max(\delta^{-(\mathrm{d}'+2)}, 1)$ with probability of at least $1 - \delta$ for all h, we have:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \mathcal{W}_1(\hat{\mu}_{\mathcal{S}}, \hat{\mu}_{\mathcal{T}}) + \sqrt{2\log(\frac{2}{\delta})/\zeta'}\left(\sqrt{\frac{1}{n} + \frac{1}{m}}\right) + \lambda, \tag{3.49}$$

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

The previous bound is the first theoretical justification for the use of optimal transport for domain adaptation. Another bound with the same generic form was given in (Shen et al., 2018) in terms of the Wasserstein distance, but without imposing any additional assumptions on the ground metric used in the definition of the Wasserstein distance. This bound makes use of the absolute value loss in the following way.

Theorem 3.44 *Let $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}(\mathcal{X})$ be two probability measures. Assume the hypothesis $h \in \mathcal{H}$ are all K-Lipschitz continuous for some K and l is the loss function defined by $l(y, y') = |y - y'|$. Then for every $h \in \mathcal{H}$ the following holds:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + 2K\mathcal{W}_1(\hat{\mu}_{\mathcal{S}}, \hat{\mu}_{\mathcal{T}}) + 2K\sqrt{2\log(\frac{2}{\delta})/\zeta'}\left(\sqrt{\frac{1}{n} + \frac{1}{m}}\right) + \lambda, \tag{3.50}$$

In the bound above, the hypothesis space is composed of K-Lipschitz continuous functions for some K. Although this may appear to be too limiting, when the hypothesis are implemented by neural networks, the basic linear mapping functions and activation functions such as sigmoid and relu are all Lipschitz continuous, so the hypothesis is not so strong and can be satisfied. Moreover, the weights in the neural networks are always regularized to avoid overfitting, which means that the constant K will not be too large.

Additionally, (Courty et al., 2017) proposed another generalization bound for domain adaptation based on the Wasserstein distance. This bound introduced the Wasserstein distance between $\mathcal{S}$ and a pseudo-labeled version of the target domain $\tilde{\mathcal{T}}$, with an additional term related to the Probabilistic Transfer Lipschitzness assumption, which is a modified version of the Probabilistic Lipschitzness. The Probabilistic Lipschitzness assumption is a relaxation of the classic deterministic Lipschitzness of a function. It was theoretically studied in (Urner et al., 2011) for semi-supervised learning and in (Ben-David and Urner, 2014) for domain adaptation, we give here the definition used in (Ben-David and Urner, 2014).

Definition 3.45 (Probabilistic Lipschitzness) *Let $(\mathcal{X}, d)$ be a metric space and let consider a function $\phi : \mathcal{X} \to [0, 1]$. We say that $f : \mathcal{X} \to \mathbb{R}$ is ϕ-Lipschitz with respect to a distribution $\mathcal{P}$ over $\mathcal{X}$ if, for all $\lambda > 0$, we have:*

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left[\, \exists x' \in \mathcal{X} \mid |f(x) - f(x')| > \lambda d(x, x') \,\right] \leq \phi(\lambda). \tag{3.51}$$

Deterministic K-Lipschitzness is a particular case of the definition provided above as it corresponds to the setting $\phi(\lambda) = [\lambda < K]$ where $[\cdot]$ denotes the Iverson bracket

for indicator functions, 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 hypothesis, deterministic λ-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.

The Probabilistic Transfer Lipschitzness assumption was proposed in (Courty et al., 2017) in the following way.

Definition 3.46 (Probabilistic Transfer Lipschitzness) *Let $(\mathcal{X}, d)$ be a metric space and let consider a function $\phi : \mathcal{X} \to [0, 1]$. We say that $f : \mathcal{X} \to \mathbb{R}$ is ϕ-Lipschitz with respect to a distribution $\mathcal{P}$ over $\mathcal{X}^2$ and we write $f \in \mathrm{PTL}_\phi(\mathcal{P})$ if, for all $\lambda > 0$, we have:*

$$\mathbb{P}_{(x,x')\sim\mathcal{P}} \left[|f(x) - f(x')| > \lambda d(x, x') \right] \leq \phi(\lambda). \tag{3.52}$$

Based on the Probabilistic Transfer Lipschitzness assumption, (Courty et al., 2017) provided the following generalization bound.

Theorem 3.47 *Let l be an K-Lipschitz loss function for some $K > 0$, assumed to verify the triangle inequality. Let $h \in \mathcal{H}$ and let $\tilde{\mathcal{T}}$ be the target domain distribution with labels predicted by h. For $\alpha > 0$, let $d_{\alpha,l}$ be the cost function defined over $(\mathcal{X} \times \mathcal{Y})^2$ as:*

$$d_{\alpha,l}((x, y), (x', y')) = \alpha d(x, x') + l(y, y'), \tag{3.53}$$

and $\mathcal{W}_1$ its induced Wasserstein distance of order 1 between $\mathcal{S}$ and $\tilde{\mathcal{T}}$. Let $\gamma^* \in \Pi(\mathcal{S}, \tilde{\mathcal{T}})$ be a transport plan defining $\mathcal{W}_1$. Assume there exists a Lipschitz continuous function $f^* \in \mathcal{H}$ such that:

$$f^* = \underset{f^* \in \mathcal{H} \cap \mathrm{PTL}_\phi(\gamma^*)}{argmin} \epsilon_{\mathcal{S}}(f^*) + \epsilon_{\tilde{\mathcal{T}}}(f^*), \tag{3.54}$$

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

$$\epsilon_{\mathcal{T}}(h) \leq \mathcal{W}_1(\hat{\mathcal{S}}, \hat{\tilde{\mathcal{T}}}) + \sqrt{2\log(\frac{2}{\delta})/\zeta'}\left(\sqrt{\frac{1}{n} + \frac{1}{m}}\right) + \epsilon_{\mathcal{S}}(f^*) + \epsilon_{\mathcal{T}}(f^*) + KM\phi(\frac{\alpha}{K}). \tag{3.55}$$

It is important to note that this bound does not have the generic form in 3.29. In fact the previous bound includes the joint error associated with the ideal joint hypothesis f^*, however, the latter is restricted to hypothesis that satisfies the Probabilistic Transfer Lipschitzness with respect to the optimal transport plane γ^*. The last term $\phi(\frac{\alpha}{K})$ assesses the probability under which the Probabilistic Lipschitzness does not hold. If the last terms are small enough, adaptation is possible if we are able to align well $\mathcal{S}$ and $\tilde{\mathcal{T}}$, provided that f^* and γ^* verify the Probabilistic Transfer Lipschitzness.

3.3.2.6 Bounds based on the MDD discrepancy

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

$$\epsilon^\beta_{\mathcal{D}}(h) = \mathbb{E}_{x\sim\mathcal{D}}[l^\beta(h(x), f_{\mathcal{D}}(x))], \tag{3.56}$$

where l^β is the ramp loss (Shalev-Shwartz and Ben-David, 2014), defined as:

$$l^\beta(t) = \begin{cases} 1 - \frac{t}{\beta}, & \text{if } 0 \leq t \leq \beta, \\ [t < 0], & \text{otherwise.} \end{cases} \tag{3.57}$$

which leads to the definition of the Margin Disparity Discrepancy (MDD) given a hypothesis space $\mathcal{H}$ of scoring functions, in the following way.

Definition 3.48 (Scoring functions) *We consider a multiclass classification with hypothesis space $\mathcal{H}$ of scoring functions $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$, where the outputs on each dimension indicate the confidence of prediction. With a little abuse of notations, we consider $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ instead, and $f(x, y)$ indicates the component of $f(x)$ corresponding to the label y. The predicted label associated to a sample x is the one resulting in the largest score. Thus it induces a labeling function space $\mathcal{F}$ containing h_f from $\mathcal{X}$ to $\mathcal{Y}$:*

$$h_f : x \mapsto \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, f(x, y). \tag{3.58}$$

Definition 3.49 (Margin Disparity Discrepancy) *Given a hypothesis space $\mathcal{H}$ of scoring functions and $h \in \mathcal{H}$, the Margin Disparity Discrepancy (MDD) is defined by:*

$$d^\beta_{h,\mathcal{H}}(\mu_\mathcal{S}, \mu_\mathcal{T}) = \sup_{h' \in \mathcal{H}} \left(\epsilon^\beta_{\mu_\mathcal{S}}(h', y(h)) - \epsilon^\beta_{\mu_\mathcal{T}}(h', y(h)) \right). \tag{3.59}$$

As it is the case with the $\mathcal{H}$-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 h and tighter than the $\mathcal{H}$-divergence for $\beta = 0$, corresponding to the $0 - 1$ loss.

The authors show that the MDD is estimable from finite samples with guarantees expressed in terms of the Rademacher complexity, the margin parameter β, the class number k, and sample sizes. Next, the Maximum Disparity Discrepancy (MDD) is used to bound the misclassification rate on the target domain as stated by the following theorem:

Theorem 3.50 *Given a label space $\mathcal{Y} = \{C_1, ..., C_k\}$ and a hypothesis space of scoring functions $\mathcal{H}$, we have for any $\beta > 0$ and $h \in \mathcal{H}$:*

$$\epsilon^{l_{0-1}}_\mathcal{T}(h) \leq \epsilon^\beta_\mathcal{S}(h) + d^\beta_{h,\mathcal{H}}(\mu_\mathcal{S}, \mu_\mathcal{T}) + \lambda^{(\beta)}, \tag{3.60}$$

where $\lambda^{(\beta)} = \inf_{h \in \mathcal{H}} \left(\epsilon^\beta_\mathcal{S}(h) + \epsilon^\beta_\mathcal{T}(h) \right)$.

This bound has the generic form presented in 3.29, with the particularity of the dependence of the divergence term on the considered hypothesis h. This bound also offers new insights into the domain adaptation problem, by introducing the margin violation rate and scoring functions that give the confidence level of belonging to a class of interest, rather than functions with binary output. However, as they bound the 0-1 loss on the target domain, it does not indicate the behavior of the margin violation rate on this latter. It is noteworthy the dependence of the non-estimable term $\lambda^{(\beta)}$ on the classification margin, highlighted by the parameter $\beta > 0$, this remains conceptually similar to the λ term of the other bounds.

In (Dhouib et al., 2020), authors provided a generalization bound using a translated version of the ramp loss given in 3.57, defined as:

$$l^{\rho,\beta}(t) = \begin{cases} 1 - \frac{t-\rho}{\beta}, & \text{if } \rho \leq t \leq \beta + \rho, \\ [t < \rho], & \text{otherwise.} \end{cases} \tag{3.61}$$

for some $\rho > 0$.

The authors proved a bound that is analogous to 3.60 , but concerning the margin violation loss $\epsilon_{\mathcal{T}}^{\rho,0}(h)$ on the target domain, as follows.

Theorem 3.51 *Assume that for any $h' \in \mathcal{H}'$, we have $\mathbb{P}_{x\sim\mu_{\mathcal{S}}}[h'(x) = 0] = \mathbb{P}_{x\sim\mu_{\mathcal{T}}}[h'(x) = 0] = 0$. Let $\rho, \beta, \alpha > 0$ be such that $\rho + \beta < \alpha < 1$. Then, for any $h \in \mathcal{H}$, the following bound holds:*

$$\epsilon_{\mathcal{T}}^{\rho,0}(h) \leq \epsilon_{\mathcal{S}}^{\frac{\rho+\beta}{\alpha},0}(h) + d_{h,\mathcal{H}'}^{\rho,\beta}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) + \lambda_{\alpha}, \tag{3.62}$$

where

$$d_{h,\mathcal{H}'}^{\rho,\beta}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = \sup_{h'\in\mathcal{H}'} \left| \epsilon_{\mu_{\mathcal{S}}}^{\rho,\beta}(h, h') - \epsilon_{\mu_{\mathcal{T}}}^{\rho,\beta}(h, h') \right|, \tag{3.63}$$

and

$$\lambda_{\alpha} = \inf_{h\in\mathcal{H}'} \epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{T}}(h) + \mathbb{P}_{x\sim\mu_{\mathcal{S}}}[|h(x)| < \alpha]. \tag{3.64}$$

Compared to the bound in 3.60, this bound is more informative on the separation quality between classes in the target domain, assessed by the margin violation risk $\epsilon_{\mathcal{T}}^{\rho,0}(h)$. Also, the divergence term is continuous in both h and h' for $\beta > 0$, which makes it more suitable for optimization algorithms. The non estimable term λ_{α} is non symmetric with respect to $\mathcal{T}$ and $\mathcal{S}$ as it involves an absolute margin violation risk only for $\mu_{\mathcal{S}}$. Finally, hypothesis space $\mathcal{H}'$ used to define the divergence and the λ_{α} term on the one hand, and the one concerning h, i.e. $\mathcal{H}$, are not necessarily equal.

3.3.2.7 Algorithmic robustness bounds based on the λ-shift

(Mansour and Schain, 2014) used the concept of algorithmic robustness (Xu and Mannor, 2012) to define the λ-shift that encodes prior knowledge of the deviation between the source and target domains. The goal of their definition was to capture the proximity of the loss associated to a hypothesis on the source and target domains in the regions defined by partitioning the joint space $\mathcal{X} \times \mathcal{Y}$. As there is usually no access to target labels, the authors proposed to consider the conditional distribution of the label in a given region, and the relation to its sampled value over the given labeled sample S. To proceed, let ρ be a distribution over the label space $\mathcal{Y}$, and let σ^y and $\sigma^{-y} = 1 - \sigma^y$ denote the probability of a given label $y \in \mathcal{Y}$ and the total probability of the other labels, respectively. The definition of the λ-shift is then given as follows.

Definition 3.52 (λ-shift) *Let σ and ρ be two distributions over $\mathcal{Y}$. ρ is the λ-shift with respect to σ, denoted by $\rho \in \lambda(\sigma)$, if for all $y \in \mathcal{Y}$ we have $\rho^y \leq \sigma^y + \lambda\sigma^{-y}$ and $\rho^y \geq \sigma^y(1 - \lambda)$. If for some $y \in \mathcal{Y}$ we have $\rho^y = \sigma^y + \lambda\sigma^{-y}$, we say that ρ is strict-λ-shift with respect to σ.*

The above definition means that the λ-shift between two distributions on $\mathcal{Y}$ implies a restriction on the deviation between the probability of a label on the distributions: this shift might be at most a λ portion of the probability of the other labels or the probability of the label.

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

Theorem 3.53 *Let $\mathcal{A}$ be an (M, ϵ)-robust algorithm with respect to a loss function l, such that $0 \leq l(h(x), y) \leq M_l$, for all $(x, y) \in (\mathcal{X} \times \mathcal{Y})$ and $h \in \mathcal{H}$. If $\mathcal{S}$ is λ-shift of $\mathcal{T}$ with respect to the partition of $\mathcal{X}$ for any $\delta \in (0, 1]$, the following bound holds with probability of at least $1 - \delta$, over the random draw of the sample S from $\mathcal{S}$, and of the sample T from $\mathcal{T}$ of size m:*

$$\forall h \in \mathcal{H},\ \epsilon_\mathcal{T}(h) \ \leq \ \sum_{i=1}^{M_x} T(\mathcal{X}_i) l_S^\lambda(h, \mathcal{X}_i) + \epsilon + M_l \sqrt{\frac{2M \ln 2 + 2 \ln \frac{1}{\delta}}{m}}\,, \tag{3.65}$$

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

$$\forall i \in \{1, \ldots, M_\mathcal{X}\}, \quad l_S^\lambda(h, \mathcal{X}_i) \leq \max_{y \in \mathcal{Y}} \left\{ l_i(h, y) \bar{\lambda}^y(\mathcal{S}_i) + \sum_{y' \neq y} l_i(h, y') \underline{\lambda}^{y'}(\mathcal{S}_i) \right\}, \tag{3.66}$$

with

$$l_i(h, y) = \begin{cases} \max_{x \in S \cap \mathcal{X}_i \times y} l(h(x), y) & \text{if } S \cap \mathcal{X}_i \times y \neq \emptyset, \\ M_l & \text{otherwise.} \end{cases} \tag{3.67}$$

The main difference between this domain adaptation result and the original robustness bound of Theorem 3.16 is seen in the first term. In the latter case, which is an upper bound on the source risk, the first term $\frac{1}{m} \sum_{(x,y) \in S} l(h_S(x), y)$ simply corresponds to the empirical error of the model learned on the source sample. In the former bound, which upper-bounds the target risk, the first term $\sum_{i=1}^{M_x} T(\mathcal{X}_i) l_S^\lambda(h, \mathcal{X}_i)$ depends also on the empirical risk on the source sample, which is a combination of the λ-shifted source risk of each region weighted by the ratio of target points in the region. This is reminiscent of the multiplicative dependence between the source error and the divergence term already mentioned in previous sections.

3.3.2.8 PAC-Bayesian bounds based on the $\mathcal{P}$-disagreement

In order to derive domain adaptation generalization bounds in the PAC-Bayesian framework, (Germain et al., 2013) introduced the $\mathcal{P}$-disagreement divergence defined as follows.

Definition 3.54 ($\mathcal{P}$-disagreement divergence) *Given a probability distribution $\mathcal{P}$ over $\mathcal{H}$, the $\mathcal{P}$-disagreement between two probability distributions $\mu_\mathcal{S}$ and $\mu_\mathcal{T}$ over $\mathcal{X}$ is defined as:*

$$\mathrm{dis}_\mathcal{P}(\mu_\mathcal{S}, \mu_\mathcal{T}) = \left| \mathop{\mathbb{E}}_{(h,h') \sim \mathcal{P}} [\epsilon_{\mu_\mathcal{S}}^{l_{0-1}}(h, h') - \epsilon_{\mu_\mathcal{T}}^{l_{0-1}}(h, h')] \right|. \tag{3.68}$$

This quantity is similar to an IPM but with an expectation over hypothesis in $\mathcal{H}$ instead of a supremum. It is worth noting that the value of $\mathrm{dis}_\mathcal{P}$ is always lower than the $\mathcal{H}\Delta\mathcal{H}$-divergence between $\mu_\mathcal{S}$ and $\mu_\mathcal{T}$. Indeed, for every $\mathcal{H}$ and $\mathcal{P}$ over $\mathcal{H}$, we have:

$$\mathrm{dis}_\mathcal{P}(\mu_\mathcal{S}, \mu_\mathcal{T}) \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mu_\mathcal{S}, \mu_\mathcal{T}) \tag{3.69}$$

From the $\mathcal{P}$-disagreement, the authors proved the domain adaptation bound given by the following theorem.

Theorem 3.55 *Let $\mathcal{H}$ a hypothesis space. For any posterior distribution $\mathcal{P}$ over $\mathcal{H}$, we have:*

$$\mathop{\mathbb{E}}_{h\sim\mathcal{P}} \epsilon_{\mathcal{T}}^{l_{0-1}}(h) \leq \mathop{\mathbb{E}}_{h\sim\mathcal{P}} \epsilon_{\mathcal{S}}^{l_{0-1}}(h) + \frac{1}{2}\mathrm{dis}_{\mathcal{P}}(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) + \lambda_{\mathcal{P}} \tag{3.70}$$

where $\lambda_{\mathcal{P}}$ is the deviation between the expected joint errors between pairs for voters on the target and source domains, defined as:

$$\lambda_{\mathcal{P}} = \left| \mathop{\mathbb{E}}_{(h,h')\sim\mathcal{P}^2} \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{S}}} l_{01}(h(x), y)) \times l_{01}(h'(x), y)) - \mathop{\mathbb{E}}_{(h,h')\sim\mathcal{P}^2} \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{T}}} l_{01}(h(x), y)) \times l_{01}(h'(x), y)) \right|. \tag{3.71}$$

This theorem seems very similar to the generic form in 3.29 with an important distinction that consists in substituting the supremum in the domain dissimilarity term with an expectation. We also note, that the posterior distribution $\mathcal{P}$ intercedes even in the non-estimable term $\lambda_{\mathcal{P}}$, which is not the case for all of the previously presented bounds.

3.3.3 Algorithmic advances

Since the launching of domain adaptation theory, a large panoply of algorithms was proposed to deal with it, and they can be roughly divided into shallow (Kouw and Loog, 2019) and deep (Wilson and Cook, 2020) approaches.

Most shallow algorithms try to solve the unsupervised domain adaptation problem in two steps by first aligning the source and target domains to make them indiscernible, which then allows to apply traditional supervised methods on the transformed data. Such an alignment is typically accomplished through sample-based approaches, which focus on correcting biases in the sampling procedure (Shimodaira, 2000; Sugiyama et al., 2007) or feature-based approaches which focus on learning domain-invariant representations (Pan et al., 2010; Long et al., 2013) and finding subspace mappings (Fernando et al., 2013; Sun and Saenko, 2015; Sun et al., 2016).

Deep domain adaptation algorithms have also gained a renewed interest due to their feature extraction ability to learn more abstract and robust representations that are both semantically meaningful and domain invariant (Long et al., 2015; Glorot et al., 2011; Ganin et al., 2016).

More recent advances in domain adaptation are due to the theory of optimal transport, which allows to learn explicitly the least cost transformation of the source distribution into the target one, and provide both shallow (Courty et al., 2016, 2017; Redko et al., 2019a) and deep algorithms (Damodaran et al., 2018; Shen et al., 2018; Dhouib et al., 2020; Rakotomamonjy et al., 2022).

Below we present the different categories of domain adaptation algorithms and the most relevant approaches within each one.

3.3.3.1 Sample-based approaches

This category of approaches relies on two assumptions: the covariate shift and the absolute continuity assumption $\mu_{\mathcal{T}} \ll \mu_{\mathcal{S}}$ described bellow.

Definition 3.56 (Covariate shift) *The covariate shift assumption corresponds to the setting where the conditional distributions of the source and target domains are equal and the shift is only due to the distribution of the covariates (the features) as follows:*

$$\mu_{\mathcal{S}} \neq \mu_{\mathcal{T}} \qquad \text{and} \qquad \mathcal{S}_{\mathcal{Y}|x} = \mathcal{T}_{\mathcal{Y}|x} \tag{3.72}$$

Definition 3.57 (Absolute continuity) *Let $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ be two probability distribution over $\mathcal{X}$. $\mu_{\mathcal{T}}$ is said to be absolutely continuous [5] with respect to $\mu_{\mathcal{S}}$, and we note $\mu_{\mathcal{T}} \ll \mu_{\mathcal{S}}$ if for all measurable subset $A \subset \mathcal{X}$:*

$$\mu_{\mathcal{S}}(A) = 0 \implies \mu_{\mathcal{T}}(A) = 0 \tag{3.73}$$

Theorem 3.58 *Let $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ be two probabilities over $\mathcal{X}$, such that $\mu_{\mathcal{T}} \ll \mu_{\mathcal{S}}$. Then there exists a measurable function $f : \mathcal{X} \to \mathbb{R}_+$, verifying for all measurable subset $A \subset \mathcal{X}$:*

$$\mu_{\mathcal{T}}(A) = \int_A f d\mu_{\mathcal{S}} \tag{3.74}$$

The function f is unique $\mu_{\mathcal{S}}$-almost everywhere and is called the Radon-Nikodym derivative of $\mu_{\mathcal{T}}$ with respect to $\mu_{\mathcal{S}}$ and we write $f = \frac{d\mu_{\mathcal{T}}}{d\mu_{\mathcal{S}}}$.

Defining the importance as $w(x) = \frac{d\mu_{\mathcal{T}}}{d\mu_{\mathcal{S}}}$, this approach can be deduced as follows:

$$\begin{aligned}
\epsilon_{\mathcal{T}}(h) &= \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{T}}}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{T}_{\mathcal{Y}|x}}(l(h(x),y))\right] && (3.75)\\
&= \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{S}}}\left[\frac{d\mu_{\mathcal{T}}}{d\mu_{\mathcal{S}}}(x)\mathop{\mathbb{E}}_{y\sim\mathcal{T}_{\mathcal{Y}|x}}(l(h(x),y))\right] && (3.76)\\
&= \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{S}}}\left[w(x)\mathop{\mathbb{E}}_{y\sim\mathcal{S}_{\mathcal{Y}|x}}(l(h(x),y))\right] && (3.77)\\
&= \mathop{\mathbb{E}}_{x\sim\mu_{\mathcal{S}}}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{S}_{\mathcal{Y}|x}}(w(x)l(h(x),y))\right] && (3.78)\\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{S}}\big[w(x)l(h(x),y)\big] && (3.79)
\end{aligned}$$

The second line is due to the absolute continuity assumption and the third line is due to the covariate shift assumption. Hence, minimizing the risk on the target domain boils down to minimizing a weighted risk on the source domain, with weights $w(x)$ that are independent of the labels.

Sample re-weighting is one of the earliest approaches in domain adaptation, and the estimation of weights $w(x)$ is the key challenge in these approaches. For this approach, the probability weights of the instances from both domains are estimated, and then the ratio is used to compute $w(x)$. The estimation of every domain's probability weights can be parametric (Shimodaira, 2000), non parametric, e.g. by kernel density estimation (Sugiyama et al., 2007; Baktashmotlagh et al., 2014) or using neural networks (de Mathelin et al., 2022).

[5] Some authors prefer to use the term "dominated by" instead of "absolutely continuous w.r.t.".

3.3.3.2 Feature-based approaches

Approaches of the previous category are limited by the assumption of the absolute continuity of the marginal source distribution with respect to the target one, an assumption that is violated in several tasks. Feature-based approaches do not require this assumption and try to match the source samples with the target ones by learning a transformation that extracts the invariant feature representation across domains. They typically create a new representation by transforming the original features into a new feature space, and then minimize the shift between the source and target domains in an optimization procedure. Below we detail the most relevant feature-based adaptation categories and discuss some related approaches.

3.3.3.2.1 Subspace mappings

In some cases, source and target domains contain domain-specific noise but common subspaces. Subspace mappings approaches would involve identifying these subspaces and matching the source and the target samples along them.

One of the most well-known methods in this category, called subspace alignment (SA) (Fernando et al., 2013), calculates the first d′ principal components in each domain to form two subspaces defined by the matrices $C_{\mathcal{S}}$ and $C_{\mathcal{T}}$, where d′ < d. A transformation matrix M^* is then computed to aligns the source components with the target ones by minimizing the following Bregman matrix divergence:

$$M^* = \underset{M}{\operatorname{argmin}} || C_{\mathcal{S}} M - C_{\mathcal{T}} ||_F^2 = C_{\mathcal{S}}^T C_{\mathcal{T}}, \tag{3.80}$$

where $||.||_F^2$ is the Frobenius norm. Then, the matrix M^* aligns $C_{\mathcal{S}}$ with $C_{\mathcal{T}}$, before projecting data in each domain to their components and training a classifier on the transformed source data.

In this spirit, the linear correlation alignment (CORAL) (Sun et al., 2016) minimized the domain shift by aligning the second-order statistics of source and target distributions, by solving the following optimization problem:

$$\min_M || C_{\hat{\mathcal{S}}} - C_{\mathcal{T}} ||_F^2 = \min_M || M^T C_{\mathcal{S}} M - C_{\mathcal{T}} ||_F^2 \tag{3.81}$$

where M is the transformation matrix, and here, $C_{\hat{\mathcal{S}}}$ is the covariance of the transformed source features, $C_{\mathcal{S}}$ and $C_{\mathcal{T}}$ are covariance matrices of source and target domains, respectively.

Other extensions are possible, such as considering the alignment on both the data distributions and the subspaces (Sun and Saenko, 2015), and training a classifier jointly with learning the subspace (Fernando et al., 2015).

Instead of subspaces, some extensions consider non-linear manifolds, such as in (Aljundi et al., 2015) where a kernelized version of subspace alignment is used. Other approaches based on non-linear manifolds aim to map the data on a Riemannian manifold and reduce the distance between the two domains on it. Some of these earliest approaches (Gopalan et al., 2011; Gong et al., 2012) learn the intermediate features between the sub-source and the sub-target domains via the geodesic (shortest path) on a Grassmannian manifold.

3.3.3.2.2 Domain-invariant spaces

Another common practice focuses on aligning the source and target samples through a learned domain-invariant space. The advantage of this approach is that classification becomes the same as standard supervised learning.

A simple approach to finding a domain-invariant space is Transfer component adaptation (TCA) (Pan et al., 2010) that aims to find common latent features having the same marginal distribution across the source and target domains while maintaining the intrinsic structure of the original samples. The latent features are learned between the source and target domains in a RKHS using the maximum mean discrepancy. In order to learn principal components C based on joint directions of variation, the joint domain kernel matrix, $K = [K_{\mathcal{S},\mathcal{S}}, K_{\mathcal{S},\mathcal{T}}, K_{\mathcal{T},\mathcal{S}}, K_{\mathcal{T},\mathcal{T}}]$ composed of kernel matrices of samples in the source domain, target domain, and cross domains, is first constructed. Data projected onto components C should have minimal distance to the empirical means in each domain. As such, components are extracted by minimizing the trade-off between the trace of the projected joint domain kernel matrix and a regularization term $\text{tr}(C^T C)$ that control the complexity of C:

$$\begin{aligned} \min_{C} \quad & \text{trace}(C^T KLKC) + \alpha \text{trace}(C^T C) \\ \text{s.t.} \quad & C^T KHKC = I \end{aligned} \tag{3.82}$$

where α is a trade-off parameter, L is the normalization matrix that divides each entry in the joint kernel by the sample size of the domain from which it originated, and H is the centering matrix. The constraint is necessary to avoid trivial solutions, such as projecting all data to 0. After finding the domain-invariant features, any classical supervised learning technique can be used to train the final target classifier.

Joint domain adaptation (JDA) (Long et al., 2013) extends (TCA) by matching simultaneously marginal and conditional distributions of the source and target domains. Principal component analysis is employed for optimization and dimensionality reduction. To address the divergence in marginal distribution between the domains, the maximum mean discrepancy distance is used to calculate the marginal distribution differences and is incorporated into the PCA optimization algorithm. The second part of the solution needs a procedure to rectify the conditional distribution differences, which requires labeled target samples. Since the target data is unlabeled, pseudo labels are formed by learning a classifier from the labeled source samples. The maximum mean discrepancy distance is changed to measure the distance between the conditional distributions and is integrated into the PCA optimization algorithm to minimize the conditional distributions:

$$\min_{C^T XHX^T C=I} \sum_{c=0}^{k} \text{trace}(C^T X M_c X^T C) + \alpha ||C||_F^2 \tag{3.83}$$

where C is an orthogonal transformation matrix, M_c are the MMD matrices involving class labels and α is a regularization parameter to guarantee the optimization problem is well-defined. Finally, the features identified by the modified PCA algorithm are used to train the final target classifier.

3.3.3.2.3 Deep domain adaptation

Firstly, (Long et al., 2015) proposed Deep Adaptation Network (DAN) that employs deep neural networks to learn transferable features across domains. DAN relies on the assumption that there is a gap between the marginal distributions while the conditional distributions stay unchanged. Consequently, it aims to match marginal distributions across domains by including several adaptation layers for task-specific representations. The adaptation layers use the multiple kernel variant of the MMD to integrate all the task-specific representations into a RKHS and align the shift between the marginal distributions. Deep Transfer Network (DTN) (Zhang et al., 2015), has been proposed later to align simultaneously both the marginal and conditional distributions.

Another category of deep adaptation networks uses autoencoders to reduce the divergence between domains by minimizing the reconstruction error and learning an invariant and transferable representation across domains. The main underlying idea behind using autoencoders in domain adaptation is to learn the encoder parameters based on the source samples and adapt the decoder to reconstruct the target samples. In this vision, (Glorot et al., 2011) proposed a deep domain adaptation network based on stacked autoencoders (SDA), in order to extract a high-level representation of source and target samples. SDA yields remarkable results, but it is computationally expensive and unscalable, especially when dealing with high-dimensional features. A marginalized version (mSDA) has been proposed in (Chen et al., 2012) to address SDA limitations by marginalizing the noise with linear denoisers to make the model learn the parameters in a closed-form solution without using stochastic gradient descent (SGD). Later on, (Ghifary et al., 2016b) proposed a Deep Reconstruction Classification Network (DRCN) that consists of a standard convolutional network (encoder) to predict the source labels and a deconvolutional networks (decoder) to reconstruct the target samples.

The popularity of adversarial learning as a strong domain-invariant feature extractor has prompted many researchers to incorporate it into deep networks. Adversarial domain adaptation approaches seek to minimize the distributional gap between domains in order to obtain transferable and domain-invariant features. The main idea of adversarial domain adaptation was inspired by generative adversarial networks (GAN) (Goodfellow et al., 2014), which aims to minimize the cross-domain discrepancy through an adversarial objective. GANs are generative models based on deep learning, consisting of a two-player game, a generator model G, and a discriminator model D. The generator seeks to output samples similar to the domain of interest from the source data and to confuse the discriminator into making a wrong decision. The discriminator then tries to discriminate between the true samples of the domain of interest and the counterfeits generated by the model G. (Ganin et al., 2016) is one of the most popular deep adaptative networks, which is directly derived from the seminal theoretical contribution in (Ben-David et al., 2006), its main idea is to embed domain adaptation into the representation learning process, so that the final classification decisions are made based on features that are both discriminative and invariant to domain changes. Later on, the work of (Shrivastava et al., 2017; Bousmalis et al., 2017; Zhang et al., 2019) provided other noteworthy adversarial learning algorithms for the problem of domain adaptation.

3.3.3.2.4 Optimal transport

More recent advances in domain adaptation are due to the theory of optimal transport that allows to learn explicitly the least cost transformation of the source distribution into the target one. This idea was first investigated in the work of (Courty et al., 2016), where authors have successfully cast the domain adaptation problem into an optimal transport one to match the shifted marginal distributions of the two domains:

$$\min_{\gamma \in U(a,b)} \langle \gamma, C \rangle_F + \varepsilon H(\gamma) + \eta \Omega_c(\gamma), \tag{3.84}$$

where $\eta > 0$ and $\Omega_c(\gamma)$ is a class-based regularization, that can be either based on group sparsity and then promotes an optimal transport plan where a given target sample receives masses from source samples that have the same labels, or alternatively based on graph Laplacian regularization and then promote a locally smooth and class-regular structure in the source transported samples. Which then allows to learn a classifier on the transported data.

Since then, several optimal transport-based domain adaptation approaches have emerged. In (Courty et al., 2017), authors proposed to avoid the two-steps adaptation procedure, by aligning the joint distributions using a coupling accounting for the marginals and the class-conditional distributions shift jointly:

$$\min_{f \in \mathcal{H}, \gamma \in U(a,b)} \sum_{i,j} \left[\alpha d(x_i, x_j) + l(y_i, f(x_j)) \right] . \gamma_{i,j} + \lambda \Omega(f) \tag{3.85}$$

where $\Omega(f)$ is a regularization term on f. Depending on how $\mathcal{H}$ is defined, a RKHS or a function space parametrized by some parameters $w \in \mathbb{R}^p$, $\Omega(f)$ can be either a non-decreasing function of the squared-norm induced by the RKHS, or a squared-norm on the vector parameter, and $\alpha, \lambda > 0$.

Authors in (Redko et al., 2019a) performed multi-source domain adaptation under the target shift assumption, by learning simultaneously the class probabilities of the unlabeled target samples and the optimal transport plan allowing to align several probability distributions.

The work of (Dhouib et al., 2020) derived an efficient optimal transport-based adversarial approach from a bound on the target margin violation rate and the more recent work of (Rakotomamonjy et al., 2022) addressed the problem of generalized target shift, where we have both label shift and class-conditional distribution shift by proposing an algorithm that minimizes importance weighted loss in the source domain and a Wasserstein distance between weighted marginals. Finally, several deep domain adaptation algorithms based on optimal transport were proposed in (Damodaran et al., 2018; Shen et al., 2018; Chen et al., 2018; Xu et al., 2020; Li et al., 2020) to name a few.

Bibliographical notes

The main reference in domain adaptation is undoubtedly the book by Redko et al. that exhaustively covers the theoretical advances in this field. While a general introduction to statistical learning theory can be found in the books by Mohri et al. and Shalev-Shwartz and Ben-David and to transfer learning in the book by Yang et al..

Chapter 4

Hierarchical Optimal Transport for Domain Adaptation

Contents

This chapter is based on the paper (El Hamri et al., 2022b) where we propose a novel approach for unsupervised domain adaptation that relates notions of optimal transport, learning probability measures, and unsupervised learning. The proposed approach, HOT-DA, is based on a hierarchical formulation of optimal transport, that leverages beyond the geometrical information captured by the ground metric, richer structural information in the source and target domains. The additional information in the labeled source domain is formed instinctively by grouping samples into structures according to their class labels. While exploring hidden structures in the unlabeled target domain is reduced to the problem of learning probability measures through Wasserstein barycenter, which we prove to be equivalent to spectral clustering. Experiments show the superiority of the proposed approach over state-of-the-art across a range of domain adaptation problems including inter-twinning moons dataset, Digits, Office-Caltech, and Office-Home. Experiments also show the robustness of our model against structure imbalance.

Without hardship everyone would reign,
generosity impoverishes and bravery kills.

Al-Mutanabbi

4.1 Introduction

Supervised learning is arguably the most widespread task of machine learning and has enjoyed much success on a broad spectrum of application domains (Kotsiantis et al., 2007). However, most supervised learning methods are built on the crucial assumption that training and test data are drawn from the same probability distribution (Pan and Yang, 2009). In real-world applications, this hypothesis is usually violated due to several application-dependent reasons: in computer vision, the presence or absence of backgrounds, the variation of acquisition devices, or the change of lighting conditions introduce non-negligible discrepancies in data distributions (Saenko et al., 2010), in product reviews classification, the drifts observed in the word distributions are caused by the difference of product category and the changes in word frequencies (Blitzer et al., 2007).

These distributional shifts will be likely to degrade significantly the generalization ability of supervised learning models. While manual labeling may appear like a feasible solution, such an approach is unreasonable in practice, since it is often prohibitively expensive to collect from scratch a new large high quality labeled dataset with the same distribution as the test data, due to lack of time, resources, or other factors, and it would be an immense waste to totally reject the available knowledge on a different, yet related labeled training set. Such a challenging situation has promoted the emergence of domain adaptation (Redko et al., 2019b), a sub-field of statistical learning theory (Vapnik, 1999), that takes into account the distributional shift between training and test data, and in which the training set and test set distributions are respectively called source and target domains.

Since the launching of domain adaptation theory, a large panoply of algorithms was proposed to deal with its unsupervised variant, and they can be roughly divided into shallow (Kouw and Loog, 2019) and deep (Wilson and Cook, 2020) approaches. Most shallow algorithms try to solve the unsupervised domain adaptation problem in two steps by first aligning the source and target domains to make them indiscernible, which then allows to apply traditional supervised methods on the transformed data. Such an alignment is typically accomplished through sample-based approaches which focus on correcting biases in the sampling procedure (Shimodaira, 2000; Sugiyama et al., 2007) or feature-based approaches which focus on learning domain-invariant representations (Pan et al., 2010; Long et al., 2013) and finding subspace mappings (Fernando et al., 2013; Sun and Saenko, 2015; Sun et al., 2016). Deep domain adaptation algorithms have also gained a renewed interest due to their feature extraction ability to learn more abstract and robust representations that are both semantically meaningful and domain invariant (Glorot et al., 2011; Long et al., 2015; Ganin et al., 2016).

More recent advances in domain adaptation are due to the theory of optimal transport, which allows to learn explicitly the least cost transformation of the source distribution into the target one. This idea was first investigated in the work of (Courty et al., 2016), where authors have successfully cast the domain adaptation problem into an optimal transport one to match the shifted marginal distributions of the two domains, which then allows to learn a classifier on the transported data. Since then, several optimal transport-based domain adaptation methods have emerged. In (Courty et al., 2017), authors proposed to avoid the two-step adaptation procedure, by aligning the joint distributions using a coupling accounting for the marginals and the class-conditional distributions shift jointly. Authors in (Redko et al., 2019a) performed multi-source domain adaptation under the target shift assumption, by learning simultaneously the class probabilities of the unlabeled target samples and the optimal transport plan allowing to align several probability distributions. The recent work of (Dhouib et al., 2020) derived an efficient optimal transport-based adversarial approach from a bound on the target margin violation rate. Finally, several deep domain adaptation algorithms based on optimal transport were proposed in (Damodaran et al., 2018; Shen et al., 2018; Chen et al., 2018; Xu et al., 2020; Li et al., 2020) to name a few.

A common denominator of these approaches is their ability to capture the underlying geometry of the data by relying on the cost function that reflects the metric of the input space. However, these optimal transport-based methods can benefit from not relying solely on such rudimentary geometrical information, since there is further important structural information that remains uncaptured directly from the ground metric, e.g., the local consistency induced by class labels in the source. The exploitation of this structural information can elicit some desired properties in domain adaptation like preserving compact classes during the transportation. It is, moreover, what led authors in (Courty et al., 2016) to propose the inclusion of this structural information by adding a group-norm regularizer. Such structures, however, could not be induced directly by the standard formulation of optimal transport. To the best of our knowledge, (Alvarez-Melis et al., 2018) is the only work that has attempted to incorporate structural information directly into the optimal transport problem without the need to add a regularization term. This approach developed a nonlinear generalization of discrete optimal transport based on submodular functions. However, the application of this method in domain adaptation only takes into account the available structures in the labeled source domain, by partitioning samples according to their class labels, while every target sample forms its own cluster. Nonetheless, richer structures in the target domain can be easily captured differently, e.g., by grouping, and the incorporation of such target structures directly into the optimal transport formulation can lead in our view to a significant improvement in the performance of domain adaptation algorithms.

Contributions: In this chapter, we address the existing limitations of the target-structure-agnostic algorithms mentioned above by proposing a principally new approach based on hierarchical optimal transport (Schmitzer and Schnörr, 2013). Hierarchical optimal transport is an effective and efficient paradigm to induce structural information into the transportation procedure. It has been recently used for different tasks such as multi-level clustering (Ho et al., 2017), multimodal distribution alignment (Lee et al., 2019), document representation (Yurochkin et al., 2019) and semi-supervised learning (Taherkhani et al., 2020). The relevance of this paradigm

for domain adaptation is illustrated in Figure 4.1, where we show that the structure-agnostic Reg-OT (Cuturi, 2013) and target-structure-agnostic OT-GL (Courty et al., 2016) algorithms fail to always restrict the transportation of mass across instances of different structures, whereas, our Hierarchical Optimal Transport for Domain Adaptation (HOT-DA) model manages to do it correctly by leveraging the source and target structures simultaneously, which will subsequently lead to a better adaptation.

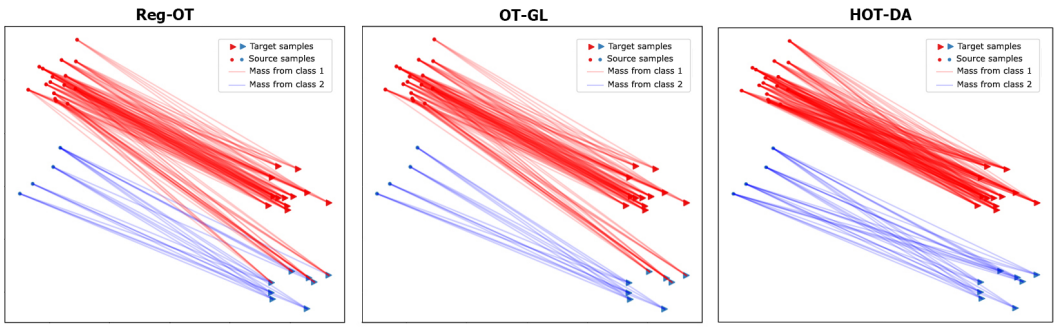

FIGURE 4.1: Illustration of the transportation obtained with structure-agnostic Reg-OT (Cuturi, 2013) and target-structure-agnostic OT-GL (Courty et al., 2016) methods, and our proposed algorithm HOT-DA.

To the best of our knowledge, the proposed approach is the first hierarchical optimal transport method for unsupervised domain adaptation, and the first work to shed light on the connection between spectral clustering and Wasserstein barycenter.

Outline: The rest of this paper is organized as follows: in the 2^{nd} section, we present the hierarchical formulation of optimal transport. In the 3^{rd} section, we elaborate the proposed approach HOT-DA. In the 4^{th} section, we evaluate our algorithm on a toy dataset and three benchmark visual adaptation problems, and we study the relevance of Wasserstein-Spectral clustering to HOT-DA as well as the sensitivity of our approach to unbalanced structures. Finally, we conclude in section 5.

4.2 Hierarchical optimal transport

Hierarchical optimal transport is an attractive formulation that offers an efficient way to induce structural information directly into the transport process (Schmitzer and Schnörr, 2013). The main underlying idea behind this formulation is to organize the data in X and Y into structures (e.g., classes or clusters), this hierarchical organization allows to look at both X and Y as a collection of structures. To compute the hierarchical optimal transport plan between these two collections, the cost function can no longer be evaluated using a distance that quantitatively defines the closeness between data, such as the Euclidean distance, we must therefore employ another metric able to measure the discrepancy between structures. Since each structure can be represented by a discrete measure, the Wasserstein distance is an evident choice. Obviously, computing the Wasserstein distance between each pair of structures requires solving a prior optimal transport problem between samples of the two structures. Therefore, if X and Y are composed of h and l structures respectively, then, the Wasserstein cost matrix would require a prior computation of $h \times l$ optimal transport problems, before solving the final optimal transport problem between classes and clusters, hence the hierarchy.

More formally, let $\mathcal{X}$ be a Polish metric space endowed with a distance d and $\mathcal{P}(\mathcal{X})$ be the space of Borel probability measures on $\mathcal{X}$ equipped with the Wasserstein distance $\mathcal{W}_p$ according to (2.11). Since $\mathcal{X}$ is a Polish metric space, then $\mathcal{P}(\mathcal{X})$ is also a Polish metric space (Parthasarathy, 2005).

By a recursion of concepts, $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ the space of Borel probability measures on $\mathcal{P}(\mathcal{X})$ is a Polish metric space and will be equipped then with the Wasserstein metric that we note $\mathcal{HW}_p$, induced this time by the Wasserstein distance $\mathcal{W}_p$ which acts as the ground metric on $\mathcal{P}(\mathcal{X})$.

Let $\Phi = \{\rho_1, ..., \rho_h\} \subset \mathcal{P}(\mathcal{X})$ and $\Psi = \{\varrho_1, ..., \varrho_l\} \subset \mathcal{P}(\mathcal{X})$ be two sets of probability measures over $\mathcal{P}(\mathcal{X})$ (each probability measure represents a structure). The empirical distributions of Φ and Ψ can be expressed respectively by $\phi, \varphi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ as $\phi = \sum_{i=1}^{h} \alpha_i \delta_{\rho_i}$ and $\varphi = \sum_{j=1}^{l} \beta_j \delta_{\varrho_j}$, where $\alpha = (\alpha_1, ..., \alpha_h)$ and $\beta = (\beta_1, ..., \beta_l)$ are vectors in the probability simplex $\sum_h$ and $\sum_l$ respectively (ϕ and φ represent the two collections of structures). The hierarchical optimal transport problem between ϕ and φ is then:

$$(\mathcal{HOT}) \quad \min_{\Gamma \in U(\alpha,\beta)} \langle \Gamma, \mathcal{W} \rangle_F \tag{4.1}$$

where the matrix $\mathcal{W} = (\mathcal{W}_p(\rho_i, \varrho_j))_{\substack{1 \leq i \leq h \\ 1 \leq j \leq l}} \in \mathcal{M}_{h \times l}(\mathbb{R}^+)$ stands for the Wasserstein cost matrix and $U(\alpha, \beta) = \{\Gamma \in \mathcal{M}_{h \times l}(\mathbb{R}^+) \mid \Gamma \mathbb{1}_l = \alpha \text{ and } \Gamma^{\mathbf{T}} \mathbb{1}_h = \beta\}$ represents the new transportation polytope. More intuitive insights are provided in Figure 4.3.

4.3 Hierarchical Optimal Transport for Domain Adaptation

In this section, we introduce the proposed HOT-DA approach, which consists of three phases, the first one aims to learn hidden structures in the unlabeled target domain using Wasserstein barycenter, which we prove can be equivalent to spectral clustering, the second phase focuses on finding a one-to-one matching between structures of the two domains through the hierarchical optimal transport formulation, and the third phase involves transporting samples of each source structure to its corresponding target structure via the barycentric mapping.

4.3.1 Learning unlabeled target structures through Wasserstein-Spectral clustering

Samples in the source domain $S = \{(x_i, y_i)\}_{i=1}^{n}$ can be grouped into structures according to their class labels, but, data in the target domain $T = \{x_j\}_{j=1}^{m}$ are not labeled to allow us to identify directly such structures. Removing this obstacle cannot be accomplished without using some additional assumptions. In fact, to exploit efficiently the unlabeled data in the target domain, the most plausible assumption stems from the structural hypothesis based on clustering, where it is assumed that the data belonging to the same cluster are more likely to share the same label. This assumption constitutes the core nucleus for the first phase of our approach, which aims to prove that spectral clustering can be cast as a problem of learning probability measures with respect to Wasserstein barycenter. Our proof is based on three key ingredients: the equivalence between the search for a 2-Wasserstein barycenter of the empirical distribution that represents unlabeled data and k-means clustering,

the analogy between traditional k-means and kernel k-means and finally the connection between kernel k-means and spectral clustering. We derive from this result a novel algorithm able to learn efficiently hidden structures of arbitrary shapes in the unlabeled target domain.

Firstly, given m unlabeled instances $\{x_1, ..., x_m\} \subset \mathcal{X}$, k-means clustering (MacQueen et al., 1967) aims to partition the m samples into k clusters $\Pi_k = \{\pi_1, ..., \pi_k\}$ in which each sample belongs to the cluster with the nearest center. This results in a partitioning of the data space into Voronoi cells $(\text{Vor}_q)_{1 \leq q \leq k}$ generated by the cluster centers $\tilde{C}_k = \{c_1, ..., c_k\}$. The goal of k-means then is to minimize the mean squared error, and its objective function is defined as:

$$\min_{c_1,...,c_k} \frac{1}{m} \sum_{i=1}^{m} \|x_i - c_j\|^2 \tag{4.2}$$

Let $\hat{\rho}_m = \sum_{i=1}^{m} \frac{1}{m} \delta_{x_i}$ be the empirical distribution of $\{x_1, ..., x_m\}$. Since $\frac{1}{m} \sum_{i=1}^{m} \|x_i - c_j\|^2 = \mathbb{E}_{x \sim \hat{\rho}_m} \|x - \tilde{C}_k\|^2$, then according to (Canas and Rosasco, 2012):

$$\frac{1}{m} \sum_{i=1}^{m} \|x_i - c_j\|^2 = \text{W}_2^2(\hat{\rho}_m, \pi_{\tilde{C}_k} \# \hat{\rho}_m) \tag{4.3}$$

where $\pi_{\tilde{C}_k} : \mathcal{X} \to \tilde{C}_k$ is the projection function mapping each $x \in \text{Vor}_q \subset \mathcal{X}$ to c_q. Since k-means minimizes (4.3), it also finds the measure that is closest to $\hat{\rho}_m$ among those with support of size k (Pollard, 1982). Which proves the equivalence between k-means and searching for a 2-Wasserstein barycenter of $\hat{\rho}_m$ in $\mathcal{P}_k(\mathcal{X})$, i.e., a minimizer in $\mathcal{P}_k(\mathcal{X})$ of:

$$f(\kappa) = \text{W}_2^2(\hat{\rho}_m, \kappa) \tag{4.4}$$

Secondly, k-means suffers from a major drawback, namely that it cannot separate clusters that are nonlinearly separable in the input space. Kernel k-means (Schölkopf et al., 1998) can overcome this limitation by mapping the input data in $\mathcal{X}$ to a high-dimensional reproducing kernel Hilbert space $\mathcal{H}$ by a nonlinear mapping $\psi : \mathcal{X} \to \mathcal{H}$, then the traditional k-means is applied on the high-dimensional mappings $\{\psi(x_1), ..., \psi(x_m)\}$ to obtain a nonlinear partition. Thus, the objective function of kernel k-means can be expressed analogously to that of traditional k-means in (4.2):

$$\min_{c_1,...,c_k} \frac{1}{m} \sum_{i=1}^{m} \|\psi(x_i) - c_j\|^2 \tag{4.5}$$

Usually, the nonlinear mapping $\psi(x_i)$ cannot be explicitly computed, instead, the inner product of any two mappings $\psi(x_i)^T \psi(x_j)$ can be computed by a kernel function $\mathcal{K}$. Hence, the whole data set in the high-dimensional space can be represented by a kernel matrix $K \in \mathcal{M}_m(\mathbb{R}^+)$, where each entry is defined as: $K_{i,j} = \mathcal{K}(x_i, x_j) = \psi(x_i)^T \psi(x_j)$.

Thirdly, according to (Zha et al., 2001), the objective function of kernel k-means in (4.5) can be transformed to the following spectral relaxed maximization problem:

$$\max_{Y^T Y = I_k, Y \geq 0} \text{trace}(Y^T K Y) \tag{4.6}$$

On the other hand, spectral clustering has emerged as a robust approach for data clustering (Shi and Malik, 2000; Ng et al., 2002). Here we focus on the normalized cut for k-way clustering objective function (Gu et al., 2001; Stella and Shi, 2003). Let $G = (V, E, \tilde{K})$ be a weighted graph, where $V = \{x_1, ..., x_m\}$ is the vertex set, E the edge set, and $\tilde{K}$ the affinity matrix defined by a kernel $\tilde{\mathcal{K}}$. The k-way normalized cut spectral clustering aims to find a disjoint partition $\{V_1, ..., V_k\}$ of the vertex set V, such that:

$$\min_{V_1,...,V_k} \sum_{l=1}^{k} \text{linkratio}(V_l, \overline{V_l}) \tag{4.7}$$

where $\text{linkratio}(V_l, \overline{V_l}) = \frac{\text{links}(V_l, \overline{V_l})}{\text{degree}(V_l)} = \frac{\sum_{i \in V_l} \sum_{j \in \overline{V_l}} \tilde{K}_{ij}}{\sum_{i \in V_l} \sum_{j \in V} \tilde{K}_{ij}}$.

Following (Dhillon et al., 2004; Ding et al., 2005), the minimization in (4.7) can be casted as:

$$\max_{Z^T Z = I_k, Z \geq 0} \text{trace}(Z^T \tilde{D}^{-1/2} \tilde{K} \tilde{D}^{-1/2} Z) \tag{4.8}$$

where $\tilde{D}$ is the degree matrix of the graph G. Thus, the maximization problem in (4.8) is identical to the spectral relaxed maximization of kernel k-means clustering in (4.6) when equipped with the kernel matrix $K = \tilde{D}^{-1/2} \tilde{K} \tilde{D}^{-1/2}$.

According to the three-dimensional analysis above, we can now give the main result in the first phase of our method:

Theorem 4.1 *Spectral clustering using an affinity matrix $\tilde{K}$ is equivalent to the search for a 2-Wasserstein barycenter of $\hat{\varrho}_m = \sum_{i=1}^{m} \frac{1}{m} \delta_{\xi(x_i)}$ in the space of probability measures with support of size k, where ξ is a nonlinear mapping corresponding to the kernel matrix $K = \tilde{D}^{-1/2} \tilde{K} \tilde{D}^{-1/2}$ and $\tilde{D}$ is the degree matrix associated to $\tilde{K}$.*

In the sequel, we will refer to the search for a 2-Wasserstein barycenter of $\hat{\varrho}_m$ as Wasserstein-Spectral clustering, and we will use it to learn k hidden structures in the unlabeled target domain T.

Complexity analysis: Wasserstein-Spectral clustering offers an alternative to the popular spectral clustering algorithm of (Ng et al., 2002) that has limited applicability to large-scale problems due to its prohibitive running time that might be cubic $\mathcal{O}(m^3)$ on the size m of the input dataset (Yan et al., 2009; Tsironis et al., 2013). In fact, there are fast and efficient algorithms to perform Wasserstein-Spectral clustering as (Cuturi and Doucet, 2014), (Kroshnin et al., 2019) which is based on accelerated gradient descent with complexity proportional to m^2/ε and (Altschuler and Boix-Adsera, 2021) which can be computed in polynomial time in fixed dimension d. Furthermore, when the barycenter is restricted to measures with support of size k, the recent work of (Izzo et al., 2021) shows that randomized dimensionality reduction can be used to map the problem to a space of dimension $\mathcal{O}(\log(k))$ independent of d and that any solution found in the reduced dimension will have its cost preserved up to arbitrary small error in the original space. The algorithmic application of this statement is that one can take any approximation algorithm or heuristic for computing Wasserstein barycenter and combine it with dimensionality reduction to cope with the curse of dimensionality burden of Wasserstein barycenter.

It is noteworthy that the computation of Wasserstein barycenter is an increasingly popular problem in the machine learning and statistics communities and our algorithm can benefit from this renewed interest to reach more faster running time.

The theoretical result in Theorem 1 is confirmed by experiments, this is illustrated in Figure 4.2, where we show that Wasserstein-Spectral clustering performs identically to the traditional spectral clustering and that both are effective at separating nonlinearly separable clusters, whereas k-means fails to separate data with non-globular structures.

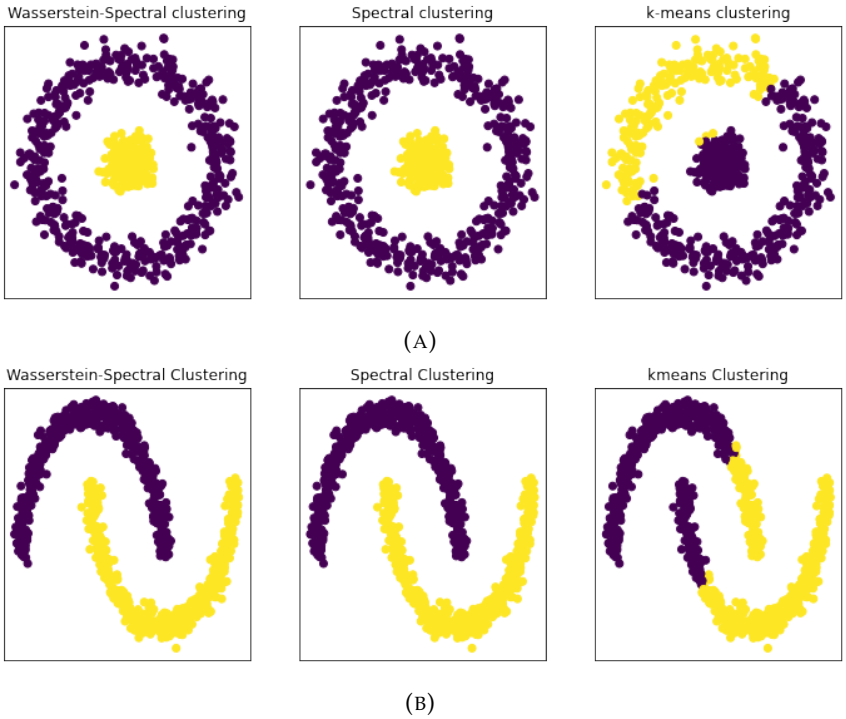

FIGURE 4.2: (a) Comparison of Wasserstein-Spectral clustering, spectral clustering, and k-means on Two-Circles dataset. (b) As for (a) but on Moons dataset.

4.3.2 Matching source and target structures through hierarchical optimal transport

Optimal transport offers a well-founded geometric way for comparing probability measures in a Lagrangian framework, and for inferring a matching between them as an inherent part of its computation. Its hierarchical formulation has inherited all these properties with the extra benefit of inducing structural information directly without the need to add any regularized term for this purpose, as well as the capability to split a sophisticated optimization surface into simpler ones that are less subject to local minima, and the ability to benefit from the entropy-regularization. Hence the key insight behind its use in the second phase of our method.

To use an appropriate formulation for hierarchical optimal transport, samples in the source domain $S = \{(x_i, y_i)\}_{i=1}^{n}$ must be partitioning according to their class labels into k classes $\{C_1, ..., C_k\}$. The empirical distributions of these structures can be expressed using discrete measures $\{\rho_1, ..., \rho_k\} \subset \mathcal{P}(\mathcal{X})$ as follows:

$$\rho_h = \sum_{i=1/x_i \in C_h}^{n} a_i \delta_{x_i}, \quad \forall h \in \{1, ..., k\} \tag{4.9}$$

Similarly, samples in the target domain $T = \{x_j\}_{j=1}^{m}$ are grouped in k clusters $\{Cl_1, ..., Cl_k\}$ using Wasserstein-Spectral clustering in the first phase. The empirical distributions of these structures can be expressed using discrete measures $\{\varrho_1, ..., \varrho_k\} \subset \mathcal{P}(\mathcal{X})$ in the following way:

$$\varrho_l = \sum_{j=1/x_j \in Cl_l}^{m} b_j \delta_{x_j}, \quad \forall l \in \{1, ..., k\} \tag{4.10}$$

Under the assumption that S and T are two sets of independent and identically distributed samples, the weights of all instances in each structure are naturally set to be equal:

$$a_i = \frac{1}{|C_h|} \quad \text{and} \quad b_j = \frac{1}{|Cl_l|}, \quad \forall h, l \in \{1, ..., k\} \tag{4.11}$$

The set S of labeled source samples and the set T of unlabeled target samples can be seen in a hierarchical paradigm as a collection of classes and clusters. Thus, the distribution of S and T can be expressed respectively as a measure of measures ϕ and φ in $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ as follows:

$$\phi = \sum_{h=1}^{k} \alpha_h \delta_{\rho_h} \quad \text{and} \quad \varphi = \sum_{l=1}^{k} \beta_l \delta_{\varrho_l} \tag{4.12}$$

where $\alpha = (\alpha_1, ..., \alpha_k)$ and $\beta = (\beta_1, ..., \beta_k)$ are vectors in the probability simplex $\sum_k$. The weights α_h and β_l are set to be equal to deal with the problem of structure imbalance, in the following way:

$$\alpha_h = \frac{1}{k} \quad \text{and} \quad \beta_l = \frac{1}{k}, \quad \forall h, l \in \{1, ..., k\} \tag{4.13}$$

To learn the correspondences between classes and clusters, we formulate an entropy-regularized hierarchical optimal transport problem between ϕ and φ in the following way:

$$(\mathcal{HOT}\text{-}\mathcal{DA}) \quad \min_{\Gamma \in U(\alpha,\beta)} \langle \Gamma, \mathcal{W} \rangle_F - \varepsilon \mathcal{H}(\Gamma) \tag{4.14}$$

where $U(\alpha, \beta) = \{\Gamma \in \mathcal{M}_k(\mathbb{R}^+) \mid \Gamma \mathbb{1}_k = \alpha \text{ and } \Gamma^{\mathbf{T}} \mathbb{1}_k = \beta\}$ represents the transportation polytope and $\mathcal{W} = (\mathcal{W}_{h,l})_{1 \leq h,l \leq k} \in \mathcal{M}_k(\mathbb{R}^+)$ stands for the Wasserstein cost matrix, whose each matrix-entry $\mathcal{W}_{h,l}$ is defined as the 2-Wasserstein distance between the measures ρ_h and ϱ_l:

$$\mathcal{W}_{h,l}^2 = \mathcal{W}_2^2(\rho_h, \varrho_l) = \langle \gamma_{h,l}^{*,\varepsilon\prime}, \mathcal{C}_{h,l} \rangle_F \tag{4.15}$$

where $\mathcal{C}_{h,l}$ is the cost matrix of pairwise squared-Euclidean distances between elements of C_h and Cl_l, and $\gamma_{h,l}^{*,\varepsilon\prime}$ is the regularized optimal transport plan between ρ_h and ϱ_l.

The optimal transport plan Γ_ε^* in (4.14) can be interpreted as a soft multivalued matching between ϕ and φ as it provides the degree of association between classes $\{C_1, ..., C_k\}$ in the source domain S and clusters $\{Cl_1, ..., Cl_k\}$ in the target domain T. Then, the one-to-one matching relationship ($\widehat{=}$) between each class C_h and its corresponding cluster Cl_l can be inferred by hard assignment from Γ_ε^*, in the following way:

$$C_h \widehat{=} Cl_l \mid l = \underset{j=1,...,k}{\operatorname{argmax}} \Gamma_\varepsilon^*(h, j), \quad \forall h \in \{1, ..., k\} \tag{4.16}$$

4.3.3 Transporting source to target structures through the barycentric mapping

Besides being a means of comparison and matching, optimal transport has the asset of performing thanks to its intrinsic quiddity of transport an alignment between source and target structures. Hence the main underlying idea of this phase.

Once the correspondence between source and target structures has been determined according to the one-to-one matching relationship ($\widehat{=}$) in (4.16), the source samples in each class C_h have to be transported to the target samples in the corresponding cluster Cl_l. This transportation can be handily expressed for each instance x_i in C_h with respect to the instances in Cl_l as the following barycentric mapping (Reich, 2013; Ferradans et al., 2014; Courty et al., 2016):

$$\widetilde{x_i} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sum_{j=1/x_j \in Cl_l}^{m} \gamma_{h,l}^{*,\varepsilon\prime}(i,j)\|x - x_j\|^2 \tag{4.17}$$

where $\widetilde{x_i}$ is the image of x_i in the region occupied by Cl_l on the target domain, and $\gamma_{h,l}^{*,\varepsilon\prime}$ is the optimal transport plan between ρ_h and ϱ_l already computed in (4.15). The barycentric mapping can be formulated for each class C_h as follows:

$$\widetilde{C_h} = diag(\gamma_{h,l}^{*,\varepsilon\prime}\mathbb{1}_{|Cl_l|})^{-1}\gamma_{h,l}^{*,\varepsilon\prime}Cl_l, \quad \forall h \in \{1,...,k\} \tag{4.18}$$

While samples in C_h and Cl_l are drawn i.i.d. from ρ_h and ϱ_l, then this mapping can be cast as a linear expression:

$$\widetilde{C_h} = |C_h|\gamma_{h,l}^{*,\varepsilon\prime}Cl_l, \quad \forall h \in \{1,...,k\} \tag{4.19}$$

After the alignment of each class C_h with its corresponding cluster Cl_l has been done as suggested in (4.19), a classifier η can be learned on the transported labeled source data $\widetilde{S} = \cup_{q=1}^{k}\widetilde{C_q}$ and evaluated on the unlabeled target data T.

The proposed HOT-DA approach is formally summarized in Algorithm 4.1:

Algorithm 4.1 HOT-DA

Input : $S = \{(x_i, y_i)\}_{i=1}^{n}, T = \{x_j\}_{j=1}^{m}$
Parameter: $\varepsilon, \varepsilon\prime$ Form $\rho_h, \varrho_l \quad \forall h, l \in \{1,...,k\}$ (4.9)(4.10)
Form ϕ, φ (4.12)
Solve the HOT-DA problem between ϕ and φ (4.14)
Get the one-to-one matching between structures (4.16)
Transport the source structures to the target ones to get $\widetilde{S}$ (4.19)
Train a classifier η on $\widetilde{S}$ and evaluate it on T
return $\{y_j\}_{j=1}^{m}$

Figure 4.3 below provides an overview of the HOT-DA approach and gives more intuitive insights on hierarchical optimal transport.

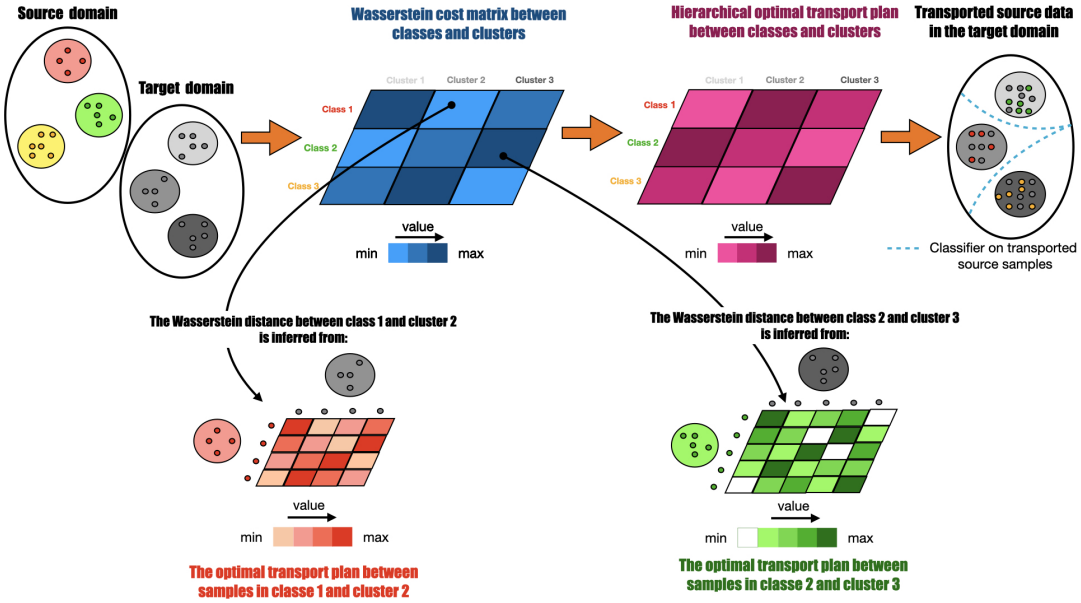

FIGURE 4.3: Wasserstein-Spectral clustering is used to learn hidden structures in the target domain as a seminal step before performing hierarchical optimal transport to align the source and target domains. The optimal plan of this hierarchical transport (in purple) is calculated from the Wasserstein cost matrix (in blue) that measures the distance between the source classes and the target clusters. The distance between each pair of structures is computed through the optimal transport plan of their points (e.g., orange and green).

4.4 Experimental results

In this section, we evaluate our method on a toy dataset and three challenging real-world visual adaptation problems.

4.4.1 Inter-twinning moons dataset

In the first experiment, we carry on moons dataset, the source domain is the classical binary two inter-twinning moons centered at the origin (0,0) and composed of 300 instances, where each class is associated to one moon of 150 samples. We consider 7 different target domains by rotating anticlockwise the source domain around its center according to 7 angles. Naturally, the greater is the angle, the harder is the adaptation. The experiments were run by setting $\varepsilon = \varepsilon\prime = 0.1$, and an SVM with a Gaussian kernel as classifier to cope with the non-linearity of this dataset. The width parameter of the SVM was chosen as $\sigma = \frac{1}{2\mathbb{V}}$, where $\mathbb{V}$ is the variance of the transported source samples. Our algorithm is compared to an SVM classifier with a Gaussian kernel trained on the source domain (without adaptation), PBDA (Germain et al., 2013) and four optimal transport based domain adaptation methods, OT-GL (Courty et al., 2016), JDOT (Courty et al., 2017), HiWA (Lee et al., 2019) and MADAOT (Dhouib et al., 2020), with the hyperparameter ranges suggested in the respective articles. To assess the generalization ability of the compared methods, they are tested on an independent set of 1000 instances that follow the same distribution as the target domain. The experiments are conducted 10 times, and the average accuracy is considered as a comparison criterion. The results are presented in Table 4.1 and the decision boundary of HOT-DA is illustrated in Figure 4.4.

TABLE 4.1: Average accuracy over moons dataset for 7 rotation angles.

Angle (°)	10°	20°	30°	40°	50°	70°	90°
SVM	**1**	0.896	0.760	0.688	0.600	0.266	0.172
PBDA	**1**	0.906	0.897	0.775	0.588	0.374	0.313
OT-GL	**1**	**1**	**1**	0.987	0.804	0.622	0.492
JDOT	0.989	0.955	0.906	0.865	0.815	0.705	0.600
HiWA	0.575	0.579	0.514	0.579	0.579	0.552	0.399
MADAOT	0.995	0.993	0.996	0.996	0.989	0.770	0.641
HOT-DA	**1**	**1**	**1**	**1**	**1**	**1**	**0.997**

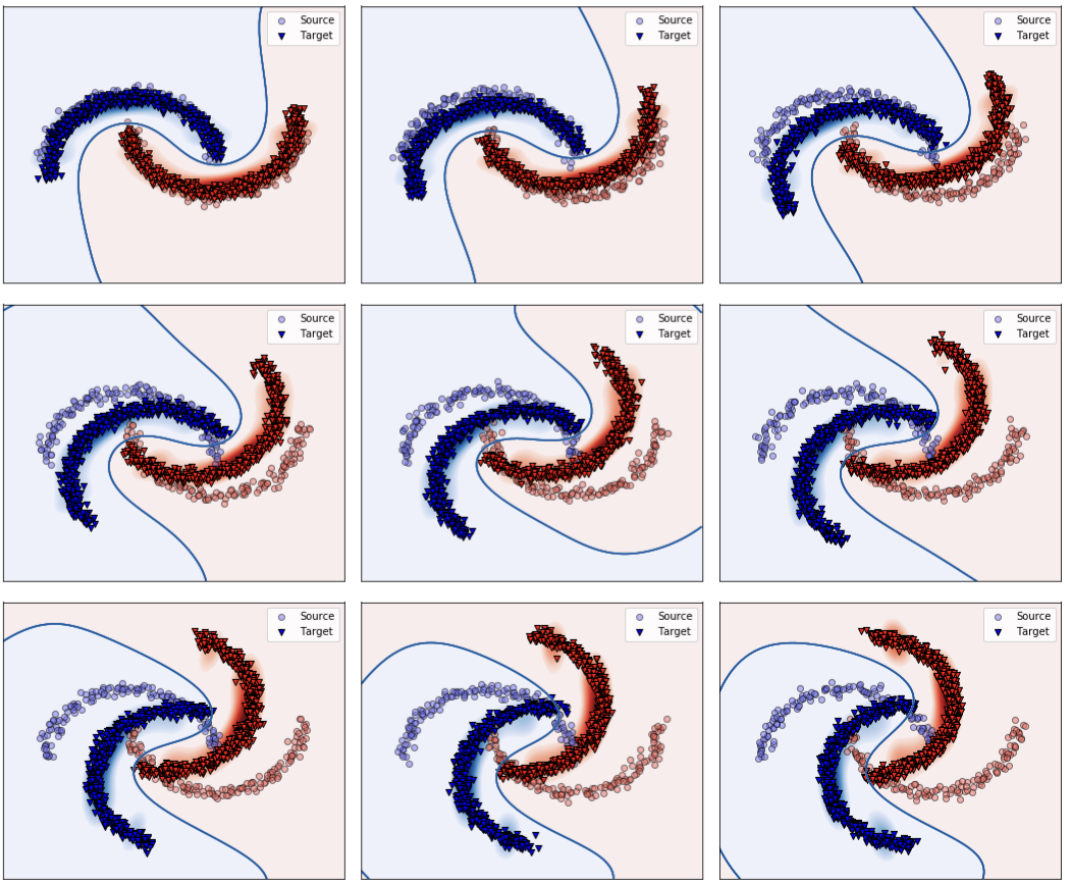

FIGURE 4.4: Illustration of the decision boundary of HOT-DA over moons problem for increasing rotation angles (10° to 90°).

We remark that all the considered algorithms based on optimal transport (except for HiWa) manage to achieve an almost perfect score on the angles from $10°$ to $40°$, which is rational, as for these small angles the adaptation problem remains quite easy. However, the SVM without adaptation has experienced a decline of almost one-third of its accuracy from $30°$. This proves that moons dataset presents a difficult adaptation problem that goes beyond the generalization ability of standard supervised learning models. For the strongest deformation, from $50°$ and up to $90°$, the proposed method HOT-DA, always provides an almost perfect score, while a big deterioration in the performance of PBDA and considerable deterioration in the performance of OT-GL and JDOT from $50°$ was observed, for MADAOT, a significant deterioration of performances starts from $70°$. In short, structures leveraged by

HOT-DA are highlighted by eliminating the increasing difficulty of this adaptation task, the constancy of the excellent performances of our approach speaks for itself, while the poor performances of HiWa, which is a multimodal distribution alignment method that seeks to jointly learn the alignment and the structure-correspondences is rather surprising, considering that this approach also relies on hierarchical optimal transport.

4.4.2 Visual adaptation datasets

We now evaluate our method on three challenging visual adaptation datasets. We start by presenting the details of these benchmark datasets, the experimental protocol, the hyper-parameter tuning, and finish by providing and discussing the obtained results.

Datasets: We consider three visual adaptation datasets: Digits (Hull, 1994; LeCun, 1998), Office-Caltech (Fei-Fei et al., 2004; Saenko et al., 2010) and Office-Home dataset (Venkateswara et al., 2017). A detailed description of each dataset is given in Table 4.2.

TABLE 4.2: Description of the visual adaptation datasets.

Dataset	Domains	#Samples	#Features	#Classes	Abbr.
Digits	USPS	1800	256	10	U
	MNIST	2000	256	10	M
Office-Caltech	Caltech	1123	4096	10	C
	Amazon	958	4096	10	A
	Webcam	295	4096	10	W
	DSLR	157	4096	10	D
Office-Home	Art	2427	2048	65	Ar
	Clipart	4365	2048	65	Cl
	Product	4439	2048	65	Pr
	Real-World	4357	2048	65	Rw

Experimental protocol: For the problem of Digits recognition, 2000 and 1800 images are randomly selected respectively from the original MNIST and USPS datasets. Then, the selected MNIST images are resized to the same 16×16 resolution as USPS ones. For the second visual adaptation problem, Office-Caltech dataset is used, where we randomly sampled a collection of 20 images per class from each domain, except for DSLR where only 8 images per class are selected. To represent these images, 4096 DeCaf6 features are used (Donahue et al., 2014). For the last problem, the more complex Office-Home dataset (Venkateswara et al., 2017) is employed. This dataset contains 15588 images from four visually very different domains: Artistic images, Clip Art, Product images, and Real-world images. For this problem, ResNet-50 was used to extract 2048 features (He et al., 2016).

As a classifier for our approach, we use 1-Nearest Neighbor classifier (1NN) on the three visual adaptation datasets, which has the advantage of being parameter free.

For the problem of Digits recognition, the comparison is conducted using 1NN classifier (without adaptation) and five domain adaptation methods, SA (Fernando et al., 2013) with a linear SVM, JDA (Long et al., 2013) with 1NN classifier, SCA (Ghifary

et al., 2016a) with 1NN classifier, OT-GL with 1NN classifier (Courty et al., 2016) and JDOT with a linear SVM (Courty et al., 2017). Concerning Office-Caltech dataset, the comparison is performed with the same competitors as for Digits in addition to DeepJDOT (Damodaran et al., 2018). Regarding the more voluminous and challenging Office-Home dataset, the choice is made to conduct the comparison with five deep learning approaches to prove the scalability of our method, and its capability to compete with deep learning models. The competitors are: ResNet-50 (without adaptation), DAN (Long et al., 2015), DANN (Ganin et al., 2016), JAN (Long et al., 2017) and DeepJDOT (Damodaran et al., 2018).

Hyper-parameter tuning: For the problem of Digits recognition, the experiments were performed by setting $\varepsilon = \varepsilon\prime = 0.1$. For Office-Caltech dataset, each target domain is equitably splited into a validation and test set. The validation set is used to select the best hyper-parameters $\varepsilon, \varepsilon\prime$ in the range of $\{1, ..., 100\}$. The accuracy is then evaluated on the test set, with the chosen hyper-parameters. The experimentation is performed 10 times, and the mean accuracy in % is reported as in (Courty et al., 2016). For Office-Home dataset, all labeled source samples and unlabeled target samples are used, and the average classification accuracy in % is computed based on three random experiments as in (Ganin and Lempitsky, 2015). The best hyper-parameters $\varepsilon, \varepsilon\prime$ are selected in the range of $\{1, ..., 100\}$.

Results: The results of our experiments are reported in Table 4.3, Table 4.4, and Table 4.5. For each task, we use bold and underlined fonts to indicate the best and second best results respectively.

TABLE 4.3: Accuracy on Digits dataset.

Task	1NN	JDA	SA	SCA	OT-GL	JDOT	**HOT-DA**
M → U	58.33	60.09	67.71	65.10	<u>69.96</u>	64.00	**76.39**
U → M	39.00	54.52	49.85	48.00	<u>57.85</u>	56.00	**63.20**
average	48.66	57.30	58.73	56.55	<u>63.90</u>	60.00	**69.79**

From Table 4.3, we can see that the proposed approach HOT-DA significantly outperforms the other domain adaptation methods on both tasks of Digits recognition problem.

Table 4.4 shows that HOT-DA surpasses the other competitors on 5 out of 12 tasks in Office-Caltech dataset, and has the second best accuracy on another task. Tables 4.3 and 4.4 also present the average results of each algorithm, where we observe a slight advance in favor of our method compared to competitors, notably JDOT and DeepJDOT. Therefore, we attribute this gain to the effectiveness of our Wasserstein-Spectral clustering that succeeds in learning hidden structures in the target domain even if they do not have globular shapes, which is the case of these two challenging visual adaptation datasets. Furthermore, the hierarchical formulation incorporates efficiently these structures, which allows the preservation of compact classes during the transportation and limits the mass splitting across different target structures. However, we see that DeepJDOT significantly outperforms HOT-DA in the three tasks where Caltech (C) is the target domain, this is explained by the difficulty we encountered to produce clusters similar to the unknown real classes in this domain.

TABLE 4.4: Accuracy on Office-Caltech dataset (Decaf6 features).

Task	1NN	JDA	SA	SCA	OT-GL	JDOT	DeepJDOT	**HOT-DA**
A → C	22.25	81.28	79.20	78.80	85.51	85.22	**87.40**	80.00
A → D	20.38	86.25	83.80	85.40	85.00	87.90	88.50	**92.53**
A → W	23.51	88.33	74.60	75.90	83.05	84.75	86.70	**96.74**
C → A	20.54	88.04	89.30	89.50	92.08	91.54	**92.30**	92.19
C → D	19.62	84.12	74.40	87.90	87.25	89.91	92.00	**96.27**
C → W	18.94	79.60	88.50	85.40	84.17	88.81	85.30	**95.11**
D → A	27.10	91.32	79.00	90.00	**92.31**	88.10	91.50	91.33
D → C	23.97	81.13	**92.25**	78.10	84.11	84.33	85.30	78.48
D → W	51.26	97.48	79.20	98.60	96.29	96.61	**98.70**	96.33
W → A	23.19	90.19	55.00	86.10	90.62	90.71	86.60	**91.86**
W → C	19.29	81.97	**99.60**	74.80	81.45	82.64	84.70	78.20
W → D	53.62	98.88	81.65	**100.00**	96.25	98.09	98.70	94.61
average	28.47	86.72	81.65	85.90	88.18	89.05	89.80	**90.30**

TABLE 4.5: Accuracy on Office-Home dataset (ResNet-50 features).

Task	ResNet-50	DAN	DANN	JAN	DeepJDOT	**HOT-DA**
Ar → Cl	34.9	43.6	45.6	45.9	**50.7**	48.0
Ar → Pr	50.0	57.0	59.3	61.2	68.6	**69.0**
Ar → Rw	58.0	67.9	70.1	68.9	74.4	**75.3**
Cl → Ar	37.4	45.8	47.0	50.4	59.9	**61.7**
Cl → Pr	41.9	56.5	58.5	59.7	**65.8**	63.2
Cl → Rw	46.2	60.4	60.9	61.0	**68.1**	67.4
Pr → Ar	38.5	44.0	46.1	45.8	**55.2**	54.1
Pr → Cl	31.2	43.6	43.7	43.4	**46.3**	39.7
Pr → Rw	60.4	67.7	68.5	70.3	73.8	**75.3**
Rw → Ar	53.9	63.1	63.2	63.9	66.0	**67.6**
Rw → Cl	41.2	51.5	51.8	52.4	**54.9**	47.9
Rw → Pr	59.9	74.3	76.8	76.8	78.3	**78.5**
average	46.1	56.3	57.6	58.3	**63.5**	62.4

The experimental results on Office-Home dataset are shown in Table 4.5. We observe that HOT-DA outperforms the other methods on 6 out of 12 tasks, while DeepJDOT performs better in the remaining 6 tasks. DeepJDOT is in the second place 6 times compared to 3 times for HOT-DA, which experienced a drop in performance in the 3 tasks where Clipart is the target domain. This behavior led to a slight difference in their average accuracy on Office-Home dataset in favor of DeepJDOT. This is rather surprising considering that the competitors rely on neural networks to learn the final classifier and these latter are expected to have higher discriminative power than the 1-Nearest Neighbor classifier used in our approach. Consequently, we attribute this competitiveness to the efficiency of our hierarchical optimal transport formulation that manages to better align the two distributions, and that can be seen as an "implicit regularized" optimal transport. This implicit regularization heavily relies on "a priori knowledge" (clustering), which leads to the injection of structural information directly into the transport problem.

Globally, the mean accuracy of HOT-DA is 0.5% higher than DeepJDOT on Office-Caltech. In parallel, DeepJDOT shows an improvement of 1.1% compared to our method on Office-Home. Roughly speaking, the set of experiments shows good behavior with respect to state-of-the-art methods, especially JDOT and DeepJDOT, which however manage to outperform our algorithm on several tasks. This competitive behavior is, we believe, due to the commonality between JDOT and DeepJDOT on the one hand and HOT-DA on the other hand. The former methods design a simultaneous optimization problem to find the coupling between the joint distribution of the source and target domains and the labeling function that solves the transfer problem. While the second method tries to address the same task sequentially by first finding the target structures, which is equivalent to performing a pseudo-labeling in the target domain, before aligning each source structure with its corresponding target structure, which can be seen as an alignment of the joint distributions.

4.4.3 Relevance of Wasserstein-Spectral clustering to HOT-DA

The first step of HOT-DA is not directly integrated into the domain adaptation process, and it is questionable whether other well-known clustering algorithms such as k-means (MacQueen et al., 1967), DBSCAN (Ester et al., 1996) or HDBSCAN (Campello et al., 2013) can be used to learn the target structures instead of Wasserstein-Spectral clustering (W-SC).

k-means suffers from several drawbacks, notably its inability to identify clusters with non-convex shapes, as shown in Figure 4.2. This incapacity can significantly reduce the performance of HOT-DA on several unsupervised domain adaptation problems where clusters do not have globular shapes in the target domain. These problems include but are not limited to, the inter-twinning moons dataset.

On the other hand, DBSCAN relies on detecting areas where points are closely packed together (points with many nearby neighbors) and marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN does not require specifying the number of clusters a priori, instead, it requires two parameters: minimum number of neighbors $minpts$ and minimum radius Eps. Therefore, for clustering high-dimensional data, it becomes very difficult to tune these parameters to get the desired number of clusters, even using heuristic methods (Musdholifah et al., 2013). Which can lead to finding a number of clusters very larger or very smaller than the number of classes k in the source domain, and then to poor adaptation results. Regarding HDBSCAN, which is a conversion of DBSCAN into a hierarchical clustering algorithm, from which a simplified hierarchy composed only of the most significant clusters can be easily extracted. It can find clusters of varying densities, unlike DBSCAN and it performs well on low to medium dimensional data. However, its performance tends to decrease as the dimension increases. In general, the performance of HDBSCAN can see significant decreases already with tens of dimensions (Campello et al., 2020). The unsupervised domain adaptation settings can be beneficial for clustering algorithms that require the number of clusters k to the detriment of DBSCAN and HDBSCAN which do not benefit from this available information, especially for high-dimensional data (e.g., visual domain adaptation datasets using ResNet-50 or DeCaf features) where it becomes quite difficult to tune these parameters to get the desired number of clusters k.

This analysis is the main motivation behind replacing k-means, DBSCAN, or HDBSCAN with spectral clustering which is able to find exactly k clusters, even with non-globular shapes. This choice was reconsidered for complexity reasons as discussed in 4.3.1, which led to the establishment of an equivalent algorithm: Wasserstein-Spectral clustering, which furthermore allows unifying the different steps of our algorithm under the aegis of optimal transport.

To confirm the insights above, we reproduce the experiments on the following datasets: Moons, Office-Caltech, and Office-Home using four variants of our algorithm, the first one uses Wasserstein-Spectral clustering, the second one uses k-means, the third one is based on DBSCAN and the fourth one is rather based on HDBSCAN. The results of these experiments are given in Figure 4.5 using Kiviat diagram.

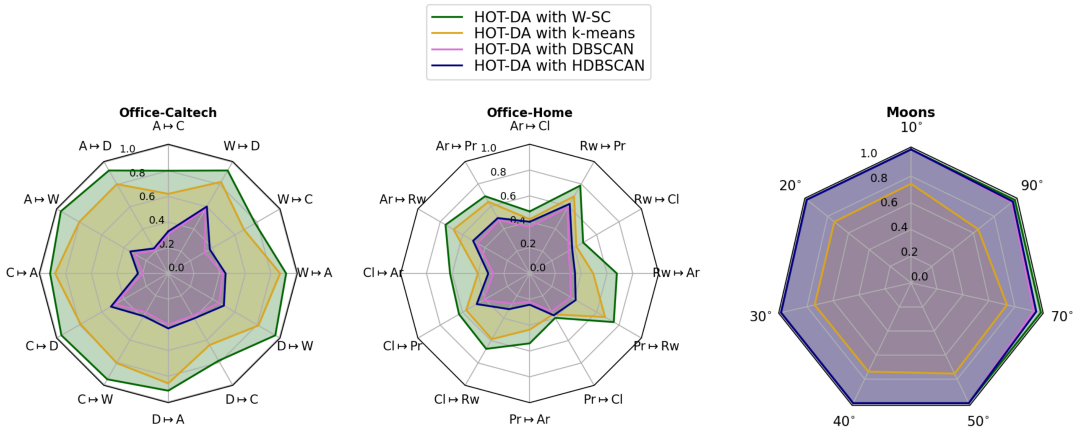

FIGURE 4.5: Kiviat's accuracy diagram for the four variants of HOT-DA on Office-Caltech, Office-Home, and Moons datasets. The radar corresponding to the variant based on Wasserstein-Spectral clustering dominates the other radars on the three datasets

Figure 4.5 indicates ostensibly that the radar corresponding to the variant with W-SC encompasses the other radars on the three datasets. On the 7 rotation problems of moons dataset, the variant of HOT-DA based on Wasserstein-Spectral clustering performs slightly better than the other variants based on DBSCAN and HDBSCAN and all manage to make a nearly perfect adaptation. This is due to the ability of Wasserstein-Spectral clustering to capture the structure of the two moons, and the ease of tuning the parameters for DBSCAN and HDBSCAN to find the desired number of clusters in a small dimensional space (d = 2). While the variant of HOT-DA based on k-means has much poorer performance due to the inability of k-means to correctly explore the two inter-twinning moons. Regarding Office-Caltech and Office-Home, the high-dimensionality of these datasets (d = 4096 for Office-Caltech and d=2048 for Office-Home) has strongly impacted the performance of DBSCAN and HDBSCAN, which fail to find exactly the desired number of clusters (k = 10 for Office-Caltech and k = 65 for Office-Home), while k-means and Wasserstein-Spectral clustering benefit from this available information to obtain better results, with significant supremacy for this latter.

The above empirical experiments strengthen our choice of Wasserstein-Spectral clustering and clearly demonstrate that it is a well-suited candidate for these unsupervised domain adaptation settings.

4.4.4 Structure imbalance sensitivity analysis

The problem of structure imbalance where an uneven distribution of samples occurs among a variety of structures can lead to pathological behavior of the mass transportation, by showing favoritism towards majority target structures in spite of minority ones which may receive no mass due to the thresholding performed in (4.16). Fortunately, the choice made to give the same mass to each structure, allows HOT-DA to avoid this behavior and to achieve the right matching between source and target structures. The intuition behind this choice is to consider each structure as an independent entity and to remove the bias induced by its cardinality, which is quite natural since a class in the source domain and its corresponding cluster in the target domain do not necessarily have the same proportion of points.

To evaluate the behavior of HOT-DA with respect to the problem of structure imbalance, an experiment is conducted on a toy dataset composed of two structures in each domain as shown in Figure 4.6. The experiment is designed to compare the performance of our proposed approach with Reg-OT (Cuturi, 2013), OT-GL (Courty et al., 2016) and GW (Gromov-Wasserstein) (Mémoli, 2011; Sturm, 2006), in three scenarios: balanced structures, moderately unbalanced structures and, extremely unbalanced structures.

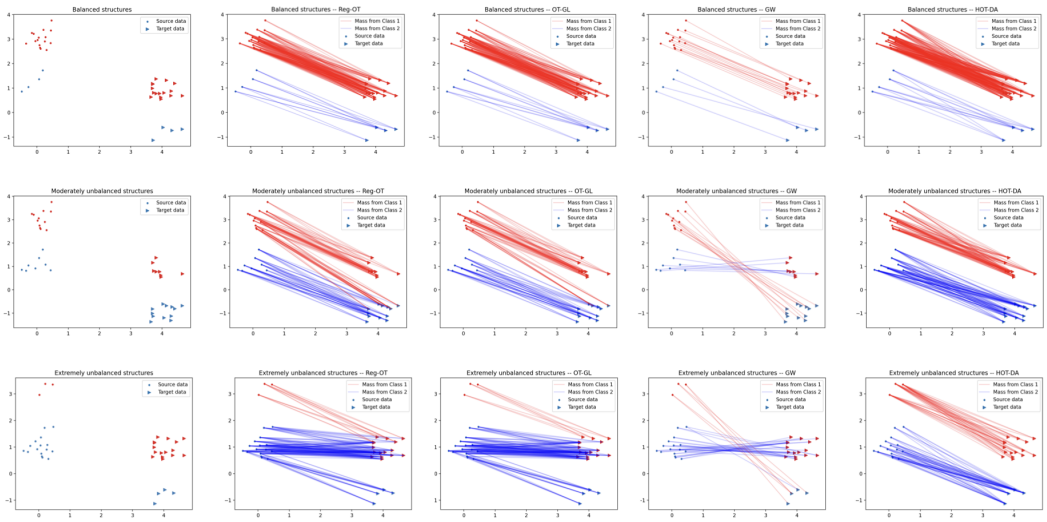

FIGURE 4.6: Behavior of Reg-OT, OT-GL, GW, and, HOT-DA towards the problem of structure imbalance.

The first part of the experiment concerning the case of balanced structures shows an ideal behavior of the four methods. The situation begins to change slightly in the second case of moderately unbalanced structures, where Reg-OT and OT-GL make some mistakes because of the extra-mass of the red source structure that has to be sent to the blue target structure, while GW reverses the matching due to this moderate imbalance. However, our approach still achieves an uncontested matching. The third part concerning the most complicated scenario of extremely unbalanced structures, demonstrates a catastrophic deterioration in the results of the three methods Reg-OT, OT-GL, and GW, while our HOT-DA approach continues to provide a flawless result. This proves that HOT-DA is a robust and non-sensitive algorithm to this kind of imbalance, unlike other approaches. It is noteworthy that our model is less

sensitive than other optimal transport methods to changes in the value of the entropy regularization parameter thanks to the thresholding carried out by the hard assignment in (4.16).

4.5 Software

We make our code and the used datasets publicly available at:

https://github.com/MouradElHamri/HOT-DA

4.6 Conclusion and future perspectives

In this chapter, we proposed HOT-DA, a novel approach dealing with unsupervised domain adaptation, by leveraging the ability of hierarchical optimal transport to induce structural information directly into the transportation process. We also proved theoretically the equivalence between spectral clustering and the problem of learning probability measures through Wasserstein barycenter, this latter was used to derive Wasserstein-Spectral clustering, a new alternative of spectral clustering able to learn hidden structures of arbitrary shapes in the unlabeled target domain, as a seminal step before performing hierarchical optimal transport to align the source and target domains. The proposed approach has been shown to be efficient on both simulated and real-world problems compared to several state-of-the-art methods, in addition to being able to cope with structure imbalance.

The work of this chapter can be extended in different directions:

- From an algorithmic standpoint, we plan to take advantage of the unification of the different steps of our approach under the banner of optimal transport, in order to jointly learn the target structures and the optimal transport plan that aligns them with the source classes and to investigate a possible application of the proposed approach to multi-source domain adaptation settings.
- From a theoretical standpoint, future work will include the development of generalization bounds that take into account the hierarchical organization of source and target samples in structures. These bounds will reflect explicitly both the excess clustering risk in the target domain and which structures must be aligned to lead to a good adaptation.

CHAPTER 5

THEORETICAL GUARANTEES WITH HIERARCHICAL OPTIMAL TRANSPORT

Contents

Recent theoretical advances show that the success of domain adaptation algorithms heavily relies on their ability to minimize the divergence between the probability distributions of the source and target domains. However, minimizing this divergence cannot be done independently of the minimization of other key ingredients such as the source risk or the combined error of the ideal joint hypothesis. The trade-off between these terms is often ensured by algorithmic solutions that remain implicit and not directly reflected by the theoretical guarantees. To get to the bottom of this issue, we propose in this chapter based on (El Hamri et al., 2022d) a new theoretical framework of domain adaptation through hierarchical optimal transport. This framework provides more explicit generalization bounds and allows us to consider the natural hierarchical organization of samples in both domains into classes or clusters. Additionally, we provide a new divergence measure between the source and target domains called Hierarchical Wasserstein distance that indicates under mild assumptions, which structures have to be aligned to lead to a successful adaptation.

> And a voice shouts in my heart: We fell,
> and a voice shouts in my heart: Get up!
>
> Mourid Al-Barghouti

5.1 Introduction

In domain adaptation theory, existing generalization bounds on the target risk of a given hypothesis are often stated in a generic form implying the source risk, a divergence measure between the source and target domains, and a term assessing the ability of the given hypothesis space to successfully resolve the problem of adaptation (3.29). The source risk is estimable from finite samples and can be minimized by learning the hypothesis from the available source labeled data. Similarly, the divergence is estimable from the observed data and is intended to be slight if the two domains are nearby. While the last term is non-estimable and is usually formulated as the combined error of the ideal joint hypothesis.

In the pioneering theoretical work of (Ben-David et al., 2006), the domain adaptation problem was carefully addressed using the total variation TV distance, but its employment as a divergence measure between the marginal distributions of the source and target domains presents two major weaknesses. First, the TV distance is not directly related to the concerned hypothesis space, which results in loose generalization bounds, and secondly, it is not estimable from finite samples drawn from arbitrary probability distributions (Batu et al., 2000). To overcome these limitations, (Ben-David et al., 2010) introduced a classifier-induced divergence called the $\mathcal{H}\Delta\mathcal{H}$-divergence, based on the $\mathcal{A}$-divergence provided in (Kifer et al., 2004). Indeed, the $\mathcal{H}\Delta\mathcal{H}$-divergence explicitly considers the given hypothesis space $\mathcal{H}$, which guarantees that the generalization bounds stay relevant and decidedly linked to the learning problem in question, and, for a given hypothesis space $\mathcal{H}$ of finite Vapnik-Chervonenkis dimension, the $\mathcal{H}\Delta\mathcal{H}$-divergence can be estimated from finite samples. Furthermore, the $\mathcal{H}\Delta\mathcal{H}$-divergence is always smaller than the TV distance for any hypothesis space $\mathcal{H}$, which results in tighter bounds. Nevertheless, an obvious shortcoming of the $\mathcal{H}\Delta\mathcal{H}$-divergence is its reliance on the 0 - 1 loss function. Whereas, it might be desirable to have generalization bounds for a more generic domain adaptation framework, where any arbitrary loss function with some suitable properties can be considered. To address this concern, (Mansour et al., 2009) introduced the discrepancy distance $disc_l$ that expands the previous theoretical analysis of domain adaptation for any arbitrary loss function, which is symmetric, bounded, and obeys the triangle inequality. Additionally, the discrepancy distance $disc_l$ relies on the hypothesis space $\mathcal{H}$, but the complexity term is rather related to the Rademacher complexity of $\mathcal{H}$. This distinctive refinement provides data-dependent bounds that are commonly sharper than those derived from Vapnik–Chervonenkis theory.

Despite their numerous advantages, both the $\mathcal{H}\Delta\mathcal{H}$-divergence and the discrepancy distance $disc_l$ suffer from a computational burden related to their estimation. In such a circumstance, it was natural to look for other metrics with some appealing computational properties to quantify the divergence between the two domains. Following

this trend, (Redko, 2015) appealed to the Maximum Mean Discrepancy (MMD) distance to infer generalization bounds analogous to that of (Ben-David et al., 2010). These bounds turned out to be remarkably meaningful since an unbiased estimator of the squared MMD distance can be computed in linear time, and the complexity term does not depend on the Vapnik–Chervonenkis dimension but on the empirical Rademacher complexities of the hypothesis space with respect to the source and target samples. Some time later, (Redko et al., 2017) presented generalization bounds in terms of the Wasserstein distance $\mathcal{W}_1$ as a theoretical analysis of the seminal domain adaptation algorithm based on optimal transport (Courty et al., 2016). This analysis proved to be very fruitful for several reasons. First, the Wasserstein distance is computationally attractive, particularly in virtue of the entropic regularization introduced in (Cuturi, 2013). Furthermore, the Wasserstein distance has the ability to capture the underlying geometry of the data in both domains. Moreover, the Wasserstein distance is quite strong, and according to (Villani, 2009), it is not so hard to associate the convergence information in the Wasserstein distance with certain smoothness bound to obtain convergence in stronger distances. This powerful asset of the Wasserstein distance gives tighter bounds compared to other results in state-of-the-art.

Under the above generic form of generalization bounds, it is clear that minimizing the previous distances between the marginal distributions of the source and target domains cannot be performed separately from minimizing the source risk and the ability term. For instance, the minimization of the Wasserstein distance results from the transport of the source to the target samples such that $\mathcal{W}_1$ becomes quite low when computing between the newly transported source samples and the target instances. Nevertheless, by minimizing the Wasserstein distance only, the obtained transformation may transport some source samples of different labels to the same target samples, and thus, the empirical source error cannot be adequately minimized. Moreover, the joint error will be negatively affected since no classifier will be capable of separating these source instances. We may also consider an ironically extreme situation of binary classification task where the transport plan sends the source data of each class to the target data of the inverse class. In such a case, the joint error will be drastically impacted. To avoid these pathological scenarios, a possible remedy was then to promote group sparsity in the optimal transport plan in order to restrict the source instances of different classes to be transported to the same target points. This algorithmic solution is implemented through a group-norm regularizer in (Courty et al., 2016). From a theoretical point of view, this regularization constitutes an arrangement to control the trade-off between the three terms of the bound. However, this trade-off remains imperceptible, and the bound does not reflect it explicitly.

Recently, (El Hamri et al., 2022b) [1] proposed a new domain adaptation algorithm based on a hierarchical formulation of optimal transport that leverages beyond the geometrical information captured by the ground metric, richer structural information in the source and target domains. The exploitation of this structural information elicited some desired properties in domain adaptation like preserving compact classes during the transportation, which provided an alternative algorithmic solution to restrict the source instances of different classes to be transported to the same

[1] See the previous chapter.

target points. The main underlying idea behind the hierarchical formulation of optimal transport is to organize samples in the source domain into structures according to their class labels, and samples in the target domain into structures by clustering. This organization offers a new paradigm of perceiving each domain as a measure of measures. Rigorously, each domain can be seen as a distribution over structures, where the structures are also distributions, but over samples. Hierarchical optimal transport attempts then to align the structures of both domains while minimizing the total cost of the transportation quantified by the Wasserstein distance, which acts as the ground metric. While presenting very interesting empirical performances, it turns out that the work of (El Hamri et al., 2022b) has no theoretical guarantees, despite it may be an untapped potential solution to avoid the limitations listed above, specifically the one concerning the imperceptible trade-off between the three terms of the bound.

Contributions: In this chapter, we address the aforementioned limitations by providing new generalization bounds based on hierarchical optimal transport. The main underlying idea behind these bounds is to decompose the two domains into structures and then indicate explicitly which structures should be aligned together to lead to a good adaptation.

This paper's contributions are threefold:

1. We provide a theoretical analysis of the work of (El Hamri et al., 2022b), which justifies the use of hierarchical optimal transport for domain adaptation.

2. We consider the usual hierarchical organization of data into structures and introduce a new divergence measure to quantify the similarity between source and target domains in light of this hierarchy, which we call the Hierarchical Wasserstein distance. We relate the proposed distance to the classical Wasserstein distance.

3. We derive generalization bounds on the target risk based on the Hierarchical Wasserstein distance, for the three domain adaptation settings: unsupervised, semi-supervised, and multi-source domain adaptation. The proposed generalization bounds indicate the distance between which structures should be really minimized to lead to a good adaptation. This makes the trade-off between the three terms of the bound more explicit and may suggest the minimization of each term independently of the others.

Outline: The rest of this paper is organized as follows. The 2^{nd} section introduces the Hierarchical Wasserstein distance as a divergence measure between the source and target domains and introduces the link with the classical Wasserstein distance. The 3^{rd} section proves generalization bounds based on the Hierarchical Wasserstein distance for three scenarios, unsupervised, semi-supervised, and multi-source domain adaptation. Finally, we discuss conclusions and future research directions in section 4.

5.2 Hierarchical Wasserstein distance

This section is dedicated to constructing the Hierarchical Wasserstein distance $\mathcal{HW}_p$ on the space $\mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ that will serve as a divergence measure to quantify the closeness between the source and target domains.

The Hierarchical Wasserstein distance will allow us to introduce several generalization bounds on the target risk in the next section where we will study three scenarios of domain adaptation as we will see later.

First, let us present the following Theorem from (Villani, 2009) that will be useful for establishing the $\mathcal{HW}_p$ distance.

Theorem 5.1 (Topology of the Wasserstein space) *Let $(\mathcal{X}, d)$ be a Polish metric space and let $p \in [1, \infty[$. Then the Wasserstein space $\mathcal{P}_p(\mathcal{X})$ metrized by the Wasserstein distance $\mathcal{W}_p$ is itself a Polish metric space.*

In the following Lemma we show that $\mathcal{HW}_p$ is effectively a distance on the space $\mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$.

Lemma 5.2 (Hierarchical Wasserstein distance) *Let $(\mathcal{X}, d)$ be a Polish metric space and let $p \in [1, \infty[$. For any two probability measures $\phi, \varphi \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$, the Hierarchical Wasserstein distance of order p between ϕ and φ is defined by:*

$$\mathcal{HW}_p(\phi, \varphi) = \left(\inf_{\eta \in \Pi(\phi, \varphi)} \int_{\mathcal{P}_p(\mathcal{X})^2} \mathcal{W}_p(\rho, \varrho)^p \, d\eta(\rho, \varrho) \right)^{1/p}. \tag{5.1}$$

Furthermore $\mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ metrized by $\mathcal{HW}_p$ is a Polish metric space.

Proof Let $p \in [1, \infty[$. Since $(\mathcal{X}, d)$ is a Polish metric space, then $(\mathcal{P}_p(\mathcal{X}), \mathcal{W}_p)$ is itself a Polish metric space by Theorem 5.1. By a recursion of concepts, Definition 2.18 ensures that $\mathcal{HW}_p$ defines a distance on the space $\mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ and Theorem 5.1 holds that $(\mathcal{P}_p(\mathcal{P}_p(\mathcal{X})), \mathcal{HW}_p)$ is a Polish metric space. ■

The following Corollary from (Villani, 2009) is of particular interest for the statement of the link between $\mathcal{W}_p$ and $\mathcal{HW}_p$.

Corollary 5.3 (Measurable selection of optimal plans) *Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function, inf $c > -\infty$. Let Υ be a measurable space and let $\upsilon \mapsto (\mu_\upsilon, \nu_\upsilon)$ be a measurable function $\Upsilon \to \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$. Then there is a measurable choice $\upsilon \mapsto \pi_\upsilon$ such that for each υ, π_υ is an optimal transport plan between μ_υ and ν_υ.*

In the following Lemma, we prove that the Wasserstein distance enjoys an interesting property that we call hierarchical monotonicity.

Lemma 5.4 (Hierarchical monotonicity of Wasserstein distance) *Let $(\mathcal{X}, d)$ be a Polish metric space and let $p \in [1, \infty[$. Let $\phi, \varphi \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ and let $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ such that $\mu = \int_{\mathcal{P}_p(\mathcal{X})} X d\phi$ and $\nu = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi$ for some generic measure-valued random variable X. The following holds,*

$$\mathcal{W}_p(\mu, \nu) \leq \mathcal{HW}_p(\phi, \varphi). \tag{5.2}$$

Proof Let $\phi, \varphi \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ and let consider an arbitrary $\eta \in \Pi(\phi, \varphi)$, then:

$$\begin{aligned}
\int_{\mathcal{P}_p(\mathcal{X})^2} \mathcal{W}_p^p(\rho, \varrho) d\eta(\rho, \varrho) &= \int_{\mathcal{P}_p(\mathcal{X})^2} \left(\int_{\mathcal{X}^2} d(x, y)^p \pi_{\rho,\varrho}(dx, dy) \right) d\eta(\rho, \varrho) && (5.3)\\
&= \int_{\mathcal{X}^2} \left(\int_{\mathcal{P}_p(\mathcal{X})^2} d(x, y)^p \pi_{\rho,\varrho} d\eta(\rho, \varrho) \right) (dx, dy) && (5.4)\\
&= \int_{\mathcal{X}^2} d(x, y)^p \left(\int_{\mathcal{P}_p(\mathcal{X})^2} \pi_{\rho,\varrho} d\eta(\rho, \varrho) \right) (dx, dy) && (5.5)\\
&= \int_{\mathcal{X}^2} d(x, y)^p \pi(dx, dy) && (5.6)\\
&\geq \mathcal{W}_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} d(x, y)^p \, \pi(dx, dy) && (5.7)
\end{aligned}$$

First line is obtained by the definition of the Wasserstein distance and by using the measurable selection of optimal plans, so $\pi_{\rho,\varrho}$ is an optimal transport plan between ρ and ϱ that is chosen in a measurable way according to Corollary 5.3. Second line is due to Fubini's theorem. Third line is trivial. Fourth line follows from the fact that $\int_{\mathcal{P}_p(\mathcal{X})^2} \pi_{\rho,\varrho} d\eta(\rho, \varrho) = \pi$ for some valid transport plan $\pi \in \Pi(\mu, \nu)$, this becomes clear by marginalizing out y and marginalizing out x, respectively:

$$\begin{aligned}
\forall \mathcal{A} \subset \mathcal{X} : \int_{\mathcal{P}_p(\mathcal{X})^2} \pi_{\rho,\varrho}(\mathcal{A} \times \mathcal{X}) d\eta(\rho, \varrho) &= \int_{\mathcal{P}_p(\mathcal{X})^2} \rho(\mathcal{A}) d\eta(\rho, \varrho) && (5.8)\\
&= \int_{\mathcal{P}_p(\mathcal{X})} \rho(\mathcal{A}) d\phi && (5.9)\\
&= \mu(\mathcal{A}) && (5.10)\\
\forall \mathcal{B} \subset \mathcal{X} : \int_{\mathcal{P}_p(\mathcal{X})^2} \pi_{\rho,\varrho}(\mathcal{X} \times \mathcal{B}) d\eta(\rho, \varrho) &= \int_{\mathcal{P}_p(\mathcal{X})^2} \varrho(\mathcal{B}) d\eta(\rho, \varrho) && (5.11)\\
&= \int_{\mathcal{P}_p(\mathcal{X})} \varrho(\mathcal{B}) d\varphi && (5.12)\\
&= \nu(\mathcal{B}) && (5.13)
\end{aligned}$$

The first equalities (5.8) and (5.11) follow from the fact that $\pi_{\rho,\varrho}$ is an optimal transport plan between ρ and ϱ. Second equalities (5.9) and (5.12) follow from the fact that η is an optimal transport plan between ϕ and φ. Third equalities (5.10) and (5.13) follow from the assumptions made on ϕ and φ, respectively.

Let's get back to the core of the proof, inequality in the fifth line follows from the definition of the Wasserstein distance.

The inequality $\int_{\mathcal{P}_p(\mathcal{X})^2} \mathcal{W}_p^p(\rho, \varrho) d\eta(\rho, \varrho) \geq \mathcal{W}_p^p(\mu, \nu)$ holds for any $\eta \in \Pi(\phi, \varphi)$, then, we obtain the final result by taking the infimum over η from the left-hand side, i.e.

$$\inf_{\eta \in \Pi(\phi,\varphi)} \int_{\mathcal{P}_p(\mathcal{X})^2} \mathcal{W}_p^p(\rho, \varrho) d\eta(\rho, \varrho) \geq \mathcal{W}_p^p(\mu, \nu) \tag{5.14}$$

which gives:

$$\mathcal{HW}_p(\phi, \varphi) \geq \mathcal{W}_p(\mu, \nu) \tag{5.15}$$

■

5.3 Generalization bounds based on the Hierarchical Wasserstein distance

In this section, we introduce generalization bounds on the target risk when the divergence between the source and target domains is measured by the Hierarchical Wasserstein distance.

5.3.1 A bound for unsupervised domain adaptation

This subsection focuses on unsupervised domain adaptation where no labeled data are available in the target domain. We first present the Lemma that introduces Hierarchical Wasserstein distance to relate the source and target risks for an arbitrary pair of hypothesis.

Lemma 5.5 *Let $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{X})$ be two probability measures on a compact $\mathcal{X} \subseteq \mathbb{R}^{\mathrm{d}}$ and let $\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ be two probability measures on $\mathcal{P}_p(\mathcal{X})$ such that $\mu_{\mathcal{S}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{S}}$ and $\mu_{\mathcal{T}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{T}}$ for some generic measure-valued random variable X. Assume that the cost function $c(x, y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_k}$, where $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS) equipped with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ induced by $\phi : \mathcal{X} \to \mathcal{H}_k$ and $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_k}$. Assume further that the loss function $l_{h,f} : x \mapsto l(h(x), f(x))$ is convex, symmetric, bounded, obeys triangle equality, and has the parametric form $|h(x) - f(x)|^q$ for some $q > 0$. Assume also that the kernel k in the RKHS $\mathcal{H}_k$ is square-root integrable w.r.t. both $\mu_{\mathcal{S}}, \mu_{\mathcal{T}}$ for all $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{X})$ where $0 \leq k(x, y) \leq K, \forall x, y \in \mathcal{X}$. If $\|l\|_{\mathcal{H}_k} \leq 1$, then the following holds:*

$$\forall (h, h') \in \mathcal{H}_k^2, \quad \epsilon_{\mathcal{T}}(h, h') \leq \epsilon_{\mathcal{S}}(h, h') + \mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}). \tag{5.16}$$

Proof Under assumptions of Lemma 5.5, and according to Lemma 3.42, we have:

$$\forall (h, h') \in \mathcal{H}_k^2, \quad \epsilon_{\mathcal{T}}(h, h') \leq \epsilon_{\mathcal{S}}(h, h') + \mathcal{W}_1(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) \tag{5.17}$$

On the other hand, using the property of Hierarchical monotonicity of Wasserstein distance in Lemma 5.4 for $p = 1$, we have:

$$\mathcal{W}_1(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) \leq \mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) \tag{5.18}$$

which gives:

$$\forall (h, h') \in \mathcal{H}_k^2, \quad \epsilon_{\mathcal{T}}(h, h') \leq \epsilon_{\mathcal{S}}(h, h') + \mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) \tag{5.19}$$

■

Remark 5.6 *Lemma 5.5 and the subsequent results are established for the special case $p = 1$, but they can easily be generalized for any $p > 1$, by applying Hölder's inequality that states:*

$$p \leq q \Rightarrow \mathcal{HW}_p \leq \mathcal{HW}_q \tag{5.20}$$

Remark 5.7 *As reported in (Redko et al., 2017), the parametric form of the loss function $l_{h,f}$ as $|h(x) - f(x)|^q$ for some $q > 0$ is only an example. Following (Saitoh, 1997), we can also look at more general nonlinear transformations of h and f that satisfy the hypothesis made on $l_{h,f}$ above. These transformations can comprise a product of hypothesis and labeling functions and thus the suggested results are relevant for hinge loss too.*

Remark 5.8 *Lemma 5.5 supposes that the cost function $c(x,y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_k}$. This may seem too demanding as in several applications, the Euclidean distance $c(x,y) = \|x-y\|$ is considered as the ground metric. But fortunately, this assumption is not that restrictive and may be bypassed through the duality between RKHS and distance-based metric representations, studied by (Sejdinovic et al., 2013). In fact:*

$$\|\phi(x) - \phi(y)\|_{\mathcal{H}_k} = \sqrt{\langle \phi(x) - \phi(y), \phi(x) - \phi(y) \rangle_{\mathcal{H}_k}} = \sqrt{k(x,x) - 2k(x,y) + k(y,y)}. \quad (5.21)$$

Thus, the Euclidean distance can be recovered by considering the kernel provided by the covariance function of the fractional Brownian motion:

$$k(x,y) = \frac{1}{2}\left(\|x\|^2 - \|x-y\|^2 + \|y\|^2\right) \quad (5.22)$$

We report now some preliminary results to show the convergence of an empirical measure to its true associated measure with respect to the Wasserstein distance. These results can be extended to the Hierarchical Wasserstein distance, which allows to provide generalization bounds for finite samples rather than true population measures. First, let's define Talagrand inequalities T_p as in (Villani, 2009).

Definition 5.9 (T_p inequality) *Let $(\mathcal{X}, d)$ be a Polish metric space and let $p \in [1, \infty[$. Let ν be a reference probability measure in $\mathcal{P}_p(\mathcal{X})$ and let $\zeta > 0$. It is said that ν satisfies $T_p(\zeta)$ inequality if:*

$$\forall \mu \in \mathcal{P}_p(\mathcal{X}) \quad \mathcal{W}_p(\nu, \mu) \leq \sqrt{\frac{2H(\nu|\mu)}{\zeta}} \quad (5.23)$$

where H is the relative entropy: $H(\nu|\mu) = \int \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} d\mu$.

We shall say that ν satisfies a T_p inequality if it satisfies $T_p(\zeta)$ for some constant $\zeta > 0$.

Probability measures verifying T_1 inequality have a characteristic property related to the existence of a square-exponential moment, as shown in (Bolley and Villani, 2005).

Theorem 5.10 (Characteristic property of T_1 inequality) *Let $\mathcal{X}$ be a measurable space equipped with a measurable distance d, let ν be a reference probability measure on $\mathcal{X}$, and let x_0 be any element of $\mathcal{X}$. Then ν satisfies T_1 inequality if and only if, for some $\alpha > 0$:*

$$\int_{\mathcal{X}} e^{\alpha d(x_0,x)^2} d\nu(x) < +\infty, \quad (5.24)$$

In (Bolley et al., 2007), authors assume a T_p inequality for the measure μ, and derive an upper bound in $\mathcal{W}_p$ distance, we present here the case $p = 1$.

Theorem 5.11 (Upper bound in $\mathcal{W}_1$) *Let $(\mathcal{X}, d)$ be a Polish metric space. Let μ be a probability measure on $\mathcal{X}$ so that for some $\alpha > 0$, we have for any $x_0 \in \mathcal{X}$, $\int_{\mathcal{X}} e^{\alpha d(x_0,x)^2} d\mu(x) < +\infty$, and let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be its associated empirical measure defined on a sample of independent variables $\{x_i\}_{i=1}^n$ all distributed according to μ. Then for any $\mathrm{d}' > dim(\mathcal{X})$ and $\zeta' < \zeta$, there exists some constant N_0 depending on d', ζ' and some square exponential moment of μ, such that for any $\varepsilon > 0$ and $N \geq N_0 \max(\varepsilon^{-(\mathrm{d}'+2)}, 1)$*

$$\mathbb{P}\left[\mathcal{W}_1(\mu, \hat{\mu}) > \varepsilon\right] \leq \exp\left(\frac{-\zeta'}{2} N \varepsilon^2\right). \quad (5.25)$$

Remark 5.12 *The original version of Theorem 5.11 is established for $\mathcal{X} = \mathbb{R}^{\mathrm{d}}$ as stated in Theorem 3.41, but we can find the generalization above for any metric space $(\mathcal{X}, d)$ in (Courty et al., 2017).*

Using Lemma 5.5 and Theorem 5.11, we are now ready to give a generalization bound on the target risk in terms of the Hierarchical Wasserstein distance we have constructed.

Theorem 5.13 *Under the assumptions of Lemma 5.5, let $\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ satisfying a $T_1(\zeta)$ inequality and let $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{X})$ such that $\mu_{\mathcal{S}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{S}}$ and $\mu_{\mathcal{T}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{T}}$. Let S and T be two sets of size n and m drawn i.i.d. from $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ respectively and let $\hat{\mu}_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\hat{\mu}_{\mathcal{T}} = \frac{1}{m}\sum_{j=1}^{m} \delta_{x_j}$ be their associated empirical measures. Assume further that samples in S and T are grouped respectively in k classes and k clusters, such that, the empirical measures of $\varphi_{\mathcal{S}}$ and $\varphi_{\mathcal{T}}$ can be expressed as $\hat{\varphi}_{\mathcal{S}} = \sum_{h=1}^{k} \frac{1}{k}\delta_{\rho_h}$ and $\hat{\varphi}_{\mathcal{T}} = \sum_{l=1}^{k} \frac{1}{k}\delta_{\varrho_l}$, where $\rho_h = \sum_{i=1/x_i \in C_h}^{n} \frac{1}{|C_h|}\delta_{x_i}$ and $\varrho_l = \sum_{j=1/x_j \in Cl_l}^{m} \frac{1}{|Cl_l|}\delta_{x_j}$ are the empirical measure of the h^e class C_h and l^e cluster Cl_l respectively. Then for any $\mathrm{d}' > dim(\mathcal{P}_p(\mathcal{X}))$ and $\zeta' < \zeta$, there exists some constant k_0 depending on d', such that for any $\delta > 0$ and $k \geq k_0 \max(\delta^{-(\mathrm{d}'+2)}, 1)$ with probability of at least $1 - \delta$ for all h, the following holds:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \mathcal{HW}_1(\hat{\varphi}_{\mathcal{S}}, \hat{\varphi}_{\mathcal{T}}) + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} + \lambda\,, \tag{5.26}$$

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

Proof

$$\begin{aligned}
\epsilon_{\mathcal{T}}(h) &\leq \epsilon_{\mathcal{T}}(h, h^*) + \epsilon_{\mathcal{T}}(h^*, f_{\mathcal{T}}) && (5.27)\\
&= \epsilon_{\mathcal{T}}(h, h^*) + \epsilon_{\mathcal{T}}(h^*, f_{\mathcal{T}}) + \epsilon_{\mathcal{S}}(h, h^*) - \epsilon_{\mathcal{S}}(h, h^*) && (5.28)\\
&\leq \epsilon_{\mathcal{T}}(h, h^*) + \epsilon_{\mathcal{T}}(h^*) + \epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{S}}(h^*) - \epsilon_{\mathcal{S}}(h, h^*) && (5.29)\\
&\leq \epsilon_{\mathcal{S}}(h) + \mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \epsilon_{\mathcal{S}}(h^*) + \epsilon_{\mathcal{T}}(h^*) && (5.30)\\
&\leq \epsilon_{\mathcal{S}}(h) + \mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda && (5.31)\\
&\leq \epsilon_{\mathcal{S}}(h) + \mathcal{HW}_1(\varphi_{\mathcal{S}}, \hat{\varphi}_{\mathcal{S}}) + \mathcal{HW}_1(\hat{\varphi}_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda && (5.32)\\
&\leq \epsilon_{\mathcal{S}}(h) + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} + \mathcal{HW}_1(\hat{\varphi}_{\mathcal{S}}, \hat{\varphi}_{\mathcal{T}}) + \mathcal{HW}_1(\hat{\varphi}_{\mathcal{T}}, \varphi_{\mathcal{T}}) + \lambda && (5.33)\\
&\leq \epsilon_{\mathcal{S}}(h) + \mathcal{HW}_1(\hat{\varphi}_{\mathcal{S}}, \hat{\varphi}_{\mathcal{T}}) + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} + \lambda && (5.34)
\end{aligned}$$

First and third lines are obtained using the triangular inequality applied to the error function. Fourth line is a consequence of Lemma 5.5. Fifth line follows from the definition of λ, sixth, seventh and eighth lines use the fact that Hierarchical Wasserstein metric is a proper distance and the Theorem 5.11 for $\mathcal{HW}_1$ applied to $\varphi_{\mathcal{S}}$ and $\varphi_{\mathcal{T}}$. ■

A straightforward implication of this theorem is that it justifies the application of hierarchical optimal transport in unsupervised domain adaptation. A similar result is Theorem 3.47, where authors in (Courty et al., 2017) use the Wasserstein distance to measure the similarity between the joint distribution of the source domain and an estimated joint distribution of the target one. Even if this bound does not have

the generic form in (3.29), it suggests the minimization of the Wasserstein distance between the joint distributions, which is very close to the minimization of the Hierarchical Wasserstein distance between classes and clusters in our bound.

Other similar results can be found in Theorem 3.43 of (Redko et al., 2017) and Theorem 3.44 of (Shen et al., 2018). The only distinction is the use of the Wasserstein distance in these bounds to measure the similarity between the marginal distributions of both domains rather than the Hierarchical Wasserstein distance in our case. Although one might think, due to the inequality in Lemma 5.4 that the proposed bound is less tight than the one in Theorem 3.43, but our bound has a major advantage, as shown below.

Indeed, the following Corollary gives a more explicit bound based on the development of the $\mathcal{HW}_1$ distance.

Corollary 5.14 *Under the assumptions of Theorem 5.13, let* $\Gamma^* = \underset{\Gamma \in U(\alpha,\beta)}{argmin} \langle \Gamma, \mathcal{W}_1 \rangle_F$ *be the optimal transport plan between* $\hat{\varphi}_\mathcal{S}$ *and* $\hat{\varphi}_\mathcal{T}$*, with probability of at least* $1-\delta$ *for all* h*, we have:*

$$\epsilon_\mathcal{T}(h) \leq \epsilon_\mathcal{S}(h) + \sum_{h=1}^{k} \mathcal{W}_1(\rho_h, \varrho_{\sigma(h)}) + k(k-1)\iota + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} + \lambda\,, \tag{5.35}$$

where $\sigma : \{1, ..., k\} \to \{1, ..., k\}$, $h \mapsto l^* = \underset{l}{argmax}\, \Gamma^*_{h,l}$ *and* $\iota = \max_{h, l \neq \sigma(h)} \mathcal{W}_1(\rho_h, \varrho_l)$.

Proof

$$\mathcal{HW}_1(\hat{\varphi}_\mathcal{S}, \hat{\varphi}_\mathcal{T}) = \sum_{h=1}^{k}\sum_{l=1}^{k} \mathcal{W}_1(\rho_h, \varrho_l)\Gamma^*_{h,l} \tag{5.36}$$

$$\leq \sum_{h=1}^{k}\sum_{l=1}^{k} \mathcal{W}_1(\rho_h, \varrho_l) \tag{5.37}$$

$$= \sum_{h=1}^{k} \mathcal{W}_1(\rho_h, \varrho_{\sigma(h)}) + \sum_{h=1}^{k}\sum_{l=1/l\neq\sigma(h)}^{k} \mathcal{W}_1(\rho_h, \varrho_l) \tag{5.38}$$

$$\leq \sum_{h=1}^{k} \mathcal{W}_1(\rho_h, \varrho_{\sigma(h)}) + k(k-1)\iota \tag{5.39}$$

First line follows from the definition of the Hierarchical Wasserstein distance. Second line uses the fact that $\Gamma^* \in U(\alpha, \beta)$, then we can bound each $\Gamma^*_{h,l}$ by 1 for simplicity [2]. Third and fourth lines are trivial. ■

The work of (El Hamri et al., 2022b) is based on the minimization of the Wasserstein distance between each class and its corresponding cluster, i.e. $\sum_{h=1}^{k} \mathcal{W}_1(\rho_h, \varrho_{\sigma(h)})$. The minimization of this amount leads eventually to the minimization of the Hierarchical Wasserstein distance in (5.26).

But also, when it is accompanied by a high-quality clustering in the target domain, it leads to the transportation of labeled source data of each class together without

[2] A tighter bound can be obtained by bounding each $\Gamma^*_{h,l}$ by $\frac{1}{k}$.

splitting to the region occupied by the target data having the same class label [3]. In this sense, the algorithmic solution suggested in (El Hamri et al., 2022b) in order to preserve compact classes during the transportation is explicitly reflected by the generalization bound (5.26), unlike the other bounds by (Redko et al., 2017; Shen et al., 2018).

Furthermore, this may suggest that one can independently minimize the other terms $\epsilon_{\mathcal{S}}(h)$ and λ since there is no longer the concern of transporting source data of different labels to the same target data.

5.3.2 A bound for semi-supervised domain adaptation

In semi-supervised domain adaptation, when we have access to an additional small set of labeled instances ϑn drawn independently from $\mu_{\mathcal{T}}$ in conjunction with $(1-\vartheta)n$ instances drawn independently from $\mu_{\mathcal{S}}$ and labeled by $f_{\mathcal{T}}$ and $f_{\mathcal{S}}$, respectively. The minimization of the target risk may not be the best choice, especially if ϑ is small, which is usually the case in semi-supervised domain adaptation. Instead, we can minimize a convex combination of the empirical source and target risk, defined as follows:

$$\hat{\epsilon}_{\theta}(h) = \theta\hat{\epsilon}_{\mathcal{T}}(h) + (1-\theta)\hat{\epsilon}_{\mathcal{S}}(h) \tag{5.40}$$

where $\theta \in [0,1]$.

In this section, we bound the target risk of a hypothesis that minimizes $\hat{\epsilon}_{\theta}(h)$. The proof of the bound has two main parts, which we state as Lemmas below.

Lemma 5.15 *Under the assumptions of Lemma 5.5, let $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{X})$ and let $\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ such that $\mu_{\mathcal{S}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{S}}$ and $\mu_{\mathcal{T}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{T}}$, let D be a labeled sample of size n with ϑn points drawn from $\mu_{\mathcal{T}}$ and $(1-\vartheta)n$ from $\mu_{\mathcal{S}}$ with $\vartheta \in (0,1)$, and labeled according to $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$. Then*

$$\mid \epsilon_{\theta}(h) - \epsilon_{\mathcal{T}}(h) \mid \leq (1-\theta)(\mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda) \tag{5.41}$$

Proof

$$\begin{aligned}
\mid \epsilon_{\theta}(h) - \epsilon_{\mathcal{T}}(h) \mid &= (1-\theta) \mid \epsilon_{\mathcal{S}}(h) - \epsilon_{\mathcal{T}}(h) \mid && (5.42)\\
&\leq (1-\theta)[\mid \epsilon_{\mathcal{S}}(h) - \epsilon_{\mathcal{S}}(h,h^*) \mid + \mid \epsilon_{\mathcal{S}}(h,h^*) - \epsilon_{\mathcal{T}}(h,h^*) \mid \\
&\quad + \mid \epsilon_{\mathcal{T}}(h,h^*) - \epsilon_{\mathcal{T}}(h) \mid] && (5.43)\\
&\leq (1-\theta)[\mid \epsilon_{\mathcal{S}}(h) - \epsilon_{\mathcal{S}}(h) - \epsilon_{\mathcal{S}}(h^*) \mid + \mid \epsilon_{\mathcal{S}}(h,h^*) - \epsilon_{\mathcal{T}}(h,h^*) \mid \\
&\quad + \mid \epsilon_{\mathcal{T}}(h) + \epsilon_{\mathcal{T}}(h^*) - \epsilon_{\mathcal{T}}(h) \mid] && (5.44)\\
&\leq (1-\theta)\,[\epsilon_{\mathcal{S}}(h^*) + \mid \epsilon_{\mathcal{S}}(h,h^*) - \epsilon_{\mathcal{T}}(h,h^*) \mid + \epsilon_{\mathcal{T}}(h^*)] && (5.45)\\
&\leq (1-\theta)(\mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda) && (5.46)
\end{aligned}$$

Second and third lines follow from the triangle inequality for classification error. The last line relies on Lemma 5.5. ■

In this Lemma where we bound the difference between the target risk $\epsilon_{\mathcal{T}}(h)$ and the weighted risk $\epsilon_{\theta}(h)$, we show that as θ approaches 1, we rely increasingly on the target data, and the distance between domains matters less and less.

[3] Evidently, the class labels are unknown in the target domain.

Lemma 5.16 *For a fixed hypothesis h, if a random labeled sample of size n is generated by drawing ϑm points from $\mu_{\mathcal{T}}$ and $(1-\vartheta)m$ from $\mu_{\mathcal{S}}$, and labeling them according to $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of the samples:*

$$\mathbb{P}\left[\mid \hat{\epsilon}_\theta(h) - \epsilon_\theta(h) \mid > 2\sqrt{K/n}\left(\frac{\theta}{n\vartheta\sqrt{\vartheta}} + \frac{(1-\theta)}{n(1-\vartheta)\sqrt{1-\vartheta}}\right) + \varepsilon\right] \leq \exp\left(\frac{-\varepsilon^2 n}{2K\left(\frac{\theta^2}{\vartheta} + \frac{(1-\theta)^2}{1-\vartheta}\right)}\right) \tag{5.47}$$

The Lemma above from (Redko et al., 2017) bound the difference between the true weighted risk $\epsilon_\theta(h)$ and its empirical counterpart $\hat{\epsilon}_\theta(h)$.

Theorem 5.17 *Under the assumptions of Theorem 5.13 and Lemma 5.5, let $\mu_{\mathcal{S}}, \mu_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{X})$ and let $\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ such that $\mu_{\mathcal{S}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{S}}$ and $\mu_{\mathcal{T}} = \int_{\mathcal{P}_p(\mathcal{X})} X d\varphi_{\mathcal{T}}$, let D be a labeled sample of size n with ϑn points drawn from $\mu_{\mathcal{T}}$ and $(1-\vartheta)n$ from $\mu_{\mathcal{S}}$ with $\vartheta \in (0,1)$, and labeled according to $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$. If $\hat{h}$ is the empirical minimizer of $\hat{\epsilon}_\theta(h)$ and $h^*_{\mathcal{T}} = \min_h \epsilon_{\mathcal{T}}(h)$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of the samples:*

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K\left(\frac{(1-\theta)^2}{1-\vartheta} + \frac{\theta^2}{\vartheta}\right)\log(2/\delta)}{n}} + 4\sqrt{K/n}\left(\frac{\theta}{n\vartheta\sqrt{\vartheta}} + \frac{(1-\theta)}{n(1-\vartheta)\sqrt{1-\vartheta}}\right) + 2(1-\theta)\left(\mathcal{HW}_1(\hat{\varphi}_{\mathcal{S}}, \hat{\varphi}_{\mathcal{T}}) + \lambda + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}}\right) \tag{5.48}$$

Proof

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_\theta(\hat{h}) + (1-\theta)(\mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda) \tag{5.49}$$

$$\leq \hat{\epsilon}_\theta(\hat{h}) + \sqrt{\frac{2K\left(\frac{(1-\theta)^2}{1-\vartheta} + \frac{\theta^2}{\vartheta}\right)\log(2/\delta)}{n}} + 2\sqrt{K/n}\left(\frac{\theta}{n\vartheta\sqrt{\vartheta}} + \frac{(1-\theta)}{n(1-\vartheta)\sqrt{1-\vartheta}}\right) + (1-\theta)(\mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda) \tag{5.50}$$

$$\leq \hat{\epsilon}_\theta(h^*_{\mathcal{T}}) + \sqrt{\frac{2K\left(\frac{(1-\theta)^2}{1-\vartheta} + \frac{\theta^2}{\vartheta}\right)\log(2/\delta)}{n}} + 2\sqrt{K/n}\left(\frac{\theta}{n\vartheta\sqrt{\vartheta}} + \frac{(1-\theta)}{n(1-\vartheta)\sqrt{1-\vartheta}}\right) + (1-\theta)(\mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda) + \lambda) \tag{5.51}$$

$$\leq \epsilon_\theta(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K\left(\frac{(1-\theta)^2}{1-\vartheta} + \frac{\theta^2}{\vartheta}\right)\log(2/\delta)}{n}} + 4\sqrt{K/n}\left(\frac{\theta}{n\vartheta\sqrt{\vartheta}} + \frac{(1-\theta)}{n(1-\vartheta)\sqrt{1-\vartheta}}\right) + (1-\theta)(\mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda) \tag{5.52}$$

$$\leq \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K\left(\frac{(1-\theta)^2}{1-\vartheta} + \frac{\theta^2}{\vartheta}\right)\log(2/\delta)}{n}} + 4\sqrt{K/n}\left(\frac{\theta}{n\vartheta\sqrt{\vartheta}} + \frac{(1-\theta)}{n(1-\vartheta)\sqrt{1-\vartheta}}\right) + 2(1-\theta)(\mathcal{HW}_1(\varphi_{\mathcal{S}}, \varphi_{\mathcal{T}}) + \lambda) \tag{5.53}$$

$$\leq \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K\left(\frac{(1-\theta)^2}{1-\vartheta} + \frac{\theta^2}{\vartheta}\right)\log(2/\delta)}{n}} + 4\sqrt{K/n}\left(\frac{\theta}{n\vartheta\sqrt{\vartheta}} + \frac{(1-\theta)}{n(1-\vartheta)\sqrt{1-\vartheta}}\right) + 2(1-\theta)\left(\mathcal{HW}_1(\hat{\varphi}_{\mathcal{S}}, \hat{\varphi}_{\mathcal{T}}) + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} + \lambda\right) \tag{5.54}$$

First and fifth lines follow from Lemma 5.15. Second and fourth lines are obtained using the concentration inequality of Lemma 5.16. Third line follows from the definition of $\hat{h}$ and $h^*_{\mathcal{T}}$. Sixth line follows from Theorem 5.5. ■

This theorem demonstrates that the best hypothesis $\hat{h}$ that takes into account both source and target labeled data (i.e., $0 \leq \theta \leq 1$) performs at least as good as the best hypothesis $h^*_{\mathcal{T}}$ learned on only target data ($\theta = 1$). This result is consistent with the insight that semi-supervised domain adaptation methods are expected to be as good as or better than unsupervised methods.

5.3.3 Bounds for multi-source domain adaptation

In this section, we consider the scenario of multi-source domain adaptation, where not one but many source domains are available. More formally, we have N different source domains. For each source j, we have a labeled sample S_j of size $n_j = \vartheta_j n$ drawn from the associated unknown distribution μ_{S_j} and labeled by f_{S_j}, such that $\sum_{j=1}^{N} \vartheta_j = 1$ and $\sum_{j=1}^{N} n_j = n$.

We define the empirical weighted multi-source risk of a hypothesis h for some vector $\theta = (\theta_1, ..., \theta_N)$ as follows:

$$\hat{\epsilon}_\theta(h) = \sum_{j=1}^{N} \theta_j \hat{\epsilon}_{S_j}(h) \tag{5.55}$$

where $\sum_{j=1}^{N} \theta_j = 1$ and each θ_j represents the weight of the source domain S_j.

We present in turn two generalization bounds for the setting of multi-source domain adaptation. The first bound uses the pairwise Hierarchical Wasserstein distance between each source and the target domain, while the second bound uses the combined Hierarchical Wasserstein distance.

The proof of these bounds has a main common component, which we state as Lemma below.

Lemma 5.18 *For a fixed hypothesis h, if a random labeled sample of size n is generated by drawing $\vartheta_j n$ points from μ_{S_j} and labeled according to f_{S_j} for each $j \in \{1, ..., N\}$ and for any fixed weight vector θ. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\mathbb{P}\left[\mid \hat{\epsilon}_\theta(h) - \epsilon_\theta(h) \mid > 2\sqrt{K/n} \sum_{j=1}^{N} \frac{\theta_j}{\vartheta_j n \sqrt{\vartheta_j}} + \varepsilon \right] \leq \exp\left(\frac{-\varepsilon^2 n}{2K \sum_{j=1}^{N} \frac{\theta_j^2}{\vartheta_j}} \right). \tag{5.56}$$

This Lemma from (Redko et al., 2017) provides a uniform convergence bound for the empirical weighted risk.

5.3.3.1 A bound using pairwise Hierarchical Wasserstein distance

The first bound we present considers the pairwise Hierarchical Wasserstein distance between each source and the target domain. The term $\sum_{j=1}^{N} \theta_j \lambda_j$ that appears in this bound plays a role corresponding to λ in the previous sections.

Before presenting the bound in question, we must prove the Lemma below that bounds the difference between the target risk $\epsilon_{\mathcal{T}}(h)$ and the weighted risk $\epsilon_{\theta}(h)$.

Lemma 5.19 *Under the assumptions of Theorem 5.13 and Lemma 5.5, let D be a sample of size n, where for each $j \in \{1, ..., N\}$, $\vartheta_j n$ points are drawn from μ_{S_j} and labeled according to f_{S_j}. Then:*

$$| \epsilon_{\theta}(h) - \epsilon_{\mathcal{T}}(h) | \leq \sum_{j=1}^{N} \theta_j (\mathcal{HW}_1(\varphi_{S_j}, \varphi_{\mathcal{T}}) + \lambda_j). \tag{5.57}$$

Proof

$$| \epsilon_{\theta}(h) - \epsilon_{\mathcal{T}}(h) | = | \sum_{j=1}^{N} \theta_j \epsilon_{S_j}(h) - \epsilon_{\mathcal{T}}(h) | \tag{5.58}$$

$$\leq \sum_{j=1}^{N} \theta_j \mid \epsilon_{S_j}(h) - \epsilon_{\mathcal{T}}(h) \mid \tag{5.59}$$

$$\leq \sum_{j=1}^{N} \theta_j [\mid \epsilon_{S_j}(h) - \epsilon_{S_j}(h, h_j^*) \mid + \mid \epsilon_{S_j}(h, h_j^*) - \epsilon_{\mathcal{T}}(h, h_j^*) \mid + \mid \epsilon_{\mathcal{T}}(h, h_j^*) - \epsilon_{\mathcal{T}}(h) \mid] \tag{5.60}$$

$$\leq \sum_{j=1}^{N} \theta_j [\mid \epsilon_{S_j}(h) - \epsilon_{S_j}(h) - \epsilon_{S_j}(h_j^*) \mid + \mid \epsilon_{S_j}(h, h_j^*) - \epsilon_{\mathcal{T}}(h, h_j^*) \mid + \mid \epsilon_{\mathcal{T}}(h) + \epsilon_{\mathcal{T}}(h_j^*) - \epsilon_{\mathcal{T}}(h) \mid] \tag{5.61}$$

$$\leq \sum_{j=1}^{N} \theta_j \left[\epsilon_{S_j}(h_j^*) + \mid \epsilon_{S_j}(h, h_j^*) - \epsilon_{\mathcal{T}}(h, h_j^*) \mid + \epsilon_{\mathcal{T}}(h_j^*)\right] \tag{5.62}$$

$$\leq \sum_{j=1}^{N} \theta_j (\mathcal{HW}_1(\varphi_{S_j}, \varphi_{\mathcal{T}}) + \lambda_j) \tag{5.63}$$

Third and fourth lines follow from the triangle inequality for classification error. The last line relies on Lemma 5.5. ■

We now prove the bound that considers the data available from each source individually, ignoring the relationships between sources, using pairwise Hierarchical Wasserstein distance.

Theorem 5.20 *Under the assumptions of Theorem 5.13 and Lemma 5.5, let D be a sample of size n, where for each $j \in \{1, ..., N\}$, $\vartheta_j n$ points are drawn from μ_{S_j} and labeled according to f_{S_j}. If $\hat{h}$ is the empirical minimizer of $\hat{\epsilon}_{\theta}(h)$ and $h_{\mathcal{T}}^* = \min_h \epsilon_{\mathcal{T}}(h)$ then for any fixed θ and $\delta \in (0, 1)$ with probability at least $1 - \delta$ (over the choice of samples),*

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\mathcal{T}}(h_{\mathcal{T}}^*) + 2\sqrt{\frac{2K \sum_{j=1}^{N} \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^{N} \frac{K\theta_j}{\vartheta_j n}} + 2\sum_{j=1}^{N} \theta_j \left(\mathcal{HW}_1(\hat{\varphi}_{S_j}, \hat{\varphi}_{\mathcal{T}}) + \lambda_j + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} \right), \tag{5.64}$$

where $\lambda_j = \min_h(\epsilon_{S_j}(h) + \epsilon_{\mathcal{T}}(h))$ represents the joint error for each source domain S_j.

Proof

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_\theta(\hat{h}) + \sum_{j=1}^{N} \theta_j (\mathcal{HW}_1(\varphi_{S_j}, \varphi_{\mathcal{T}}) + \lambda_j) \tag{5.65}$$

$$\leq \hat{\epsilon}_\theta(\hat{h}) + \sqrt{\frac{2K \sum_{j=1}^{N} \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + \sqrt{\sum_{j=1}^{N} \frac{K\theta_j}{\vartheta_j n}}$$

$$+ \sum_{j=1}^{N} \theta_j (\mathcal{HW}_1(\varphi_{S_j}, \varphi_{\mathcal{T}}) + \lambda_j) \tag{5.66}$$

$$\leq \hat{\epsilon}_\theta(h^*_{\mathcal{T}}) + \sqrt{\frac{2K \sum_{j=1}^{N} \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + \sqrt{\sum_{j=1}^{N} \frac{K\theta_j}{\vartheta_j n}}$$

$$+ \sum_{j=1}^{N} \theta_j (\mathcal{HW}_1(\varphi_{S_j}, \varphi_{\mathcal{T}}) + \lambda_j) \tag{5.67}$$

$$\leq \epsilon_\theta(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K \sum_{j=1}^{N} \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^{N} \frac{K\theta_j}{\vartheta_j n}}$$

$$+ \sum_{j=1}^{N} \theta_j (\mathcal{HW}_1(\varphi_{S_j}, \varphi_{\mathcal{T}}) + \lambda_j) \tag{5.68}$$

$$\leq \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K \sum_{j=1}^{N} \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^{N} \frac{K\theta_j}{\vartheta_j n}}$$

$$+ 2\sum_{j=1}^{N} \theta_j (\mathcal{HW}_1(\varphi_{S_j}, \varphi_{\mathcal{T}}) + \lambda_j) \tag{5.69}$$

$$\leq \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K \sum_{j=1}^{N} \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^{N} \frac{K\theta_j}{\vartheta_j n}}$$

$$+ 2\sum_{j=1}^{N} \theta_j \left(\mathcal{HW}_1(\hat{\varphi}_{\mathcal{S}}, \hat{\varphi}_{\mathcal{T}}) + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} + \lambda_j \right) \tag{5.70}$$

First and fifth lines follow from Lemma 5.19. Second and fourth lines are obtained using the concentration inequality of Lemma 5.18. Fourth line is a consequence of Lemma 5.5. Third line follows from the definition of $\hat{h}$ and $h^*_{\mathcal{T}}$. Sixth line follows from Theorem 5.5. ■

5.3.3.2 A bound using combined Hierarchical Wasserstein distance

In the former bound, the Hierarchical Wasserstein distance between domains is only measured on pair, so it is not required to have a hypothesis that is valid for each source domain. The alternate bound shown in the next theorem enables us to alter the source distribution by changing θ. This has two implications. First of all, we now

need to insist that there is a hypothesis h^* which has low risk on both the θ-weighted convex combination of sources and the target domain. Secondly, we measure the Hierarchical Wasserstein distance between the target and a mixture of sources, instead of between the target and every single source.

Lemma 5.21 *Under the assumptions of Theorem 5.13 and Lemma 5.5, let D be a sample of size n, where for each $j \in \{1, ..., N\}$, $\vartheta_j n$ points are drawn from μ_{S_j} and labeled according to f_{S_j}. Then*

$$| \epsilon_\theta(h) - \epsilon_\mathcal{T}(h) | \leq \mathcal{HW}_1(\varphi_{S_\theta}, \varphi_\mathcal{T}) + \lambda_\theta. \tag{5.71}$$

Proof

$$\begin{aligned}
| \epsilon_\theta(h) - \epsilon_\mathcal{T}(h) | \leq & | \epsilon_\theta(h) - \epsilon_\theta(h, h^*) | + | \epsilon_\theta(h, h^*) - \epsilon_\mathcal{T}(h, h^*) | \\
& + | \epsilon_\mathcal{T}(h, h^*) - \epsilon_\mathcal{T}(h) | && (5.72)\\
\leq & | \epsilon_\theta(h) - \epsilon_\theta(h) + -\epsilon_\theta(h^*) | + | \epsilon_\theta(h, h^*) - \epsilon_\mathcal{T}(h, h^*) | \\
& + | \epsilon_\mathcal{T}(h) + \epsilon_\mathcal{T}(h^*) - \epsilon_\mathcal{T}(h) | && (5.73)\\
\leq & \epsilon_\theta(h^*) + | \epsilon_\theta(h, h^*) - \epsilon_\mathcal{T}(h, h^*) | + \epsilon_\mathcal{T}(h^*) && (5.74)\\
\leq & \mathcal{HW}_1(\varphi_{S_\theta}, \varphi_\mathcal{T}) + \lambda_\theta && (5.75)
\end{aligned}$$

First and second lines follow from the triangle inequality for classification error. The last line relies on Lemma 5.5. ■

We now prove the bound using combined Hierarchical Wasserstein distance.

Theorem 5.22 *Under the assumptions of Theorem 5.13 and Lemma 5.5, let D be a sample of size n, where for each $j \in \{1, ..., N\}$, $\vartheta_j n$ points are drawn from μ_{S_j} and labeled according to f_{S_j}. If $\hat{h}$ is the empirical minimizer of $\hat{\epsilon}_\theta(h)$ and $h^*_\mathcal{T} = \min_h \epsilon_\mathcal{T}(h)$ then for any fixed θ and $\delta \in (0, 1)$ with probability at least $1 - \delta$ (over the choice of samples)*

$$\begin{aligned}
\epsilon_\mathcal{T}(\hat{h}) \leq \epsilon_\mathcal{T}(h^*_\mathcal{T}) + 2\sqrt{\frac{2K \sum_{j=1}^N \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^N \frac{K\theta_j}{\vartheta_j n}} \\
+ 2\left(\mathcal{HW}_1(\varphi_{S_\theta}, \varphi_\mathcal{T}) + \lambda_\theta + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} \right).
\end{aligned} \tag{5.76}$$

where $\lambda_\theta = \min_h(\epsilon_{S_\theta}(h) + \epsilon_\mathcal{T}(h))$ represents the joint error of the target and the combination of the source domains.

Proof

$$\begin{aligned}
\epsilon_\mathcal{T}(\hat{h}) &\leq \epsilon_\theta(\hat{h}) + \mathcal{HW}_1(\varphi_{S_\theta}, \varphi_\mathcal{T}) + \lambda_\theta && (5.77)\\
&\leq \hat{\epsilon}_\theta(\hat{h}) + \sqrt{\frac{2K \sum_{j=1}^N \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + \sqrt{\sum_{j=1}^N \frac{K\theta_j}{\vartheta_j n}} \\
&\quad + \mathcal{HW}_1(\varphi_{S_\theta}, \varphi_\mathcal{T}) + \lambda_\theta && (5.78)\\
&\leq \hat{\epsilon}_\theta(h^*_\mathcal{T}) + \sqrt{\frac{2K \sum_{j=1}^N \frac{\theta_j^2}{\vartheta_j} \log(2/\delta)}{n}} + \sqrt{\sum_{j=1}^N \frac{K\theta_j}{\vartheta_j n}}
\end{aligned}$$

$$+ \mathcal{HW}_1(\varphi_{S_\theta}, \varphi_{\mathcal{T}}) + \lambda_\theta \tag{5.79}$$

$$\leq \epsilon_\theta(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K\sum_{j=1}^N \frac{\theta_j^2}{\vartheta_j}\log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^N \frac{K\theta_j}{\vartheta_j n}}$$

$$+ \mathcal{HW}_1(\varphi_{S_\theta}, \varphi_{\mathcal{T}}) + \lambda_\theta \tag{5.80}$$

$$\leq \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K\sum_{j=1}^N \frac{\theta_j^2}{\vartheta_j}\log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^N \frac{K\theta_j}{\vartheta_j n}}$$

$$+ 2\left(\mathcal{HW}_1(\varphi_{S_\theta}, \varphi_{\mathcal{T}}) + \lambda_\theta\right) \tag{5.81}$$

$$\leq \epsilon_{\mathcal{T}}(h^*_{\mathcal{T}}) + 2\sqrt{\frac{2K\sum_{j=1}^N \frac{\theta_j^2}{\vartheta_j}\log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^N \frac{K\theta_j}{\vartheta_j n}}$$

$$+ 2\left(\mathcal{HW}_1(\varphi_{S_\theta}, \varphi_{\mathcal{T}}) + 2\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\zeta' k}} + \lambda_\theta\right) \tag{5.82}$$

First and fifth lines follow from Lemma 5.21. Second and fourth lines are obtained using the concentration inequality of Theorem 5.18. Third line follows from the definition of $\hat{h}$ and $h^*_{\mathcal{T}}$. Sixth line follows from Theorem 5.5. ■

5.4 Conclusion and future perspectives

In this chapter, using hierarchical optimal transport we presented a theoretical study of domain adaptation, a problem in which we have an abundant amount of labeled samples from a source domain, but we aim to deploy a model in another target domain with a much smaller amount of labeled samples or even no labeled samples. Our main results are generalization bounds for both single and multi-source domain adaptation scenarios, where the divergence between the source and target domains is measured by the Hierarchical Wasserstein distance. Our generalization bounds justify the application of hierarchical optimal transport in the context of domain adaptation and may suggest under the assumption of successful clustering in the target domain that one can minimize the other terms $\epsilon_{\mathcal{S}}(h)$ and λ without difficulty independently from the minimization of the Hierarchical Wasserstein distance.

Future perspectives of this work are numerous and concern both the derivation of new domain adaptation algorithms and the demonstration of new generalization bounds. Indeed, the work of this chapter can be extended in different directions:

- First of all, we would like to derive a new domain adaptation algorithm based on the insights provided by the bounds in the multi-source settings.
- Secondly, we aim to produce new generalization bounds that take into account the quality of clustering in the target domain, by reflecting explicitly the excess clustering risk.
- Finally, the ability term λ is surprisingly understudied, and we would like to provide a theoretical analysis of this term using hierarchical optimal transport and investigate the possibility to estimate it from finite samples.

CHAPTER 6

OPTIMAL TRANSPORT FOR SEMI-SUPERVISED LEARNING

Contents

Marked by the necessity for an efficient way to learn hidden structures in the target domain other than clustering, and by the many drawbacks of traditional semi-supervised approaches, we tackle in this chapter based on (El Hamri et al., 2021a,b,c), the problem of transductive semi-supervised learning, that aims to obtain label predictions for the given unlabeled data according to Vapnik's principle. The proposed approach, Optimal Transport Propagation (OTP), performs in an incremental process, label propagation through the edges of a complete bipartite edge-weighted graph, whose affinity matrix is constructed from the optimal transport plan between empirical measures defined on labeled and unlabeled data. OTP ensures a high degree of prediction certitude by controlling the propagation process using a certainty score based on Shannon's entropy. We also provide a convergence analysis of our algorithm and an extension to out-of-sample data (OTI). Experiments show the superiority of the proposed approach over the state-of-the-art.

> Silence is the language of God, all else is poor translation.
>
> Jalal al-Din Rumi

6.1 Introduction

Semi-supervised learning has recently emerged as one of the most promising paradigms to alleviate the lack of massive labeled datasets, especially in learning tasks where it is prohibitively expensive to collect a large amount of high-quality labeled data due to lack of time, resources, or other factors, while unlabeled data is cheap and abundant. This is best illustrated in medicine, where measurements require expensive machinery and labels are the results of a labor and expensive expert-assisted time-consuming analysis.

Among many semi-supervised learning approaches, graph-based techniques are increasingly being studied due to their performance and to more and more real graph datasets. The problem is to predict all the unlabeled vertices in the graph based on only a small subset of vertices being observed. To date, a number of graph-based algorithms, in particular label propagation methods have been successfully applied to different fields, such as social network analysis (Boldi et al., 2011; Xie and Szymanski, 2013; Zhang et al., 2017; Jokar and Mosleh, 2019), natural language processing (Alexandrescu and Kirchhoff, 2007; Tamura et al., 2012; Barba et al., 2020), and image segmentation (Wang et al., 2007; Breve, 2019).

The performance of label propagation algorithms is often affected by the graph-construction method and the technique of inferring pseudo-labels. For graph-construction, traditional label propagation approaches are incapable of exploiting the underlying geometry of the whole input space, and the relations between labeled and unlabeled data in a global vision. Indeed, authors in (Zhu and Ghahramani, 2002; Zhou et al., 2003) have adopted pairwise relationships between instances by relying on a Gaussian function with a free parameter σ, whose optimal value can be hard to determine if only very few labeled data are available, and even a small perturbation in its value can affect significantly the classification results. Authors in (Wang and Zhang, 2007) have suggested deriving another way to avoid the use of σ, by relying on the local concept of linear neighborhood, although, the linearity assumption is intended just for computational convenience, and the variance in the neighborhood size can also drastically change the classification results. Moreover, these algorithms have the inconvenience of inferring pseudo-labels by hard assignment, ignoring the different degrees of certainty associated with each prediction. Another drawback of these algorithms is their inability to generalize for out-of-sample data.

Contributions: In this chapter, we address the existing limitations above by proposing a principally new approach based on optimal transport. Optimal transport is an efficient paradigm to capture the geometry of the data in the input space and it has found a renewed interest in semi-supervised learning community (Solomon et al., 2014; Taherkhani et al., 2020). The proposed algorithm, called Optimal Transport Propagation (OTP), has several points of differentiation from state-of-the-art approaches. Its main contributions can be summarized as follows:

1. OTP constructs a complete bipartite edge-weighted graph, to avoid adding a regularization term in the corresponding objective function for penalizing predicted labels that do not match the correct ones.

2. OTP infers an enhanced affinity matrix from the optimal transport plan between empirical measures defined on labeled and unlabeled samples, to benefit from all the geometrical information in the input space.

3. OTP performs label propagation in an incremental process to take advantage of the dependency of semi-supervised algorithms on the amount of prior information [1].

4. OTP incorporates a certainty score based on Shannon's entropy to control the certitude of the predictions during the incremental propagation process.

5. OTP can efficiently be extended to out-of-sample data, which gives rise to Optimal Transport Induction (OTI).

Outline: The rest of this chapter is organized as follows: in the 2^{nd} section, we provide a brief overview of semi-supervised learning. In the 3^{rd} section, we elaborate the proposed approach OTP. In the 4^{th} section, we extend OTP to out-of-sample data. In the 5^{th} section, we evaluate our algorithm on several real-world datasets. Finally, we conclude in section 6.

6.2 Semi-supervised learning

Semi-supervised learning (Zhu, 2005) is conceptually situated between supervised and unsupervised learning. The goal of semi-supervised learning is to use a large amount of unlabeled instances as well as a typically smaller set of labeled samples, usually assumed to be sampled from the same distribution, in order to improve the performance that can be obtained either by discarding the unlabeled data and doing classification (supervised learning) or by discarding the available labels and doing clustering (unsupervised learning).

In semi-supervised learning settings, we have access to a finite ordered set $X = \{x_1, ..., x_{l+u}\}$ of $l + u$ samples in $\mathcal{X} = \mathbb{R}^{\text{d}}$, and a discrete label set $\mathcal{Y} = \{C_1, ..., C_k\}$ of k classes. The first l points denoted by $X_L = \{x_1, ..., x_l\} \subset \mathcal{X}$ are labeled according to $Y_L = \{y_1, ..., y_l\}$, where $y_i \in \mathcal{Y}$ for every $i \in \{1, ..., l\}$, and the remaining data denoted by $X_U = \{x_{l+1}, ..., x_{l+u}\} \subset \mathcal{X}$ are unlabeled, usually $l \ll u$.

Semi-supervised learning makes use of at least one of the following assumptions (Van Engelen and Hoos, 2020):

- **Smoothness assumption:** For two samples $x, x^{'}$ that are close in the input space $\mathcal{X}$, the corresponding labels $y, y^{'}$ should be the same.

- **Low-density assumption:** The decision boundary should preferably pass through low-density regions in the input space $\mathcal{X}$.

- **Manifold assumption:** The high-dimensional input space $\mathcal{X}$ is constituted of multiple lower-dimensional substructures known as manifolds and samples lying on the same manifold should have the same label.

[1] We mean by the amount of prior information, the amount of labeled samples.

- **Cluster assumption:** Data belonging to the same cluster are likely to have the same label.

We can differentiate between two categories of semi-supervised learning: transductive and inductive semi-supervised learning (Van Engelen and Hoos, 2020). The former is solely concerned with obtaining label predictions for the given unlabeled samples, whereas the latter attempt to infer a good classifier that can estimate the label for any instance in the input space, even for previously unseen data.

The nature of transductive semi-supervised methods, make them inherently a perfect illustration of Vapnik's principle: when trying to solve some problem, one should not solve a more difficult problem as an intermediate step (Chapelle et al., 2009). This principle naturally suggests finding a way to propagate information via direct connections between samples using a graph-based approach. In fact, if we can define a graph in which similar samples are connected, information can then be propagated along its edges, which relieves us from learning a classifier.

Graph-based semi-supervised learning approaches generally involve two separate phases: graph construction and label propagation. In the first phase, vertices are connected, based on some similarity measure, and the resulting edges are weighted, such that, the stronger the similarity the higher the weight. Once the graph is constructed, it will be used in the second phase of label propagation to obtain predictions for the unlabeled data (Subramanya and Talukdar, 2014).

The common major inconvenience of transductive methods is their inability to predict labels for out-of-sample data, so when some previously unseen test data arrive, transductive learning methods need to fusion these new samples into the previous data at our disposal to reconstruct a new augmented graph based on merged data, and then perform label propagation from scratch. This process is too costly, since the presentation of even, a single new point requires rerunning all the processes already done in their entirety, which is distasteful in many real-world applications, where on-the-fly prediction for previously unseen instances is highly requested. Hence the importance of inductive semi-supervised learning.

6.3 Optimal Transport Propagation

In this section, we introduce the proposed OTP approach, which consists of two phases, the first one aims to construct a complete bipartite edge-weighted graph with an enhanced affinity matrix using optimal transport, and the second phase focuses on using an incremental process to propagate the labels from labeled to unlabeled samples.

Let $X = \{x_1, ..., x_{l+u}\}$ be a set of $l + u$ samples in the input space $\mathcal{X} = \mathbb{R}^{\mathrm{d}}$ and $\mathcal{Y} = \{C_1, ..., C_k\}$ a discrete label set consisting of k classes. The first l samples denoted by $X_L = \{x_1, ..., x_l\}$ are labeled according to $Y_L = \{y_1, ..., y_l\}$, where $y_i \in \mathcal{Y}$ for every $i \in \{1, ..., l\}$, and the remaining u samples denoted by $X_U = \{x_{l+1}, ..., x_{l+u}\}$ are unlabeled. Usually $l \ll u$. OTP aims to infer the unknown labels Y_U using all the samples in $X = X_L \cup X_U$ and labels Y_L.

To use an appropriate formulation to the paradigm of optimal transport, the empirical distribution of X_L and X_U must be expressed respectively using discrete measures as:

$$\mu = \sum_{i=1}^{l} a_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{j=l+1}^{l+u} b_j \delta_{x_j}. \tag{6.1}$$

Under the assumption that X_L and X_U are collections of i.i.d. samples, the weights of all instances in each set are naturally set to be equal:

$$a_i = \frac{1}{l}, \forall i \in \{1, ..., l\} \quad \text{and} \quad b_j = \frac{1}{u}, \forall j \in \{l+1, ..., l+u\}. \tag{6.2}$$

6.3.1 Graph construction

The objective of this phase is to construct a complete bipartite edge-weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V} = X$ is the vertex set, that can be divided into two disjoint and independent parts $\mathcal{L} = X_L$ and $\mathcal{U} = X_U$, $\mathcal{E} \subset \{\mathcal{L} \times \mathcal{U}\}$ is the edge set, and $\mathcal{W} \in \mathcal{M}_{l,u}(\mathbb{R}^+)$ is the affinity matrix that denotes the edges weights, the weight $w_{i,j}$ on the edge $e_{i,j} \in \mathcal{E}$ reflects the degree of similarity between $x_i \in X_L$ and $x_j \in X_U$.

One must take into consideration that, intuitively, we want samples that are close in the input space $\mathcal{X}$ to have similar labels (smoothness assumption). Thus, to measure quantitatively the closeness between samples, we need to use some distance over the input space. For this purpose, let's consider the matrix of pairwise squared Euclidean distances $C \in \mathcal{M}_{l \times u}(\mathbb{R}^+)$ between samples of X_L and X_U, defined by:

$$c_{i,j} = \|x_i - x_j\|^2, \ \ \forall i, j \in \{1, ..., l\} \times \{l+1, ..., l+u\}. \tag{6.3}$$

In order to construct an affinity matrix $\mathcal{W}$ that captures the underlying geometry of the whole samples in the input space and all the relations between labeled and unlabeled samples in a global vision, instead of the pairwise relationships or the local neighborhood information, a natural choice is to rely on optimal transport, which is a powerful tool for capturing the underlying geometry of the data. Since optimal transport suffers from a computational burden, we can overcome this issue by using its entropic regularized version between μ and ν, in the following way:

$$\gamma_\varepsilon^* = \underset{\gamma \in U(a,b)}{\operatorname{argmin}} \ \langle \gamma, C \rangle_F - \varepsilon H(\gamma). \tag{6.4}$$

The optimal transport plan γ_ε^* provides us the weights of associations between vertices in $\mathcal{L}$ and $\mathcal{U}$, thus, γ_ε^* can be interpreted in our context as a similarity matrix between the two parts $\mathcal{L}$ and $\mathcal{U}$ of the graph $\mathcal{G}$: similar labeled and unlabeled vertices correspond to a higher value in γ_ε^*.

To have a class probability interpretation afterwards, we column-normalize γ_ε^* to get a non-square left-stochastic affinity matrix $\mathcal{W} \in \mathcal{M}_{l,u}(\mathbb{R}^+)$, defined as follows:

$$w_{i,j} = \frac{\gamma^*_{\varepsilon_{i,j}}}{\sum_i \gamma^*_{\varepsilon_{i,j}}}, \ \ \forall i, j \in \{1, ..., l\} \times \{l+1, ..., l+u\}, \tag{6.5}$$

where $w_{i,j}$ is then the probability of jumping from the vertex $x_i \in \mathcal{L}$ to the vertex $x_j \in \mathcal{U}$.

6.3.2 Label propagation

The intuition behind this phase is to use the affinity matrix $\mathcal{W}$ to identify labeled samples that should spread their labels to similar unlabeled instances. We suggest using an incremental process to infer labels of the samples in $\mathcal{U}$. We suggest also to provide with each pseudo-label a certainty score that measures the certitude of the prediction and uses it to control the incremental label propagation process.

First, we need to construct a label matrix $U \in \mathcal{M}_{u,k}(\mathbb{R}^+)$ that indicates the probability of each unlabeled sample x_j, $j \in \{l+1, ..., l+u\}$ to belong to the class c_h, $h \in \{1, ..., k\}$. For a harmonious construction of the label matrix U with the information coming from the optimal transport plan γ_ε^*, we propose to define this probability as the sum of the similarities of x_j with the representatives of the class c_h:

$$u_{j,h} = \sum_{i/x_i \in c_h} w_{i,j}, \quad \forall j, h \in \{l+1, ..., l+u\} \times \{1, ..., k\}. \tag{6.6}$$

The matrix U is a non-square right-stochastic matrix, and can be interpreted as a vector-valued function $U : X_U \to \sum_k$, which assigns a stochastic vector $U_j \in \sum_k$ to each unlabeled sample x_j, $j \in \{l+1, ..., l+u\}$.

Traditional label propagation approaches infer simultaneously all the pseudo-labels by hard assignment, without worrying about the fact that these label predictions do not have the same degree of certainty. This issue, as mentioned by (Iscen et al., 2019), can degrade significantly the performance of the label propagation approaches. To prevent this, we suggest to associate a certainty score s_j with the label prediction of each $x_j, j \in \{l+1, ..., l+u\}$. The proposed certainty score s_j is defined in the following way:

$$s_j = 1 - \frac{\mathrm{H}(Z_j)}{log_2(k)}, \quad \forall j \in \{l+1, ..., l+u\}, \tag{6.7}$$

where $Z_j : \mathcal{Y} \to \mathbb{R}$ is a real-valued random variable, that assigns to the sample x_j the probability of belonging to a class c_h. The probability distribution of the random variable Z_j is encoded in the stochastic vector U_j:

$$\mathbb{P}(Z_j = c_h) = u_{j,h}, \quad \forall j, h \in \{l+1, ..., l+u\} \times \{1, ..., k\}, \tag{6.8}$$

and H is Shannon's entropy (Shannon, 2001), defined by:

$$\mathrm{H}(Z_j) = -\sum_{h=1}^{k} \mathbb{P}(Z_j = c_h) \log_2(\mathbb{P}(Z_j = c_h)) = -\sum_{h=1}^{k} u_{j,h} \log_2(u_{j,h}). \tag{6.9}$$

Since Shannon's entropy H is an uncertainty measure that reach its maximal value $\log_2(k)$ when all the possible events are equiprobable, i.e. $\forall h \leq k, \quad \mathbb{P}(Z_j = c_h) = \frac{1}{k}$:

$$\mathrm{H}(Z_j) = -\sum_{h=1}^{k} \frac{1}{k} \log_2(\frac{1}{k}) = \log_2(k). \tag{6.10}$$

Then, by dividing H by $\log_2(k)$ the certainty score s_j is naturally normalized between 0 and 1.

To control the certainty of the prediction resulting from the propagation process,

we define a confidence threshold $\alpha \in [0, 1]$, and for each unlabeled sample x_j, we make a comparison between α and its associated certainty score s_j. If the score s_j is greater than α, we assign to x_j a pseudo-label $\hat{y}_j$, in the following way:

$$\hat{y}_j = \underset{c_h \in \mathcal{Y}}{\operatorname{argmax}}\, u_{j,h}, \ \forall j \in \{l+1, ..., l+u\}. \tag{6.11}$$

Thus, the unlabeled instance x_j will belong to the class c_h with the highest class-probability $u_{j,h}$, in other words, to the class whose representatives possess the highest similarity with x_j. Otherwise, we do not give any label to the point x_j.

The process above corresponds to one iteration of the proposed incremental approach. At each iteration, X_L is enriched with new instances, and the number of samples in X_U is reduced:

$$X_L = X_L \cup \{x_j \in X_U \mid s_j > \alpha\} \quad \text{and} \quad Y_L = Y_L \cup \{\hat{y}_j \mid s_j > \alpha\}, \tag{6.12}$$

$$X_U = X_U \setminus \{x_j \in X_U \mid s_j > \alpha\}. \tag{6.13}$$

This modification of X_L, Y_L, and X_U at each iteration of the incremental process is of major importance in the context of label propagation, since, the effectiveness of a label propagation algorithm depends on the amount of prior information, thus, increasing the size of X_L at each iteration, will increase the performance of the proposed approach, and will make it possible to infer the label of the samples remaining in X_U with a high degree of certainty at the next iterations.

We repeat the same whole procedure at each iteration until convergence, here convergence means that all the data initially in X_U are labeled during this incremental process. The proposed OTP approach is formally summarized in Algorithm 6.1:

Algorithm 6.1 OTP

Parameters: ε, α
Input : X_L, X_U, Y_L
while *not converged* **do**
  Compute the cost matrix C (6.3)
  Solve the optimal transport problem (6.4)
  Compute the affinity matrix $\mathcal{W}$ (6.5)
  Get the label matrix U (6.6)
  for $x_j \in X_U$ **do**
    Compute the certainty score s_j (6.7)
    if $s_j > \alpha$ **then**
      Get the pseudo label $\hat{y}_j$ by (6.11)
      Inject x_j in X_L and $\hat{y}_j$ in Y_L (6.12)
      Remove x_j from X_U (6.13)
    else
      Maintain x_j in X_U
    end
  end
end
return Y_U

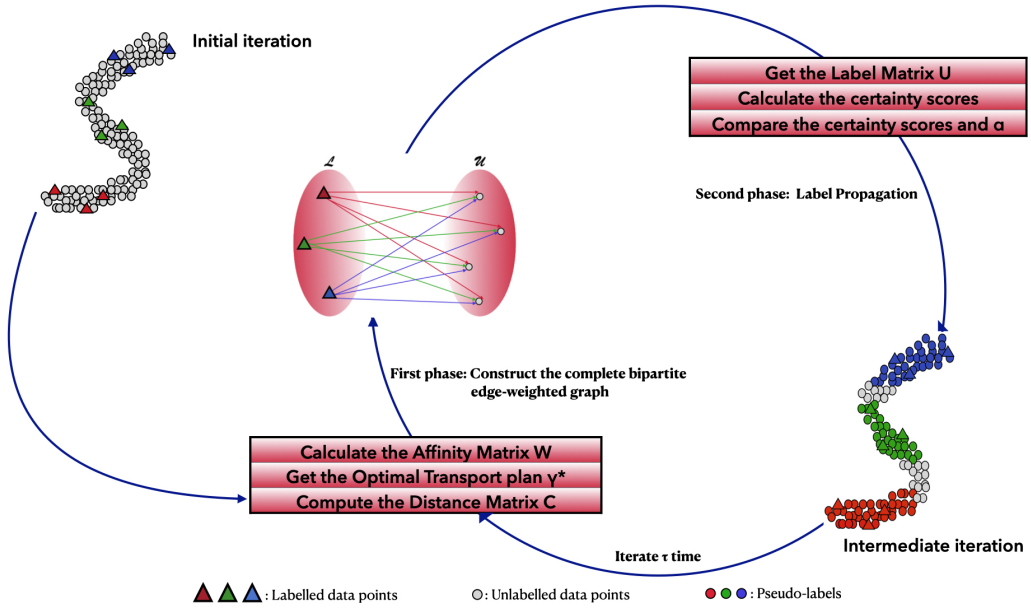

FIGURE 6.1: Overview of OTP. We initiate an incremental approach where at each iteration, we construct a complete bipartite edge-weighted graph based on the optimal transport plan between the distribution of labeled instances and unlabeled ones. Then, we propagate labels through the edges of the graph. Triangles markers correspond to the labeled instances and circles correspond to the unlabeled data which are gradually pseudo-labeled by OTP. The class is color-coded.

6.3.3 Convergence analysis

As mentioned earlier, the convergence of OTP means that all the data initially in X_U are labeled during the incremental procedure, i.e. when the set X_L absorbs all the instances initially in X_U, or in an equivalent way when X_U is reduced to the empty set $\emptyset$. To analyze the convergence of our approach, we can formulate the evolution of X_L and X_U over time, as follows :

Let m_t be the size of the set X_L at time (iteration) t, the evolution of m_t is subject to the following nonlinear dynamical system (R) :

$$(R) \ : \begin{cases} m_t = m_{t-1} + \zeta_t \\ m_0 = l \end{cases} \tag{6.14}$$

where ζ_t is the number of instances in X_U that have been labeled during the iteration t. Since in an iteration t, we can label all the instances in X_U if the parameter α is too weak, or no instance if α is very large, then the terms of the sequence $(\zeta_t)_t$ can vary between 0 and u, and we have $\sum_{t\geq 1} \zeta_t = u$.

Symmetrically, let n_t be the size of the set X_U at time t, the evolution of n_t is subject to the following nonlinear dynamical system (S) :

$$(S) \ : \begin{cases} n_t = n_{t-1} - \zeta_t \\ n_0 = u \end{cases} \tag{6.15}$$

From a theoretical point of view, our algorithm OTP must converge at the instant $t = \tau$, which verifies : $m_\tau = m_0 + \sum_{t=1}^{\tau} \zeta_t = m_0 + u = l + u$, which corresponds also to $n_\tau = n_0 - \sum_{t=1}^{\tau} \zeta_t = n_0 - u = u - u = 0$.

The question is whether OTP will reach the instant τ in a finite number of iterations. Experiments have shown that a suitable choice of α will allow us to label a large amount ζ_t of samples in X_U at each iteration t, otherwise, it suffices to decrease α in the following way: suppose that at an iteration t, we have g unlabeled samples, whose certainty score s_j is lower than the threshold α, which means that none of these examples can be labeled according to OTP procedure at the iteration t, the solution lies then in decreasing the value of α as follows :

$$\alpha \leftarrow \alpha - \min_{x_j \in [X_U]_t} (\alpha - s_j), \tag{6.16}$$

we denote by $[X_U]_t$ the set of the g points constituting X_U at iteration t. Decreasing the value of α in this way will allow the point with the greatest certainty score in $[X_U]_t$ to be labeled, and then to migrate from X_U to X_L.

Certainly, decreasing sharply the value of α will allow us to label many other instances instead of just the sample with the highest certainty score, however, this gain in terms of the number of points labeled at the same iteration will be paid out in terms of its predictions certainty. Our intuition behind the modification of α in the way above is as follows: Since moving an instance from X_L to X_U, can change the optimal transport plan between the new distributions μ and ν, and subsequently the certainty scores in the next iteration, we can try to restore the initial value of α and continue to label the other samples with the same degree of certainty as before. If the same scenario is repeated in a future iteration, we can use the same technique of decreasing α to label a new point, and so on until convergence. This reasoning shows that the proposed algorithm needs effectively a finite number of iterations to converge.

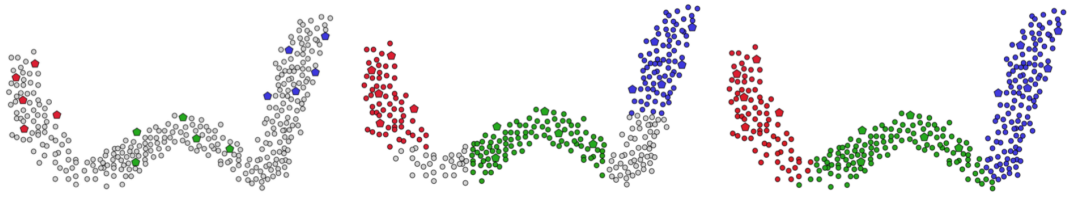

FIGURE 6.2: Illustration of the label propagation process (from the left to the right): at the initial iteration $t = 0$, at an intermediate iteration $0 < t < \tau$, and at the last iteration $t = \tau$. Pentagon markers correspond to the labeled instances and circles correspond to the unlabeled ones which are gradually pseudo-labeled by OTP. The class is color-coded.

6.4 Optimal Transport Induction

In the previous section, we have introduced the main process of OTP, but it is just for the transductive task. In a truly inductive setting, where new examples are given one after the other and a prediction must be given after each example, the use of the transductive algorithm again to get a label prediction for the new instances is very computationally costly, since it needs to be rerun in their entirety, which is unpleasant in many real-world problems, where on-the-fly classification for previously unseen instances is indispensable.

In this section, we propose an efficient way to extend OTP for out-of-sample data. In fact, we will fix the transductive predictions $\{y_{l+1}, ..., y_{l+u}\}$ and based on the objective function of our transductive algorithm OTP we will try to extend the resulting graph to predict the label of previously unseen instances.

OTP approach can be cast as the minimization of the objective function $C_{\mathcal{W},l}$ in terms of the label function values at the unlabeled samples $x_j \in X_U$:

$$C_{\mathcal{W},l}^{transduction}(f) = \sum_{x_i \in X_L} \sum_{x_j \in X_U} w_{x_i,x_j} l(y_i, f(x_j)) \tag{6.17}$$

where l is an unsupervised loss function. The objective function in (6.17) is a smoothness criterion that lies for penalizing differences in the label predictions for connected samples in the graph, which means that a good classifying function should not change too much between similar instances.

To transform the above transductive algorithm into function induction for out-of-sample data, we need to use the same type of smoothness criterion as before for a new testing instance x_{new}, and then we can optimize the objective function with respect to only the predicted label $\tilde{f}(x_{new})$ (Bengio et al., 2006). The smoothness criterion for a new test point x_{new} becomes then :

$$C_{\mathcal{W},l}^{induction}(\tilde{f}(x_{new})) = \sum_{x_i \in X_L \cup X_U} w_{x_i,x_{new}} l(y_i, \tilde{f}(x_{new})) \tag{6.18}$$

If the loss function l is convex, e.g. $l = (y_i - \tilde{f}(x_{new}))^2$, then the cost function $C_{\mathcal{W},l}^{induction}$ is also convex in $\tilde{f}(x_{new})$, the label assignment $\tilde{f}(x_{new})$ minimizing $C_{\mathcal{W},l}^{induction}$ is then given by :

$$\tilde{f}(x_{new}) = \frac{\sum_{x_i \in X_L \cup X_U} w_{x_i,x_{new}} y_i}{\sum_{x_i \in X_L \cup X_U} w_{x_i,x_{new}}} \tag{6.19}$$

6.4.1 Binary classification and multi-class settings

In a binary classification context, where $\mathcal{Y} = \{+1, -1\}$, the classification problem is transformed into a regression one, in a way that the predicted class of x_{new} is thus $sign(\tilde{f}(x_{new}))$.

$$\begin{cases} \tilde{f}(x_{new}) = +1 \text{ if } sign(\tilde{f}(x_{new})) \geq 0, \\ \tilde{f}(x_{new}) = -1 \text{ otherwise.} \end{cases}, \tag{6.20}$$

While most transductive algorithms can handle multiple classes, the inductive methods mostly only work in the binary classification setting, where $\mathcal{Y} = \{+1, -1\}$. Following the same logic as (Delalleau et al., 2005), our optimal transport approach

can be adapted and extended accurately for multi-class settings, in the following way: the label $\tilde{f}(x_{new})$ is given by the weighted majority vote of other samples in $X = X_L \cup X_U$:

$$\tilde{f}(x_{new}) = \underset{c_h \in \mathcal{C}}{argmax} \sum_{x_i \in X_L \cup X_U / y_i = c_h} w_{x_i, x_{new}} \tag{6.21}$$

The predicted class of x_{new} is then the class whose representatives have the highest similarity with x_{new}.

Equation (6.20) can be seen as a special case of (6.21) in the binary classification settings, in fact, if $\mathcal{Y} = \{+1, -1\}$, then choosing between the class that maximizes $\sum_{x_i \in X_L \cup X_U / y_i = c_k} w_{x_i, x_{new}}$ is equivalent to choosing according to the sign of $\tilde{f}(x_{new})$, since the term $\sum_{x_i \in X_L \cup X_U} w_{x_i, x_{new}}$ in (6.19) is always positive.

6.4.2 Transduction-induction consistency

It would be very interesting to see what happens when we apply the induction formula (6.18) on a point x_j of X_U. Ideally, the induction formula must be consistent with the prediction get it by the transduction formula (6.17) for an instance $x_j \in X_U$. For $x_{new} = x_j$, $j \in \{l+1, ..., l+u\}$, we have:

$$\frac{\partial C_{\mathcal{W},l}^{transduction}}{\partial f(x_j)} = -2 \sum_{x_i \in X_L} w_{i,j}(y_i - f(x_j)) \tag{6.22}$$

$C_{\mathcal{W},l}^{transduction}$ is convex in $f(x_j)$, and is minimized when :

$$\begin{aligned} f(x_j) &= \frac{\sum_{x_i \in X_L} w_{x_i, x_j} y_i}{\sum_{x_i \in X_L} w_{x_i, x_j}} \\ &= \frac{\sum_{x_i \in X_L \cup X_U} w_{x_i, x_j} y_i}{\sum_{x_i \in X_L \cup X_U} w_{x_i, x_j}} \quad \text{since } w_{x_i, x_j} = 0 \, \forall x_i \in X_U \quad (\mathcal{G} \text{ is a bipartite graph}) \\ &= \tilde{f}(x_j) \end{aligned} \tag{6.23}$$

hence the consistency.

Our proposed algorithm for the inductive task called Optimal Transport Induction (OTI), is summarized in Algorithm 6.2, where we use the algorithm (OTP) for training and (6.21) for testing.

Algorithm 6.2 OTI

Parameters: ε, α
Input : x_{new}, X_L, X_U, Y_L
(1) Training phase
Get Y_U by Algorithm 6.1
(2) Testing phase
For a new point x_{new}, compute its label $\tilde{f}(x_{new})$ (6.21)
return $\tilde{f}(x_{new})$

6.5 Experimental results

In this section, we evaluate our method on a various real-world datasets.

6.5.1 Datasets

The experiment was designed to evaluate the proposed approach on 12 benchmark datasets. Details of these datasets appear in Table 6.1.

TABLE 6.1: Experimental datasets

Datasets	#Instances	#Features	#Classes
Iris	150	4	3
Wine	178	13	3
Heart	270	13	2
Ionosphere	351	34	2
Dermatology	366	33	6
Breast	569	31	2
WDBC	569	32	2
Isolet	1560	617	26
Waveform	5000	21	3
Digits	5620	64	10
Statlog	6435	36	6
MNIST	10000	784	10

6.5.2 Evaluation measures

Three widely used evaluation measures were employed to evaluate the performance of the proposed approach: the accuracy (ACC) (Liu et al., 2019), the Normalized Mutual Information (NMI) (Dom, 2012), and the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985).

The accuracy (ACC) is the percentage of correctly classified samples, formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{6.24}$$

Normalized Mutual Information (NMI) is a normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation). In this function, mutual information is normalized by some generalized mean of true labels Y and predicted labels $\hat{Y}$.

$$\text{NMI}(Y, \hat{Y}) = \frac{2\text{I}(Y,\hat{Y})}{\text{H}(Y)+\text{H}(\hat{Y})}$$

where I is the mutual information of Y and $\hat{Y}$, defined as: $\text{I}(Y, \hat{Y}) = \text{H}(Y) - \text{H}(Y|\hat{Y})$ with H is the entropy defined by: $\text{H}(Y) = \sum_y p(y)\log(p(y))$

The Rand index is a measure of the similarity between two partitions A and B and is calculated as follows :

$$\text{Rand}(A, B) = \frac{a+d}{a+b+c+d}$$

where a is the number of pairs of elements that are placed in the same cluster in A and the same cluster in B, b denotes the number of pairs of elements in the same cluster in A but not in the same cluster in B, c is the number of pairs of elements in the same cluster in A but not in the same cluster in B and d denotes the number of pairs of elements in different clusters in both partitions. The values a and d can be interpreted as agreements, and b and c as disagreements.

The Rand index is then "adjusted for chance" into the ARI using the following scheme:

$$\text{ARI} = \frac{\text{Rand} - \text{ExpectedRand}}{\text{maxRand} - \text{ExpectedRand}}$$

The adjusted Rand index is thus ensured to have a value close to 0 for random labeling independently of the number of clusters and samples and exactly 1 when the clustering is identical (up to a permutation).

6.5.3 Experimental protocol

Experiments compared the proposed algorithm with three semi-supervised approaches, including LP (Zhou et al., 2003) and LS (Zhu and Ghahramani, 2002), which are the classical label propagation algorithms, LNP (Wang and Zhang, 2007), which is an improved label propagation algorithm with modified affinity matrix, and the spectral clustering algorithm SC (Ng et al., 2001) without prior information.

To compare these different algorithms, their related parameters were specified as follows :

- The number of clusters k for spectral clustering was set equal to the true number of classes on each dataset.
- Each of the compared algorithms LP, LS, and NLP, require a Gaussian kernel controlled by a free parameter σ to be specified to construct the affinity matrix, in the comparisons, each of these algorithms was tested with different σ values, and its best result with the highest ACC, NMI and ARI values on each dataset was selected.
- The efficiency of a semi-supervised algorithm depends on the amount of prior information. Therefore, in the experiment, the amount of prior information data was set to 15, 25, and 35 percent of the total number of samples included in a dataset.
- The effectiveness of a semi-supervised approach depends also on the quality of prior information. Therefore, in the experiment, given the amount of prior information, all the compared algorithms were run with 10 different sets of prior information to compute the average results for ACC, NMI, and ARI on each dataset.
- To give an overall vision of the best approach on all the datasets, we define the following score:

$$\text{SCORE}(A_i) = \sum_j \frac{\text{Perf}(A_i, D_j)}{\max_i \text{Perf}(A_i, D_j)} \tag{6.25}$$

  where Perf indicates the performance according to one of the three evaluation measures above of each approach A_i on each datasets D_j.

6.5.4 Results

TABLE 6.2: Performances according to Accuracy values

Datasets	Percent	LP	LS	LNP	OTP	SC
	15%	0.9437	0.9453	0.8852	**0.9507**	
Iris	25%	0.9531	0.9540	0.9261	**0.9610**	0.7953
	35%	0.9561	0.9571	0.9392	**0.9796**	
	15%	**0.9296**	0.9296	0.8462	0.9250	
Wine	25%	**0.9417**	0.9417	0.8597	0.9343	0.8179
	35%	**0.9482**	0.9482	0.8727	0.9388	
	15%	0.7261	0.7304	0.5683	**0.7696**	
Heart	25%	0.7734	0.7833	0.6826	**0.8424**	0.3411
	35%	0.8239	0.8352	0.7731	**0.8693**	
	15%	0.8300	0.8310	0.8051	**0.8796**	
Ionosphere	25%	0.8439	0.8462	0.8146	**0.8871**	0.4461
	35%	0.8458	0.8476	0.8293	**0.8978**	
	15%	0.9324	0.9327	0.8948	**0.9488**	
Dermatology	25%	0.9438	0.9438	0.9163	**0.9520**	0.4943
	35%	0.9536	0.9536	0.9428	**0.9566**	
	15%	0.9566	0.9566	0.9153	**0.9587**	
Breast	25%	0.9578	0.9578	0.9296	**0.9649**	0.7830
	35%	0.9649	0.9649	0.9427	**0.9730**	
	15%	1.0000	1.0000	0.9568	**1.0000**	
WDBC	25%	1.0000	1.0000	0.9879	**1.0000**	0.9682
	35%	1.0000	1.0000	0.9970	**1.0000**	
	15%	0.7558	0.7558	0.6519	**0.7559**	
Isolet	25%	**0.7782**	0.7782	0.6908	0.7767	0.5385
	35%	**0.8077**	0.8077	0.7249	0.8053	
	15%	0.8318	0.8334	0.7719	**0.8469**	
Waveform	25%	0.8401	0.8419	0.7892	**0.8504**	0.3842
	35%	0.8423	0.8425	0.8062	**0.8599**	
	15%	0.9589	0.9589	0.9363	**0.9678**	
Digits	25%	0.9737	0.9737	0.9571	**0.9774**	0.7906
	35%	0.9801	0.9801	0.9784	**0.9827**	
	15%	**0.8740**	0.8730	0.8249	0.8516	
Statlog	25%	**0.8779**	0.8771	0.8371	0.8533	0.6516
	35%	**0.8831**	0.8821	0.8474	0.8538	
	15%	0.9210	0.9218	0.8247	**0.9421**	
MNIST	25%	0.9460	0.9451	0.8371	**0.9540**	0.5719
	35%	0.9551	0.9571	0.8408	**0.9632**	
ALL Datasets	SCORE	35.4544	35.4975	33.3619	**35.8855**	8.1971

TABLE 6.3: Performances according to NMI values

Datasets	Percent	LP	LS	LNP	OTP	SC
	15%	0.8412	0.8442	0.7534	**0.8447**	
Iris	25%	0.8584	0.8621	0.8269	**0.8667**	0.7980
	35%	0.8621	0.8649	0.8314	**0.8852**	
	15%	**0.7821**	0.7821	0.6815	0.7384	
Wine	25%	**0.8127**	0.8127	0.7573	0.7790	0.7808
	35%	**0.8289**	0.8289	0.7897	0.7963	
	15%	0.1519	0.1575	0.1091	**0.2181**	
Heart	25%	0.2291	0.2472	0.1432	**0.3683**	0.1880
	35%	0.3313	0.3546	0.2718	**0.4374**	
	15%	0.3502	0.3535	0.3256	**0.4676**	
Ionosphere	25%	0.3848	0.3911	0.3572	**0.5000**	0.2938
	35%	0.3972	0.4014	0.3725	**0.5383**	
	15%	0.8770	0.8779	0.8349	**0.8935**	
Dermatology	25%	0.8932	0.8932	0.8692	**0.9033**	0.6665
	35%	0.9128	0.9128	0.8959	**0.9164**	
	15%	0.7340	0.7360	0.6971	**0.7449**	
Breast	25%	0.7451	0.7465	0.7192	**0.7550**	0.6418
	35%	0.7909	0.7909	0.7706	**0.8106**	
	15%	1.0000	1.0000	0.9049	**1.0000**	
WDBC	25%	1.0000	1.0000	0.9347	**1.0000**	0.9163
	35%	1.0000	1.0000	0.9715	**1.0000**	
	15%	**0.7785**	0.7785	0.7184	0.7657	
Isolet	25%	**0.7987**	0.7987	0.7503	0.7852	0.7545
	35%	**0.8210**	0.8210	0.7869	0.8077	
	15%	0.4950	0.5009	0.4628	**0.5256**	
Waveform	25%	0.5124	0.5192	0.4763	**0.5319**	0.3646
	35%	0.5192	0.5229	0.4807	**0.5421**	
	15%	0.9150	0.9150	0.8891	**0.9290**	
Digits	25%	0.9443	0.9443	0.9268	**0.9489**	0.8483
	35%	0.9570	0.9570	0.9318	**0.9607**	
	15%	**0.7396**	0.7383	0.6792	0.6753	
Statlog	25%	**0.7483**	0.7477	0.6859	0.6800	0.6139
	35%	**0.7572**	0.7571	0.6907	0.6821	
	15%	0.8019	0.8028	0.7759	**0.8177**	
MNIST	25%	0.8389	0.8367	0.7931	**0.8442**	0.6321
	35%	0.8542	0.8599	0.8136	**0.8730**	
ALL Datasets	SCORE	33.9980	34.2042	31.6326	**35.5237**	9.5770

TABLE 6.4: Performances according to ARI values

Datasets	Percent	LP	LS	LNP	OTP	SC
	15%	0.8453	0.8492	0.7861	**0.8621**	
Iris	25%	0.8680	0.8704	0.8321	**0.8884**	0.7455
	35%	0.8754	0.8783	0.8424	**0.9027**	
	15%	**0.7936**	0.7936	0.7148	0.7814	
Wine	25%	**0.8267**	0.8267	0.7346	0.8050	0.7912
	35%	**0.8455**	0.8455	0.7741	0.8192	
	15%	0.2110	0.2190	0.1562	**0.2875**	
Heart	25%	0.3176	0.2955	0.2283	**0.4662**	0.2031
	35%	0.4163	0.4464	0.3688	**0.5430**	
	15%	0.4221	0.4248	0.3998	**0.5723**	
Ionosphere	25%	0.4606	0.4673	0.4324	**0.5927**	0.3971
	35%	0.4650	0.4702	0.4418	**0.6281**	
	15%	0.8807	0.8813	0.8438	**0.8996**	
Dermatology	25%	0.8972	0.8972	0.8751	**0.9093**	0.4783
	35%	0.9146	0.9146	0.9007	**0.9218**	
	15%	0.8328	0.8327	0.7956	**0.8404**	
Breast	25%	0.8371	0.8284	0.8039	**0.8636**	0.7018
	35%	0.8632	0.8632	0.8413	**0.8940**	
	15%	1.0000	1.0000	0.9349	**1.0000**	
WDBC	25%	1.0000	1.0000	0.9691	**1.0000**	0.9565
	35%	1.0000	1.0000	0.9905	**1.0000**	
	15%	**0.6002**	0.6002	0.5064	0.5998	
Isolet	25%	**0.6333**	0.6332	0.5526	0.6299	0.5284
	35%	**0.6735**	0.6735	0.5992	0.6683	
	15%	0.5639	0.5678	0.5163	**0.5945**	
Waveform	25%	0.5819	0.5864	0.5279	**0.6031**	0.3788
	35%	0.5870	0.5880	0.5342	**0.6182**	
	15%	0.9126	0.9127	0.8993	**0.9306**	
Digits	25%	0.9432	0.9432	0.9287	**0.9508**	0.7846
	35%	0.9567	0.9567	0.9407	**0.9621**	
	15%	**0.7658**	0.7640	0.7167	0.7122	
Statlog	25%	**0.7730**	0.7714	0.7318	0.7284	0.6031
	35%	**0.7820**	0.7806	0.7391	0.7336	
	15%	0.7930	0.7944	0.7697	**0.8393**	
MNIST	25%	0.8487	0.8466	0.8152	**0.8685**	0.5153
	35%	0.8721	0.8777	0.8438	**0.8935**	
ALL Datasets	SCORE	33.9814	34.0574	31.6888	**35.7239**	8.8581

Tables 6.2, 6.3 and 6.4 list the performance of the different algorithms on all the datasets. These comparisons indicate that the proposed algorithm is superior to the spectral clustering algorithm, this suggests that prior information can improve the label propagation effectiveness, this statement is also confirmed by the fact that given the datasets, all the label propagation algorithms show growth in their performance in parallel with the increase of the amount of prior information. Furthermore, the tables show that the proposed approach is clearly more accurate than LP, LS, and NLP on most tested datasets. However, on some datasets, OTP performed slightly less accurately than LP. The tables also present the proposed score results of each algorithm, which show that the best score belongs to the proposed label propagation approach based on optimal transport, followed by LS and LP.

6.5.5 Friedman and Nemenyi tests

To confirm the superiority of our algorithm over the compared approaches, and especially LP, we suggest using the Friedman test and Nemenyi test (Demšar, 2006). First, algorithms are ranked according to their performance on each dataset, then there are as many rankings as there are datasets. The Friedman test is then conducted to test the null hypothesis under which all algorithms are equivalent, and in this case, their average ranks should be the same. If the null hypothesis is rejected, then the Nemenyi test will be performed. If the average ranks of two approaches differ by at least the critical difference (CD), then it can be concluded that their performances are significantly different. We set the significance level $\alpha = 0.05$. Figure 6.3 shows a critical diagram representing a projection of the average ranks of the algorithms on the enumerated axis. The algorithms are ordered from left (the best) to right (the worst) and a thick line connects the groups of algorithms that are not significantly different (for the significance level $\alpha = 5\%$). As shown in figure 6.3, OTP seems to achieve a significant improvement over LNP and SC, in fact, for the three evaluation measures, the statistical hypothesis test shows that OTP is more efficient. For LS and LP, we can see that OTP is ahead of them, but the difference is not statistically very significant.

6.5.6 Sensitivity analysis

To further show how our approach compares to that of our competitors, we are conducting a sensitivity analysis using the Box-Whisker plots (Turkey, 1977). Box-Whisker plots are a non-parametric method to represent graphically groups of numerical data through their quartiles, in order to study their distributional characteristics. In figure 6.4, for each evaluation measure, Box-Whisker plots are drawn from the performance of our algorithm and the compared ones over all the tested datasets. To begin with, performances are sorted. Then four equal-sized groups are made from the ordered scores. That is, 25% of all performances are placed in each group. The lines dividing the groups are called quartiles, and the four groups are referred to as quartile groups. Usually, we label these groups 1 to 4 starting at the bottom. In a Box-Whisker plot, the ends of the box are the upper and lower quartiles, so the box spans the interquartile range, the median is marked by a vertical line inside the box, the whiskers are the two lines outside the box that extend to the highest and lowest observations.

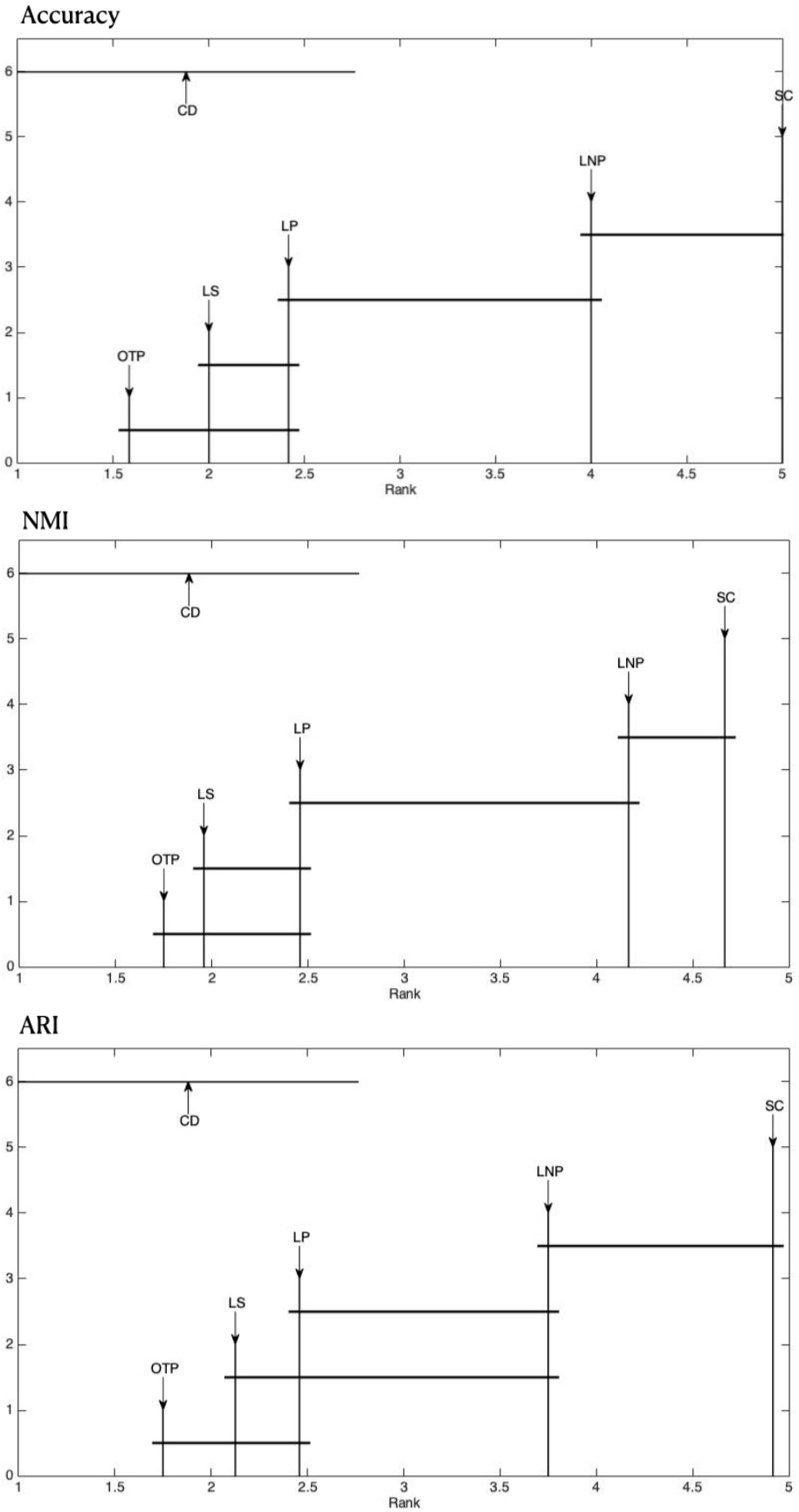

FIGURE 6.3: Friedman and Nemenyi tests: approaches are ordered from left (the best) to right (the worst).

Sensitivity Box-Whisker plots represent a synthesis of the performances into five crucial pieces of information identifiable at a glance: position measurement, dispersion, asymmetry, length of Whiskers and outliers. The position measurement is characterized by the dividing line on the median. Dispersion is defined by the length of the Box-Whiskers. Asymmetry is defined as the deviation of the median line from the center of the Box-Whiskers. The length of the Whiskers is the distance between the ends of the Whiskers to the length of the Box-Whiskers. Outliers are plotted as individual points.

Figure 6.4 shows further details on the performance of our algorithm for the three evaluation measures. Indeed, regarding the accuracy, we note that the Box-Whisker plot corresponding to OTP is comparatively short, this suggests that, overall, its performance on the different datasets has a high level of agreement with each other, implying stability comparable to that of LP and LS, and significantly better than that of LNP and SC. For NMI, the Box-Whisker plot corresponding to our approach is much higher than that of LNP and SC, also noting the presence of 2 outliers for LP and LS, these outliers correspond to Heart and Ionosphere datasets, where both approaches have achieved very low scores, on the other hand, there is an absence of outliers for OTP, these indicators confirm the improvement in terms of NMI by our approach over LP and LS. Concerning ARI, we notice that the medians of LP, LS, and OTP are all at the same level, however, the Box-Whisker of OTP is comparatively short, implying better stability.

The sensitivity analysis above confirms the superiority of our approach over LNP and SC and also shows some points of difference between our algorithm, LP and LS, which are rather in favor of OTP, such as the absence of outliers and the comparatively short length of OTP's Box-Whisker.

These results are mainly attributed to the ability of the proposed algorithm to capture much more information than the previous algorithms thanks to the enhanced affinity matrix constructed by optimal transport. It is equally noteworthy that the effectiveness of the proposed algorithm lies in the fact that the incremental process takes advantage of the dependency of semi-supervised algorithms on the amount of prior information, then the enrichment of the labeled set at each iteration with new data allows to the unlabeled instances to be labeled with high certainty. Another reason for the superiority of OTP over the other algorithms is its capacity to control the certitude of the label predictions thanks to the certainty score used, which allows instances to be labeled only if they have a high degree of prediction certainty.

6.6 Software

We make our code publicly available at:

https://github.com/MouradElHamri/OTP

The used datasets are available at:

https://archive.ics.uci.edu/

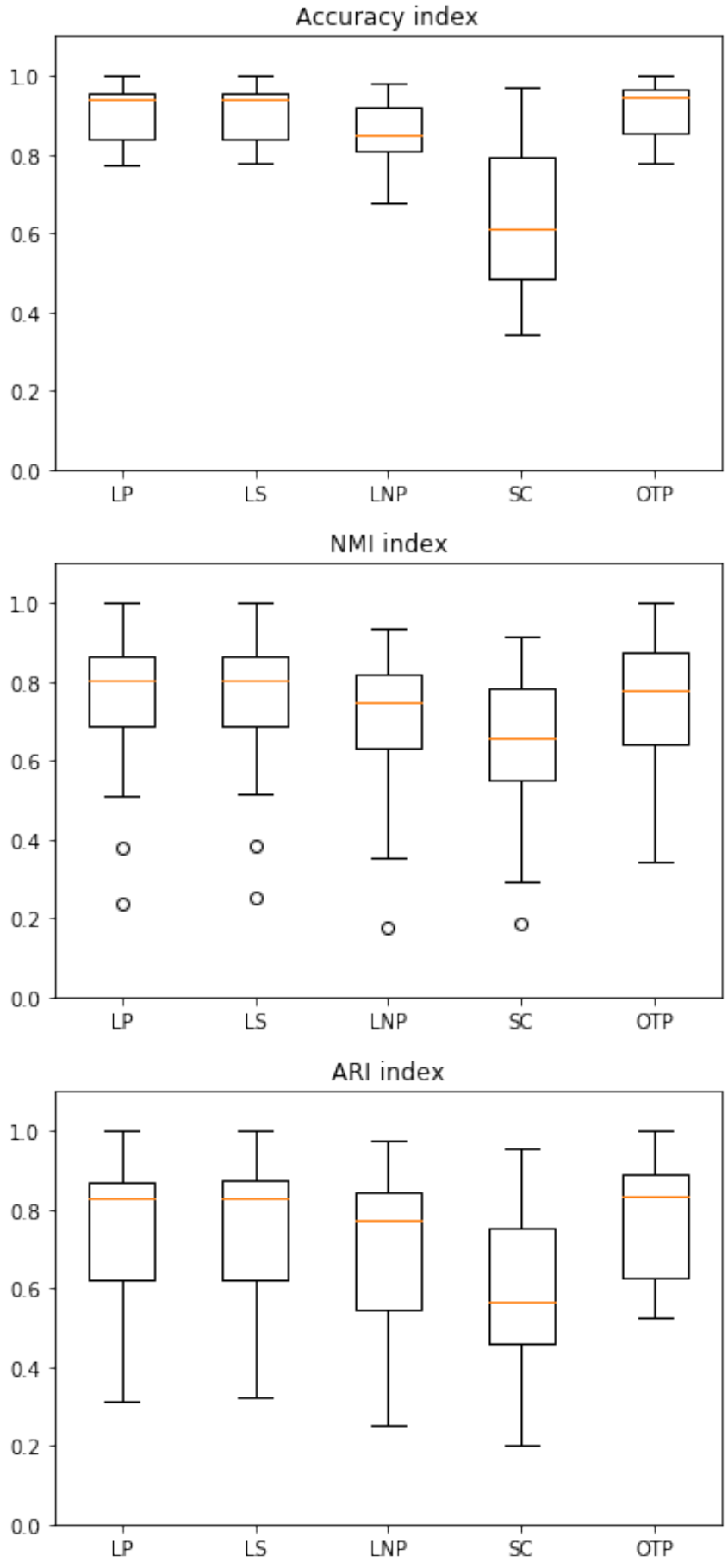

FIGURE 6.4: Sensitivity analysis using Box-Whiskers plots

6.7 Conclusion and future perspectives

Motivated by the necessity to leverage other ways of learning hidden structures in the target domain, we have addressed in this chapter the limitations of traditional label propagation methods by proposing a principally new approach based on optimal transport named OTP. The proposed approach consists in inferring an improved affinity matrix from the optimal transport plan between empirical measures defined on labeled and unlabeled data. Furthermore, to take advantage of the reliance of semi-supervised methods to the amount of prior information, we adopted an incremental process to propagate labels through the edges of a complete bipartite edge-weighted graph. To reinforce the certitude of the predictions, we incorporated a certainty score that controls the incremental propagation process. We also provided a convergence analysis for the proposed approach and an extension to out-of-sample data. Experiments have shown that OTP outperforms current state-of-the-art methods.

The work of this chapter can be extended in different directions:

- From a theoretical standpoint, we plan to develop theoretical analysis of semi-supervised learning with optimal transport.
- From an algorithmic standpoint, we intend to investigate how one can take advantage of the proposed algorithm to solve the unsupervised domain adaptation problem. Indeed we wish using OTP to perform progressive and certain labeling of the target data after projecting them in conjunction with the source data in discriminating reduced spaces.

CHAPTER 7

WHEN DOMAIN ADAPTATION MEETS SEMI-SUPERVISED LEARNING THROUGH OPTIMAL TRANSPORT

Contents

This chapter, based on (El Hamri et al., 2022a,c) deals with the problem of unsupervised domain adaptation, using the semi-supervised technique (OTP) developed in (El Hamri et al., 2021c). OTP will be used to learn hidden structures in the target domain in order to use them to incrementally create augmented source structures [1]. This will allow learning a sequence of discriminative and domain-invariant latent subspaces based on Linear Discriminant Analysis, within which it becomes easy to progressively label the target samples in a self-training fashion. A theoretical analysis of self-training methods can be used to explain the good empirical behavior of our approach.

[1] Augmented source structures are composed of labeled source samples and pseudo-labeled target samples.

> True knowledge leads to humility. The more a person knows, the more they realize they know nothing.
>
> Ibn Arabi

7.1 Introduction

A common practice of domain adaptation approaches focuses on matching marginal source and target distributions by learning a domain-invariant joint subspace. However, an exact domain-level alignment does not imply a fine-grained class-to-class overlap, since the conditional distribution of the target domain can be misaligned with that of the source domain, which implies that, the latent subspace may not only push the source and target domains closer but also confuse instances with different class labels. To prevent this, further directions were pursued by incorporating additional structural information contained in the unlabeled target domain.

Structural information in the target domain can be acquired using unsupervised learning techniques such as clustering or semi-supervised learning techniques such as pseudo-labeling. In this chapter, we focus on the second family of domain adaptation approaches based on pseudo-labels (Long et al., 2013; Wang et al., 2018; Xie et al., 2018).

The problem with these pseudo-labeling methods is their heavy reliance on the assumption that correctly pseudo-labeled data can reduce the bias caused by falsely pseudo-labeled ones. Whereas, in reality, falsely pseudo-labeled instances in the early iterations of the learning process can potentially lead to catastrophic damage due to the accumulation of errors in the subsequent iterations.

To address this issue, selective pseudo-labeling was employed in (Chen et al., 2019; Wang and Breckon, 2020; Gallego et al., 2020). Selective pseudo-labeling takes into account the certainty of predictions in the target domain. Precisely, these methods operate in the following manner: a small amount of target instances are selected to be assigned with pseudo-labels, and only these selected pseudo-labeled target samples are integrated with the labeled source data in the next iteration of the learning process. This makes these methods a specific form of self-training, which is a popular technique that has proven to be very effective for learning with unlabeled data.

Despite their very intuitive nature and good empirical performance, these methods have bottlenecks related to the amount and the way the pseudo-labels are selected. Moreover, they lack theoretical guarantees.

Contribution: To address these limitations, we propose in this chapter to learn the hidden structures in the target domain using the label propagation approach in (El Hamri et al., 2021c). From these, we select a subset of pseudo-labeled samples to create augmented source structures that will be used to learn incrementally a sequence of discriminative and domain-invariant latent subspaces, within which it becomes easy to progressively label the target samples. The proposed approach is backed up by a theoretical analysis of self-training.

Outline: The rest of this chapter is organized as follows: in the 2^{nd} section, we elaborate the proposed approach OTP-DA. In the 3^{th} section, we present a theoretical analysis of self-training methods that can be extended to the proposed approach. In the 4^{th} section, we evaluate our algorithm on two benchmark datasets. Finally, we conclude in section 5.

7.2 Optimal Transport Propagation for Domain Adaptation

The proposed approach OTP-DA aims to learn a joint subspace from the source and target domains such that the projected data into this subspace are domain invariant and well separated. To accomplish this aim, linear discriminant analysis (LDA) appears to be a good candidate for many reasons, principally for its capacity to find a linear combination of features, which separates two or more classes of data no matter the domain they come from, providing an appropriate approach for the unsupervised domain adaptation problem, where the source and target data come from different distributions. Nonetheless, LDA needs labeled data to learn the projection matrix. To surmount this challenge, we use pseudo-labels in the target domain produced by OTP (El Hamri et al., 2021c).

The reason for choosing OTP is its ability to capture the geometry of data thanks to optimal transport. Furthermore, OTP falls into the class of selective pseudo-labeling methods, so it avoids mislabeled target instances from impeding the subspace learning process by spreading the errors to the next iteration, which can reduce the robustness of the learned classifier.

Thus, we use the labeled source data and the selected pseudo-labeled target data provided by OTP to incrementally learn a sequence of lower-dimensional domain-invariant and discriminative latent subspaces where a classifier can progressively label the target samples in a self-training fashion.

7.2.1 Domain Alignment via Linear Discriminant Analysis

To learn a domain-invariant and discriminative subspace $\tilde{\mathcal{X}}$ from $\mathcal{X}$, we use Linear Discriminant Analysis (LDA) (Fisher, 1936), which is a common technique used for dimensionality reduction. LDA can also provide class separability by drawing a decision region between the different classes.

Let $X \in \mathcal{M}_{d,n}(\mathbb{R})$ be a labeled data matrix composed of n samples. Basically, LDA seeks to find a projection matrix W for which the low-dimensional projection of X yields a cloud of points that are close when they are in the same class relative to the overall spread. This projection matrix can be found by maximizing the Rayleigh quotient of the within scatter matrix S_w and between scatter matrix S_b:

$$W = \underset{V}{\text{argmax}} \frac{|V^T S_b V|}{|V^T S_w V|} \tag{7.1}$$

The maximization problem in (7.1) is equivalent to the following generalized eigenvalue problem:

$$S_b w = \lambda S_w w \tag{7.2}$$

The eigenvectors of (7.2) represent the directions of the lower-dimensional feature space learned by LDA, and the corresponding eigenvalues represent the ability of the eigenvectors to discriminate between different classes, i.e. increase the between-class variance, and decrease the within-class variance of each class. The eigenvectors with the d_1 highest eigenvalues give us the LDA projection matrix $W_{=}[w_1, ..., w_{d_1}] \in \mathcal{M}_{d,d_1}(\mathbb{R})$, from which we can learn the lower-dimensional discriminant representation $\tilde{X} \in \mathcal{M}_{d_1,n}(\mathbb{R})$:

$$\tilde{X} = W^T X \tag{7.3}$$

7.2.2 Self-Training via Optimal Transport Propagation

To learn a domain-invariant and discriminative subspace $\tilde{\mathcal{X}}$ from $\mathcal{X}$ using the projection matrix W of LDA we need labeled data as stated above. Nevertheless, in unsupervised domain adaptation settings, labeled data in the target domain are unavailable. To address this limitation, we propose to use Optimal Transport Propagation (OTP) to perform selective pseudo-labeling in the target domain.

Once the LDA projection matrix W is learned [2], the projection of both source samples S and target samples T in the joint subspace can be obtained as follows:

$$\tilde{S} = W^T S \quad \text{and} \quad \tilde{T} = W^T T \tag{7.4}$$

Pseudo-labeling in the target domain can then be performed using OTP considering that:

$$X_L = \tilde{S} \quad \text{and} \quad X_U = \tilde{T} \tag{7.5}$$

The intuition behind the use of OTP as a pseudo-labeling technique is its capability to capture the geometry of the underlying subspace and its selective ability based on the incorporated certainty score which makes it closely related to entropy minimization, where the model's predictions are encouraged to be low-entropy (i.e., high-confidence) on unlabeled data.

Thus, instead of using all the pseudo-labeled target samples to learn the next projection, we incrementally select a subset $\tilde{T}_p \subset \tilde{T}$ that contains an amount of p pseudo-labeled target samples with the highest certainty score.

Nevertheless, this technique has the potential risk of only selecting instances from particular classes and overlooking the other ones. To prevent this issue, we conduct a class-wise selection in order to ensure that pseudo-labeled target samples of each class have an equal opportunity to be selected. Precisely, for each class c_h, $\forall h \in \{1, ..., k\}$ we select p/k target samples pseudo-labeled as class c_h.

Thereafter, the projected source data is combined with the selected pseudo-labeled target data to form augmented source structures, simultaneously, the pseudo-labeled target data must be removed from the target domain in the following way:

$$S \leftarrow \tilde{S} \cup \tilde{T}_p \quad \text{and} \quad T \leftarrow \tilde{T} \setminus \tilde{T}_p \tag{7.6}$$

[2] At the first iteration, the projection matrix is learned using only the labeled source data

Equations (7.6) are used to incrementally update the source and target domains. At each iteration, a classifier η is trained on the augmented source samples in a self-training manner (using labeled source data in conjunction with selected pseudo-labeled target data).

The intuition behind this idea is that at each iteration the classifier becomes more and more robust since it is trained on both the source data and the selected pseudo-labeled target data so that in the last iteration, it will be trained on the source samples and the totality of pseudo-labeled target instances.

The overall OTP-DA algorithm is summarized in Algorithm 7.1.

Algorithm 7.1 OTP-DA

Parameters: Dimensionality of LDA d_1, sampling rate p
Input : Labeled source data S, Unlabeled target data T
while *not converged* **do**
  Learn the projection W using source data S (7.2)
  Get the projected source and target samples $\tilde{S}$ and $\tilde{T}$ (7.4)
  Assign pseudo-labels for the projected target data $\tilde{T}$ using OTP (7.5)
  Select a subset of pseudo-labeled target data $\tilde{T}_p$
  Update the source domain $S \leftarrow \tilde{S} \cup \tilde{T}_p$ (7.6)
  Update the target domain $T \leftarrow \tilde{T} \setminus \tilde{T}_p$ (7.6)
  Learn a classifier η on S
end
return *Predicted labels of the original target data* T *using* η

Figure 7.1 below provides an overview of the OTP-DA approach. For the sake of clarity, we omit the incremental aspect.

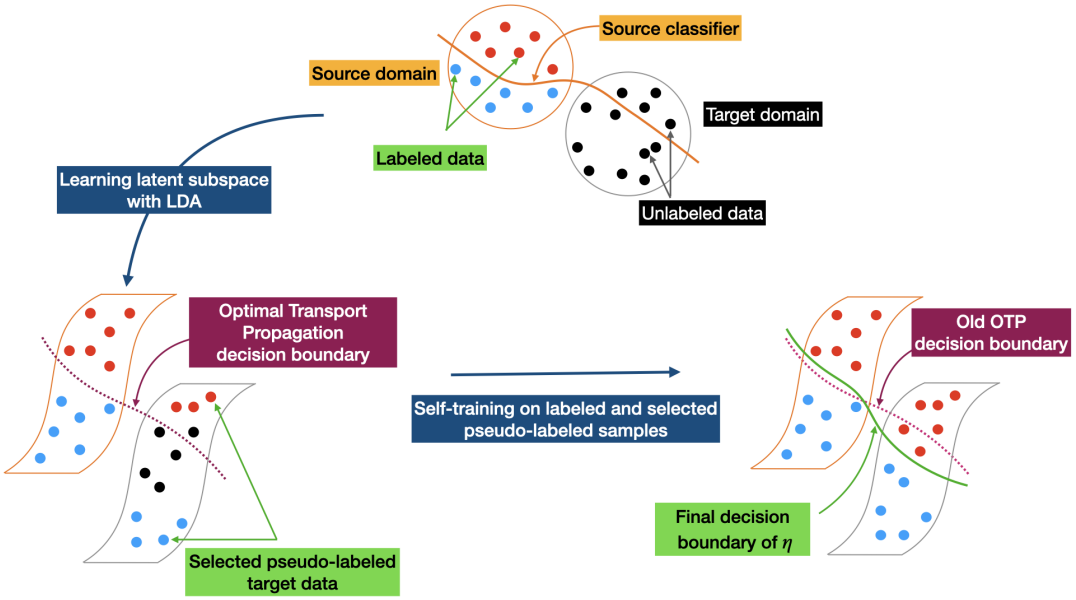

FIGURE 7.1: Overview of OTP-DA. We initiate an incremental approach where at each iteration, we learn a latent subspace using LDA. In the latent subspace, we perform selective pseudo-labeling with OTP. The selected pseudo-labeled target data are used in combination with labeled source data to learn a new decision boundary in a self-training fashion.

7.3 Theoretical Analysis

Self-training methods train a model to fit pseudo-labels, that is, predictions on unlabeled data made by a previously-learned model. The empirical phenomenon that self-training on pseudo-labels often improves over the pseudo-labeler F_{pl} despite no access to true labels, has been explained in (Wei et al., 2020) by Theorem 7.7. We first need the following definitions and assumptions.

Definition 7.1 (Transformation set) *Let $\mathbb{T}$ be the set of some transformations obtained via data augmentation, the transformation set of x is defined as:*

$$\mathbb{B}(x) = \{x' \mid \exists Tr \in \mathbb{T} \text{ such that } \| x' - Tr(x) \| \leq r\} \tag{7.7}$$

$\mathbb{B}(x)$ is the set of points with distance r from some data augmentation of x.

Definition 7.2 (Neighborhood) *The neighborhood of x denoted by $\mathcal{N}(x)$ is the set of points whose transformation sets overlap with that of x:*

$$\mathcal{N}(x) = \{x' : \mathbb{B}(x) \cap \mathbb{B}(x') \neq \emptyset\} \tag{7.8}$$

For $S \subset \mathcal{X}$, the neighborhood of S is defined as the union of neighborhoods of its elements: $\mathcal{N}(S) = \underset{x \in S}{\cup} \mathcal{N}(x)$.

Assumption 7.3 ((a,c)-expansion) *Let $\mathcal{P}$ be the distribution of unlabeled target data, and $\mathcal{P}_i$ for $i \leq k$ be the class-conditional distribution of $x \in \mathcal{X}$ conditioned on the class C_i. We say that the class-conditional distribution $\mathcal{P}_i$ satisfies (a, c)-expansion if for all $V \subset \mathcal{X}$ with $\mathcal{P}_i(V) \leq a$, the following holds:*

$$\mathcal{P}_i(\mathcal{N}(V)) \geq \min\{c\mathcal{P}_i(V), 1\} \tag{7.9}$$

If $\mathcal{P}_i$ satisfies (a, c)-expansion for all $i \leq k$, then we say $\mathcal{P}$ satisfies (a, c)-expansion.

This assumption states that a low-probability subset of the data must expand to a neighborhood with large probability relative to the subset.

Definition 7.4 (Population consistency loss) *We define the population consistency loss $R_{\mathbb{B}(x)}(F)$ as the fraction of examples where a classifier F is not robust to input transformations:*

$$R_{\mathbb{B}}(F) = \mathbb{E}_{\mathcal{P}}[\mathbb{1}(\exists x' \in \mathbb{B}(x) \text{ such that } F(x') \neq F(x))] \tag{7.10}$$

Assumption 7.5 (Separation) *We assume $\mathcal{P}$ is $\mathbb{B}$-separated with probability $1 - \delta$ by ground-truth classifier F^*, as follows: $R_{\mathbb{B}}(F^*) \leq \delta$.*

Assumption 7.6 *Define $\overline{a} = \max_{i \leq k}\{P_i(\mathcal{M}(F_{pl}))\}$ to be the maximum fraction of incorrectly pseudo-labeled examples in any class: $\mathcal{M}(F_{pl}) = \{x \mid F_{pl}(x) \neq F^*(x)\}$. We assume that $\overline{a} < \frac{1}{3}$ and P satisfies $(\overline{a}, \overline{c})$-expansion for $\overline{a} > 3$. .*

Theorem 7.7 *Define $c = \min\{\frac{1}{\overline{a}}, \overline{c}\}$. Suppose Assumptions 7.5 and 7.6 hold. Then for any minimizer $\tilde{F}$ of $\mathcal{L}(F) = \frac{c+1}{c-1} L_{0-1}(F, F_{pl}) + \frac{2c}{c-1} R_{\mathbb{B}}(F) - Err(F_{pl})$, which fits the classifier to the pseudo-labels while regularizing input consistency, we have:*

$$Err(\tilde{F}) \leq \frac{2}{c-1} Err(F_{pl}) + \frac{2c}{c-1}\delta. \tag{7.11}$$

Which explains the perhaps surprising fact that self-training with pseudo-labeling often improves over the pseudo-labeler F_{pl} even though no additional information about true labels is provided. This result is based on a simple and realistic expansion assumption that intuitively states that the data distribution has good continuity within each class.

This theoretical analysis [3] can be extended to our proposed approach that trains a classifier η in each iteration using the labeled source data and a small portion of pseudo-labeled target data provided by OTP.

7.4 Experiments

In this section, we provide empirical experimentation for the proposed algorithm and the compared algorithms.

7.4.1 Datasets

We adopt two datasets that are benchmarks in domain adaptation: ImageCLEF-DA and Office31.

ImageCLEF-DA dataset (Caputo et al., 2014) consists of four domains. We use three of them in our experiments: Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). There are 12 classes and 50 images for each class in each domain.

Office31 dataset (Saenko et al., 2010) composed of 4110 images. The dataset consists of three domains: Amazon, Webcam, and DSLR, 31 common classes from the three domains are used.

7.4.2 Experimental Protocol

We use ResNet50 features (d = 2048) for ImageCLEF-DA and Office31 datasets (He et al., 2016). Our proposed approach consists of two hyper-parameters, the dimensionality d_1 of LDA that we set equal to 128 and the sampling rate p that we set equal to 48 for ImageCLEF-DA and 62 for Office31 dataset. We use an SVM with a Gaussian kernel as a classifier (Benabdeslem and Bennani, 2006). The width parameter of the SVM was chosen as $\sigma = \frac{1}{2\mathbb{V}}$, where $\mathbb{V}$ is the variance of the source samples.

Following the standard protocol (Gong et al., 2012), the comparison is conducted using three deep learning models RTN (Long et al., 2016), MADA (Pei et al., 2018) and iCAN (Zhang et al., 2018), and with a manifold embedded distribution alignment technique based on deep features MEDA (Wang et al., 2018). We use the average accuracy as the evaluation metric in all our experiments.

7.4.3 Results

We use bold and underlined fonts to indicate the best and the second best results respectively.

[3] A theoretical study adapted to our algorithm is being prepared for an invited submission to International Journal of Neural Systems.

TABLE 7.1: Accuracy on ImageCELF-DA dataset (ResNet50 features).

Task	RTN	MADA	iCAN	MEDA	OT-DA
I → P	75.6	75.0	79.5	**79.7**	78.9
P → I	86.8	87.9	89.7	**92.5**	91.8
I → C	95.3	96.0	94.7	95.7	**97.8**
C → I	86.9	88.8	89.9	92.2	**92.6**
C → P	72.7	75.2	**78.5**	**78.5**	78.2
P → C	92.2	92.2	92.0	95.5	**95.8**
average	84.9	85.8	87.4	89.0	**89.4**

TABLE 7.2: Accuracy on Office31 dataset (ResNet50 features).

Task	RTN	MADA	iCAN	MEDA	OTP-DA
A → W	84.5	90.0	92.5	86.2	**93.3**
D → W	96.8	97.4	98.8	97.2	**99.0**
W → D	99.4	99.6	**100.0**	99.4	99.6
A → D	77.5	87.8	90.1	85.3	**90.7**
D → A	66.2	70.3	72.1	72.4	71.9
W → A	64.8	66.4	69.9	**74.0**	71.3
average	81.6	85.2	87.2	85.7	**87.6**

The classification accuracy of our proposed approach and other baseline methods are illustrated in Table 7.1 and Table 7.2, from which we can see that our proposed approach achieves the highest average accuracy over the two benchmark datasets.

Specifically, OTP-DA achieves an average accuracy of 89.4% on ImageCELF-DA dataset (Table 7.1), slightly better than MEDA which has an average accuracy of 89.0%. On the Office31 dataset (Table 7.2), OTP-DA outperforms all other baseline models with an average accuracy of 87.6% against 85.7% by MEDA and 87.2% by iCAN. Besides, OTP-DA achieves the best performance in three out of six tasks and the second-best results in two other tasks for both datasets.

In summary, the proposed approach is highly competitive compared to several state-of-the-art methods, and can outperform both deep learning models and traditional feature transformation approaches on many tasks of the two domain adaptation problems. These results are mainly attributed to the capacity of OTP to capture much more information than the other methods of pseudo-labeling thanks to the enhanced affinity matrix constructed by optimal transport and to its intrinsic property of selectivity which make it a good candidate for pseudo-labeling target data.

7.5 Software

We make our code and the used datasets publicly available at:

https://github.com/MouradElHamri/OTP-DA

7.6 Conclusion and future perspectives

With the possibility of exploring hidden structures in the target domain using Optimal Transport Propagation, we proposed to use them to incrementally create augmented source structures, composed of labeled source data and selected pseudo-labeled target data. The augmented source structures allow learning a sequence of discriminative and domain-invariant latent subspaces, using Linear Discriminate Analysis, within which it becomes convenient to progressively label the data of the target domain in a self-training manner.

The work of this chapter can be extended in different directions:

- From a theoretical standpoint, we are currently writing an extended version of this work, in which we develop a theoretical analysis specific to our algorithm based on the notion of weak learner (Freund et al., 1996). Indeed, there are few theoretical studies of domain adaptation methods based on self-labeling, and the published theoretical studies suffer from several limitations, for example, (Habrard et al., 2013) deals with a limited setting where a random selection of pseudo-labeled target examples is performed at each iteration. The current work attempts to overcome this non-optimal setup and proposes a theoretical analysis in which pseudo-labels are selected according to a deterministic procedure that reflects the functioning of OTP.
- From an algorithmic standpoint, it would be interesting to go beyond the domain adaptation special case we considered. A direct extension would be to include some target labels in a semi-supervised setting, we believe that this latter will be more advantageous for the proposed approach since the unlabeled target data will be drawn from the same distribution as the small amount of the available labeled ones, which will lead to learning latent subspaces with a higher discriminatory power.

CHAPTER 8

CONCLUSION AND PERSPECTIVE FOR FURTHER WORKS

> The worst thing one can tell you is that you're intelligent.
>
> My mom to her eldest son, 2002

Throughout this dissertation, we tackled the challenging problem of domain adaptation, and we provided contributions to it using different approaches that may be unified under the banner of structural optimal transport.

Motivated by the lack of methods that leverages structures in both domains, our first contribution considers a hierarchical formulation of optimal transport that aligns the source structures with the target ones. Structures in the source domain are formed by grouping samples according to their class labels, while structures in the target domain are learned using Wasserstein-Spectral clustering, an algorithm derived from the equivalence we proved between the problem of learning probability measures through Wasserstein barycenter and spectral clustering. Incorporating the structures of the two domains into the hierarchical formulation of optimal transport yielded good empirical results in domain adaptation.

The need to give a theoretical cover to the first contribution, led us in the second, to prove novel generalization bounds on the target risk for three scenarios, unsupervised, semi-supervised, and multi-source domain adaptation, where the divergence between the source and target domains is measured by the Hierarchical Wasserstein distance. Our generalization bounds justify the use of hierarchical optimal transport for domain adaptation and indicate under mild assumptions, which structures have to be aligned to lead to a good adaptation. These generalization bounds explicitly reflect, unlike the other state-of-the-art bounds, the algorithmic solution that was used to lead to a successful adaptation.

Marked by the necessity for an efficient way to learn hidden structures in the target domain in lieu of clustering, and by the many drawbacks of traditional semi-supervised approaches, we developed in the third contribution a label propagation technique based on optimal transport. The proposed approach captures the geometry of the input space and the relationships between labeled and unlabeled samples in a global level. This approach incrementally performs a label propagation process controlled by a score that watches over the certainty of predictions. This approach has shown good empirical performance compared to the state-of-the-art methods.

Finally, having the possibility of exploring hidden structures in the target domain using the developed label propagation technique, we proposed to use them to incrementally create augmented source structures, composed of labeled source data and selected pseudo-labeled target data. The augmented source structures allow learning a sequence of discriminative and domain-invariant latent subspaces, within which it becomes convenient to gradually label the data of the target domain.

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 the alignment process is done in several steps, it is first necessary to learn the hidden structures in the target domain and then to find the correspondences with the source structures. We would like to avoid this, by proposing to jointly learn the target structures and the optimal transport plan that aligns them with the source classes. In addition, we want to extend the proposed approach to a multi-source domain adaptation setting.

Second, our theoretical study of domain adaptation through the hierarchical formulation of optimal transport encourages us to attempt to derive other generalization bounds that take into account the quality of clustering in the target domain, by reflecting explicitly the excess clustering risk. Recently, (Li and Liu, 2021) proposed a unified clustering framework that encompasses k-means, kernel k-means, soft k-means, neural network clustering, and spectral clustering, and investigated its excessive risk bounds, obtaining state-of-the-art upper bounds under mild assumptions. The equivalence between Wasserstein-spectral clustering and spectral clustering will allow us to use this framework to derive excess clustering risk bounds for the former and then try to find a means to merge them into the domain adaptation generalization bounds.

The third contribution leads us to the question of whether it is possible to establish generalization bounds for semi-supervised learning using optimal transport. For the last contribution, we are currently writing an extended version, in which we develop a theoretical analysis specific to our algorithm based on the notion of weak learner (Freund et al., 1996). Indeed, there are few theoretical studies of domain adaptation methods based on self-labeling, and the published theoretical studies suffer from several limitations, for example, (Habrard et al., 2013) deals with a limited setting where a random selection of pseudo-labeled target examples is performed at each iteration. The current work attempts to overcome this non-optimal setup and proposes a theoretical analysis in which pseudo-labels are selected according to a deterministic procedure that reflects the functioning of OTP.

From a more high-level perspective, many questions remain open as the possibility to extend our methods and theoretical guarantees to data living in incomparable spaces, using in particular approaches developed by (Vayer et al., 2019, 2020). We also aim to extend our approaches to the multi-source setting by leveraging our other work on collaborative and federated learning (Ben Bouazza et al., 2022). Regarding applications, we are at the beginning of a collaboration with a team of surgeons from Avicenne Hospital to apply our algorithms in various medical problematics, for example, we are currently considering the possibility to deploy a model trained using postmortem imaging during intra-operative surgery.

List of Publications

1. **M. El Hamri**, Y. Bennani, and I. Falih. Hierarchical optimal transport for unsupervised domain adaptation. Machine Learning, pages 1–24. [2022b]

2. **M. El Hamri**, Y. Bennani, and I. Falih. When domain adaptation meets semi-supervised learning through optimal transport. In IFIP International Conference on Artificial Intelligence Applications and Innovations, pages 58–69. Springer. [2022a]

3. **M. El Hamri**, Y. Bennani, and I. Falih. Incremental unsupervised domain adaptation through optimal transport. In 2022 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE. [2022c]

4. **M. El Hamri**, Y. Bennani, and I. Falih. Label propagation through optimal transport. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE. [2021c]

5. **M. El Hamri**, Y. Bennani, and I. Falih. Inductive semi-supervised learning through optimal transport. In International Conference on Neural Information Processing, pages 668–675. Springer. [2021b]

6. **M. El Hamri**, Y. Bennani, and I. Falih. Apprentissage semi-supervisé transductif basé sur le transport optimal. In 9e Conférence Internationale Francophone sur la Science des Données. [2021a]

7. F.-E. Ben Bouazza, Y. Bennani, and **M. El Hamri**. An optimal transport framework for collaborative multi-view clustering. In Recent Advancements in Multi-View Data Analytics, pages 131–157. Springer [2022]

8. F.-E. Ben Bouazza, Y. Bennani, **M. El Hamri**, G. Cabanes, B. Matei, and A. Touzani. Multi-view clustering through optimal transport. Aust. J. Intell. Inf. Process. Syst., 15(3):1–9. [2019]

9. **M. El Hamri**, Y. Bennani, and I. Falih. Theoretical guarantees for domain adaptation with hierarchical optimal transport. Under review in Machine Learning. [2022d]

10. **M. El Hamri**, Y. Bennani, and I. Falih. Domain adaptation via incremental confidence samples with optimal transport. Invited submission to International Journal of Neural Systems.

11. **M. El Hamri**, Y. Bennani, and I. Falih. Apprentissage semi-supervisé transductif basé sur le transport optimal. To appear in Revue des Nouvelles Technologies de l'Information (Extended version of [2021a]).

APPENDIX A

SOME PREREQUISITES

A.1 Probabilities

Definition A.1 (σ-algebra) *Let Ω be a set. A subset $\mathcal{F}$ of the power set $P(\Omega)$ is called a σ-algebra if it satisfies the following conditions:*

1. $\Omega \in \mathcal{F}$.
2. $\forall A \in \mathcal{F}$, $\Omega \setminus A \in \mathcal{F}$, *(closed under complementation).*
3. *For any sequence $(A_i)_{i\in\mathbb{N}}$ such that $A_i \in \mathcal{F}$, we have $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$ (closed under countable unions).*

In this case, $(\Omega, \mathcal{F})$ is called a measurable space.

Definition A.2 (Measurable function) *With the previous notations, let $(E, \mathcal{K})$ be a measurable space. A map $f : \Omega \to E$ is said to be measurable if: $\forall B \in \mathcal{K}$, $f^{-1}(B) \in \mathcal{F}$.*

Definition A.3 (Probability measure) *With the previous notations, let $\mathbb{P} : \mathcal{F} \to \mathbb{R}_+$. $\mathbb{P}$ is a probability if:*

1. $\mathbb{P}(\Omega) = 1$
2. *For any sequence $(A_i)_{i\in\mathbb{N}}$ such that $A_i \in \mathcal{F}$ and $A_i \cap A_j =$ if $i \neq j$, we have: $\mathbb{P}\big(\bigcup_{i\in\mathbb{N}} A_i\big) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$*

In this case, $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space.

A.2 Topology

Definition A.4 (Metric) *For a set $\mathcal{X}$, an application $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a metric if it verifies the following properties:*

1. $\forall x, x' \in \mathcal{X}$, $d(x, x') = 0 \iff x = x'$ *(separation).*
2. $\forall x, x' \in \mathcal{X}$, $d(x, x') = d(x', x)$ *(symmetry).*
3. $\forall x, x', x'' \in \mathcal{X}$, $d(x, x') \leq d(x, x'') + d(x'', x')$ *(triangle inequality).*

In this case, $(\mathcal{X}, d)$ is called a metric space.

Definition A.5 (Lipschitzness) *Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be two metric spaces. An application $f : \mathcal{X} \to \mathcal{Y}$ verifies the Lipschitz property if there is some $K > 0$ such that for all $x, x' \in \mathcal{X}$, we have:*

$$d_\mathcal{Y}(f(x), f(x')) \leq K.d_\mathcal{X}(x, x') \tag{A.1}$$

In this case we also say that f is K-Lipschitz continuous.

Definition A.6 (Norm) *For a real vector space E, an application $||.|| : E \to \mathbb{R}_+$ is a norm if it verifies the following properties:*

1. $\forall x \in E, ||x|| = 0 \iff x = 0$ *(Positive definiteness).*
2. $\forall \lambda \in \mathbb{R}, \forall x \in E,\ ||\lambda x|| = \lambda ||x||$ *(Absolute homogeneity).*
3. $\forall x, x' \in E, ||x + x'|| \leq ||x|| + ||x'||$ *(triangle inequality).*

In this case, $(E, ||.||)$ is called a normed vector space.

Proposition A.7 *Any normed space $(E, ||.||)$ is a metric space for $d : (x, x') \mapsto ||x - x'||$.*

Definition A.8 (Complete space) *We say that a metric space $(\mathcal{X}, d)$ is complete if every Cauchy sequence in $\mathcal{X}$ has a limit in $\mathcal{X}$, i.e., every Cauchy sequence is convergent.*

Definition A.9 (Separable space) *Let $\mathcal{X}$ be a metric space. A set B is dense in $\mathcal{X}$ if $cl(B) = \mathcal{X}$. We say that a metric space is separable if it has a countable dense subset.*

Definition A.10 (Polish space) *A topological space $\mathcal{X}$ is*

1. *completely metrizable if there is a metric d defining the topology of $\mathcal{X}$ such that $(\mathcal{X}, d)$ is complete*
2. *Polish if it is separable and completely metrizable*

Definition A.11 (Lower semi-continuity) *Let $(\mathcal{X}, d)$ be a polish metric space, $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous at $x \in \mathcal{X}$ if and only if, for every sequence $(x_n)_{n\in\mathbb{N}}$ converged to x, we have*

$$f(x) \leq \varliminf_{n\to+\infty} f(x_n)$$

f is lower semi-continuous if it is lower semi-continuous at each point in $\mathcal{X}$.

A.3 Functional analysis

Definition A.12 (Hilbert space) *A Hilbert space is a complete inner product space with respect to the norm defined by the inner product.*

Definition A.13 (RKHS) *A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if the evaluation functional $ev_x : f \mapsto f(x)$ is continuous $\forall x \in \mathcal{X}$.*

Definition A.14 (Reproducing kernel) *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies:*

1. $\forall x \in \mathcal{X},\ k(., x) \in \mathcal{H}$
2. $\forall x \in X, \forall f \in \mathcal{H},\ f(x) = \langle f, k(., x)\rangle_{\mathcal{H}}$ *(the reproducing property).*

Theorem A.15 *$\mathcal{H}$ is a reproducing kernel Hilbert space (i.e., its evaluation functionals are continuous), if and only if $\mathcal{H}$ has a reproducing kernel.*

References

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

A. Alexandrescu and K. Kirchhoff. Data-driven graph construction for semi-supervised graph-based learning in nlp. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 204–211, 2007.

R. Aljundi, R. Emonet, D. Muselet, and M. Sebban. Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 56–63, 2015.

J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

J. M. Altschuler and E. Boix-Adsera. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *J. Mach. Learn. Res.*, 22:44–1, 2021.

D. Alvarez-Melis, T. Jaakkola, and S. Jegelka. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1771–1780. PMLR, 2018.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

P. Appell. Mémoire sur les déblais et les remblais des systemes continus ou discontinus. *Mémoires présentes par divers Savants à l'Académie des Sciences de l'Institut de France*, 29:1–208, 1887.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

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.

M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Domain adaptation on the statistical manifold. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2488, 2014.

E. Barba, L. Procopio, N. Campolungo, T. Pasini, and R. Navigli. Mulan: Multilingual label propagation for word sense disambiguation. In *Proc. of IJCAI*, pages 3837–3844, 2020.

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.

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. F. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. 2018.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library. Princeton University Press, 1961. ISBN 9780691079011.

F. E. Ben Bouazza, Y. Bennani, M. El Hamri, G. Cabanes, B. Matei, and A. Touzani. Multi-view clustering through optimal transport. *Aust. J. Intell. Inf. Process. Syst.*, 15 (3):1–9, 2019.

F.-E. Ben Bouazza, Y. Bennani, and M. El Hamri. An optimal transport framework for collaborative multi-view clustering. In *Recent Advancements in Multi-View Data Analytics*, pages 131–157. Springer, 2022.

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, and F. Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

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

S. Ben-David, D. Loker, N. Srebro, and K. Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. *ICML*, 2012.

K. Benabdeslem and Y. Bennani. Dendrogram-based svm for multi-class classification. *Journal of Computing and Information Technology*, 14(4):283–289, 2006.

J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Y. Bengio, O. Delalleau, and N. Le Roux. 11 label propagation and quadratic criterion. 2006.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *45th annual meeting of the ACL*, 2007.

P. Boldi, M. Rosa, M. Santini, and S. Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the 20th international conference on World wide web*, pages 587–596, 2011.

F. Bolley and C. Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, number 3, pages 331–352, 2005.

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, M. Van De Panne, S. Paris, and W. Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

N. Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Y. Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

F. Breve. Interactive image segmentation using label propagation through complex networks. *Expert Systems With Applications*, 123:18–33, 2019.

L. A. Caffarelli. Some regularity properties of solutions of monge ampere equation. Technical report, 1991.

R. J. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

R. J. Campello, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):e1343, 2020.

G. Canas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

B. Caputo, H. Müller, J. Martinez-Gomez, M. Villegas, B. Acar, N. Patricia, N. Marvasti, S. Üsküdarlı, R. Paredes, M. Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation*

Forum for European Languages, pages 192–211. Springer, 2014.

O. Catoni. Pac-bayesian supervised classification: The thermodynamics of statistical learning. institute of mathematical statistics lecture notes—monograph series 56. *IMS, Beachwood, OH. MR2483528*, 5544465, 2007.

O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019.

M. Chen, W. EDU, and Z. E. Xu. Marginalized denoising autoencoders for domain adaptation. 2012.

Q. Chen, Y. Liu, Z. Wang, I. Wassell, and K. Chetty. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7976–7985, 2018.

L. Chizat. *Unbalanced optimal transport: Models, numerical methods, applications*. PhD thesis, Université Paris sciences et lettres, 2017.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

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. *Advances in Neural Information Processing Systems*, 30, 2017.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

B. B. 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.

G. B. Dantzig. Programming of interdependent activities: Ii mathematical model. *Econometrica, Journal of the Econometric Society*, pages 200–211, 1949.

G. B. Dantzig. Application of the simplex method to a transportation problem. *Activity analysis and production and allocation*, 1951.

A. de Mathelin, F. Deheeger, M. Mougeot, and N. Vayatis. Fast and accurate importance weighting for correcting sample bias. 2022.

O. Delalleau, Y. Bengio, and N. Le Roux. Efficient non-parametric function induction in semi-supervised learning. In *AISTATS*, volume 27, page 100. Citeseer, 2005.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

I. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth international conference on Knowledge discovery and data mining*, 2004.

S. Dhouib, I. Redko, and C. Lartizien. Margin-aware adversarial domain adaptation with optimal transport. In *International Conference on Machine Learning*, pages 2514–2524. PMLR, 2020.

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

C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the SIAM international conference on data mining*, 2005.

B. E. Dom. An information-theoretic external cluster-validity measure. *arXiv preprint arXiv:1301.0565*, 2012.

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655. PMLR, 2014.

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

M. El Hamri, Y. Bennani, and I. Falih. Apprentissage semi-supervisé transductif basé sur le transport optimal. In *9 e Conférence Internationale Francophone sur la Science des Données*, 2021a.

M. El Hamri, Y. Bennani, and I. Falih. Inductive semi-supervised learning through optimal transport. In *International Conference on Neural Information Processing*, pages 668–675. Springer, 2021b.

M. El Hamri, Y. Bennani, and I. Falih. Label propagation through optimal transport.

In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021c.

M. El Hamri, Y. Bennani, and I. Falih. When domain adaptation meets semi-supervised learning through optimal transport. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 58–69. Springer, 2022a.

M. El Hamri, Y. Bennani, and I. Falih. Hierarchical optimal transport for unsupervised domain adaptation. *Machine Learning*, pages 1–24, 2022b.

M. El Hamri, Y. Bennani, and I. Falih. Incremental unsupervised domain adaptation through optimal transport. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022c.

M. El Hamri, Y. Bennani, and I. Falih. Theoretical guarantees for domain adaptation with hierarchical optimal transport. *arXiv preprint arXiv:2210.13331*, 2022d.

M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE ICCV*, pages 2960–2967, 2013.

B. Fernando, T. Tommasi, and T. Tuytelaars. Joint cross-domain classification and subspace learning for unsupervised adaptation. *Pattern Recognition Letters*, 65:60–66, 2015.

S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

A. Figalli and F. Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. 2021.

A. Figalli, F. Maggi, and A. Pratelli. A mass transportation approach to quantitative isoperimetric inequalities. *Inventiones mathematicae*, 182(1):167–211, 2010.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, et al. Pot: Python optimal transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021.

L. R. Ford and D. R. Fulkerson. Flows in networks. In *Flows in Networks*. Princeton university press, 1962.

Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015.

A.-J. Gallego, J. Calvo-Zaragoza, and R. B. Fisher. Incremental unsupervised domain-adversarial training of neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4864–4878, 2020.

Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17 (1):2096–2030, 2016.

M. Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

A. Genevay. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE), 2019.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.

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. PMLR, 2013.

P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J.-F. Roy. Risk bounds for the

majority vote: From a pac-bayesian analysis to a learning algorithm. *arXiv preprint arXiv:1503.08329*, 2015.

M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016a.

M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pages 597–613. Springer, 2016b.

X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

A. V. Goldberg and R. E. Tarjan. Finding minimum-cost circulations by canceling negative cycles. *Journal of the ACM (JACM)*, 36(4):873–886, 1989.

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.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS'14*, 2014.

R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pages 999–1006. IEEE, 2011.

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.

M. Gu, H. Zha, C. Ding, X. He, H. Simon, and J. Xia. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering. 2001.

A. Habrard, J.-P. Peyrache, and M. Sebban. Iterative self-labeling domain adaptation for linear structured image classification. *International Journal on Artificial Intelligence Tools*, 22(05):1360005, 2013.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

F. L. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of mathematics and physics*, 20(1-4):224–230, 1941.

N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via wasserstein means. In *International Conference on Machine Learning*, pages 1501–1509, 2017.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5070–5079, 2019.

Z. Izzo, S. Silwal, and S. Zhou. Dimensionality reduction for wasserstein barycenter. *Advances in Neural Information Processing Systems*, 34, 2021.

E. Jokar and M. Mosleh. Community detection in social networks based on improved label propagation algorithm and balanced link density. *Physics Letters A*, 383 (8):718–727, 2019.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

H. Karcher. Riemannian center of mass and so called karcher mean. *arXiv preprint arXiv:1407.2087*, 2014.

D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pages 180–191. Toronto, Canada, 2004.

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

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

S. B. Kotsiantis, I. Zaharakis, P. Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, pages 3–24, 2007.

W. M. Kouw and M. Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785, 2019.

A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe. On the complexity of approximating wasserstein barycenters. In *International conference on machine learning*, pages 3530–3540. PMLR, 2019.

M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

J. Lee, M. Dabagia, E. Dyer, and C. Rozell. Hierarchical optimal transport for multimodal distribution alignment. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

M. Li, Y.-M. Zhai, Y.-W. Luo, P.-F. Ge, and C.-X. Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13944, 2020.

S. Li and Y. Liu. Sharper generalization bounds for clustering. In *International Conference on Machine Learning*, pages 6392–6402. PMLR, 2021.

X. Liu, H.-M. Cheng, and Z.-Y. Zhang. Evaluation of community detection methods. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

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. PMLR, 2015.

M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016.

M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967.

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. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

D. A. McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

R. J. McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

F. Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

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. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

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

A. Musdholifah, S. Z. M. Hashim, and S. Zaiton. Cluster analysis on high-dimensional data: A comparison of density-based clustering algorithms. *Australian Journal of Basic and Applied Sciences*, 7(2):380–389, 2013.

K. Nadjahi, A. Durmus, U. Simsekli, and R. Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14:849–856, 2001.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 2009.

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, 2010.

K. R. Parthasarathy. *Probability measures on metric spaces*, volume 352. American Mathematical Soc., 2005.

Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2):199–205, 1982.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

A. Rakotomamonjy, R. Flamary, G. Gasso, M. E. Alaya, M. Berar, and N. Courty. Optimal transport for conditional domain matching and label shift. *Machine Learning*, 111(5):1651–1670, 2022.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

I. Redko. *Nonnegative matrix factorization for transfer learning*. PhD thesis, Sorbonne Paris Cité, 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, N. Courty, R. Flamary, and D. Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd AISTATS*, pages 849–858. PMLR, 2019a.

I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. *Advances in domain adaptation theory*. Elsevier, 2019b.

S. Reich. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

S. Saitoh. *Integral transforms, reproducing kernels and their applications*, volume 369. CRC Press, 1997.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55 (58-63):94, 2015.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

B. Schmitzer and C. Schnörr. A hierarchical approach to optimal transport. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 452–464. Springer, 2013.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

E. Schrödinger. *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u . . . , 1931.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291, 2013.

T. Séjourné, F.-X. Vialard, and G. Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, 1997.

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.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

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.

A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pages 306–314. PMLR, 2014.

X. Y. Stella and J. Shi. Multiclass spectral clustering. In *Computer Vision, IEEE International Conference on*, volume 2, pages 313–313. IEEE Computer Society, 2003.

K.-T. Sturm. On the geometry of metric measure spaces. *Acta mathematica*, 196(1): 65–131, 2006.

K.-T. Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

A. Subramanya and P. P. Talukdar. Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(4):1–125, 2014.

M. Sugiyama, S. Nakajima, H. Kashima, P. Von Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2007.

B. Sun and K. Saenko. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, volume 4, pages 24–1, 2015.

B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

F. Taherkhani, A. Dabouei, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Transporting labels via hierarchical optimal transport for semi-supervised learning. In *ECCV*, pages 509–526, 2020.

A. Tamura, T. Watanabe, and E. Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36, 2012.

V. Titouan, R. Flamary, N. Courty, R. Tavenard, and L. Chapel. Sliced gromov-wasserstein. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Tolstoi. Methods of finding the minimal total kilometrage in cargo transportation planning in space. *TransPress of the National Commissariat of Transportation*, 1:23–55, 1930.

S. Tsironis, M. Sozio, M. Vazirgiannis, and L. Poltechnique. Accurate spectral clustering for community detection in mapreduce. In *Advances in Neural Information Processing Systems (NIPS) Workshops*, page 8. Citeseer, 2013.

J. Turkey. Exploratory data analysis, vol. 2, 1977.

L. Q. Uddin. Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. *Nature Reviews Neuroscience*, 22(3):167–179, 2021.

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. *Communications of the ACM*, 27(11):1134–1142, 1984.

J. E. Van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

V. Vapnik. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

V. Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

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. In *Measures of complexity*, pages 11–30. Springer, 2015.

V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.

T. Vayer. A contribution to optimal transport on incomparable spaces. *arXiv preprint arXiv:2011.04447*, 2020.

T. Vayer, R. Flamary, R. Tavenard, L. Chapel, and N. Courty. Sliced gromov-wasserstein. *arXiv preprint arXiv:1905.10124*, 2019.

T. Vayer, I. Redko, R. Flamary, and N. Courty. Co-optimal transport. *Advances in Neural Information Processing Systems*, 33:17559–17570, 2020.

H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):55–67, 2007.

F. Wang, X. Wang, and T. Li. Efficient label propagation for interactive image segmentation. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, pages 136–141. IEEE, 2007.

J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 402–410, 2018.

Q. Wang and T. Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6243–6250, 2020.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

C. Wei, K. Shen, Y. Chen, and T. Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

J. Xie and B. K. Szymanski. Labelrank: A stabilized label propagation algorithm for community detection in networks. In *2013 IEEE 2nd Network Science Workshop (NSW)*, pages 138–143. IEEE, 2013.

S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pages 5423–5432. PMLR, 2018.

H. Xu and S. Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

R. Xu, P. Liu, L. Wang, C. Chen, and J. Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4403, 2020.

D. Yan, L. Huang, and M. I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916, 2009.

Q. Yang, Y. Zhang, W. Dai, and S. J. Pan. *Transfer learning*. Cambridge University Press, 2020.

M. Yurochkin, S. Claici, E. Chien, F. Mirzazadeh, and J. M. Solomon. Hierarchical

optimal transport for document representation. In *Advances in Neural Information Processing Systems*, 2019.

H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon. Spectral relaxation for k-means clustering. In *Advances in neural information processing systems*, pages 1057–1064, 2001.

W. Zhang, W. Ouyang, W. Li, and D. Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3801–3809, 2018.

X. Zhang, F. X. Yu, S.-F. Chang, and S. Wang. Deep transfer network: Unsupervised domain adaptation. *arXiv preprint arXiv:1503.00591*, 2015.

X.-K. Zhang, J. Ren, C. Song, J. Jia, and Q. Zhang. Label propagation algorithm for community detection based on node importance and label influence. *Physics Letters A*, 381(33):2691–2698, 2017.

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. PMLR, 2019.

D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16:321–328, 2003.

X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

X. J. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.

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

Titre: Transport optimal structurel pour l'adaptation de domaine avec garanties théoriques

Mots clés: Adaptation de domaine, Transport optimal, Bornes de généralisation

Résumé: La théorie du transport optimal permet non seulement de définir une distance entre les mesures de probabilité, mais offre également un moyen géométrique de transporter un ensemble de points vers un autre selon le principe du moindre effort. Ce double aspect a laissé la porte grande ouverte pour les applications en adaptation de domaine, une branche de l'apprentissage statistique qui tient compte du changement de distributions entre les données d'apprentissage et les données de test, respectivement appelées domaines source et cible. Toutefois, il existe souvent dans les deux domaines un biais structurel sur la représentation des données ou des structures latentes qui ne sont pas prises en compte par la formulation classique du transport optimal, et l'incapacité à incorporer pleinement ces structures peut entraver le succès de l'adaptation de domaine. Cette thèse présente plusieurs approches pour incorporer les informations structurelles au sein du transport optimal. La première contribution s'appuie sur une formulation hiérarchique du transport optimal pour aligner les structures sources et cibles. Les structures sources sont formées instinctivement en regroupant les données en classes selon leurs étiquettes, tandis que l'apprentissage des structures cachées dans le domaine cible est réduit au problème d'apprentissage de mesures de probabilité via le barycentre de Wasserstein, dont nous prouvons l'équivalence avec le clustering spectral. Notre deuxième contribution est une analyse théorique de l'adaptation de domaine à travers le transport optimal hiérarchique, où nous fournissons des bornes de généralisation pour trois scénarios, à savoir, l'adaptation de domaine non supervisé, semi-supervisé et multi-sources. Ces bornes de généralisation sont basées sur une nouvelle mesure de divergence que nous appelons la distance de Wasserstein Hiérarchique, qui indique, sous des hypothèses modérées, quelles structures doivent être alignées pour mener à une adaptation réussie. Dans notre troisième contribution, nous élargissons le cadre d'apprentissage des structures cibles en dehors du clustering, en développant une approche de propagation de labels basée sur le transport optimal. L'intérêt du transport optimal dans ce contexte est de capturer la géométrie de l'espace d'entrée dans son intégralité. Cette approche effectue une propagation incrémentale de labels, contrôlée par un score qui surveille la certitude des prédictions. Enfin, en s'appuyant sur ce nouvel algorithme de propagation de labels, nous présentons la dernière contribution, qui permets de créer de manière progressive des structures sources augmentées, permettant l'apprentissage d'une suite de sous-espaces latents domaine-invariants et discriminants, au sein desquels il devient facile d'étiqueter graduellement les données du domaine cible.

Title: Structural Optimal Transport for Domain Adaptation with Theoretical Guarantees

Keywords: Domain Adaptation, Optimal Transport, Generalization bounds

Abstract: Optimal transport theory not only defines a distance between probability measures but also provides a geometric way to transport a set of points to another according to the principle of least effort. This dual aspect has left the door wide open for applications in domain adaptation, a subfield of statistical learning theory that takes into account the change in distributions between training and test data, respectively called source and target domains. However, there is often a structural bias on the data representation or latent structures in both domains that are not captured by the classical optimal transport formulation, and the inability to fully incorporate these structures can hinder the success of domain adaptation. This thesis presents several approaches to incorporating structural information into the optimal transport problem. The first contribution relies on a hierarchical formulation of optimal transport to align source and target structures. The source structures are formed instinctively by grouping data into classes according to their labels while learning hidden structures in the target domain is reduced to the problem of learning probability measures through Wasserstein barycenter, which we prove to be equivalent to spectral clustering. Our second contribution is a new theoretical framework of domain adaptation through hierarchical optimal transport, where we provide generalization bounds for three scenarios, namely, unsupervised, semi-supervised, and multi-source domain adaptation. These generalization bounds are based on a new divergence measure that we call Hierarchical Wasserstein distance, indicating, under mild assumptions, which structures need to be aligned to lead to successful adaptation. In our third contribution, we extend the framework of learning target structures outside of clustering, by developing a label propagation approach based on optimal transport. The appeal of optimal transport in this context is to capture the geometry of the input space in its entirety. This approach performs incremental label propagation, controlled by a score that watches over the certainty of predictions. Finally, based on this new label propagation algorithm, we present the last contribution, which allows the progressive creation of augmented source structures, allowing to learn a sequence of latent domain-invariant and discriminative subspaces, within which it becomes easy to gradually label the target data.

Université Sorbonne Paris Nord

UNIVERSITÉ SORBONNE PARIS NORD