

On the Use of the Kantorovich-Rubinstein Distance for Dimensionality Reduction

Gaël Giordano

Thesis submitted to the Faculty of Science in partial fulfillment of the requirements
for the degree of
Doctorate in Philosophy Mathematics and Statistics¹

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

© Gaël Giordano, Ottawa, Canada, 2023

¹The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

The goal of this thesis is to study the use of the Kantorovich-Rubinstein distance as to build a descriptor of sample complexity in classification problems. The idea is to use the fact that the Kantorovich-Rubinstein distance is a metric in the space of measures that also takes into account the geometry and topology of the underlying metric space. We associate to each class of points a measure and thus study the geometrical information that we can obtain from the Kantorovich-Rubinstein distance between those measures. We show that a large Kantorovich-Rubinstein distance between those measures allows to conclude that there exists a 1-Lipschitz classifier that classifies well the classes of points. We also discuss the limitation of the Kantorovich-Rubinstein distance as a descriptor.

Résumé

Dans cette thèse, on étudie l'utilisation de la distance de Kantorovich-Rubinstein afin de construire des descripteurs de complexité. Ces descripteurs aident à juger de la difficulté à séparer un échantillon lors d'un problème de classification. On utilise le fait que la distance de Kantorovich-Rubinstein est une distance dans l'espace des mesures qui prend en compte la géométrie et la topologie de l'espace métrique sous-jacent. On associe à chacune des deux classes de points une mesure et on étudie l'information géométrique que l'on obtient à partir de la distance de Kantorovich-Rubinstein entre ces mesures. On montre qu'une grande distance de Kantorovich-Rubinstein permet de conclure qu'il existe un classificateur 1-lipschitzien qui sépare avec une grande précision les deux classes de points. On termine la thèse par une discussion sur les limites de l'utilisation de la distance de Kantorovich-Rubinstein comme descripteur de complexité.

Acknowledgement

Merci à tous ceux qui m'ont soutenu durant les très nombreuses années nécessaires à la rédaction de cette thèse. Un merci tout spécial à ma famille qui a trouvé la patience pour me soutenir du premier au dernier jour. Je tiens aussi à souligner l'aide apportée par mes deux superviseurs, Dr Pestov et Dr Wells.

Contents

List of Tables	ix
Preface	x
1 Kantorovich minimisation problem and Duality Theorems	1
1.1 Couplings	2
1.2 Kantorovich Minimisation Problem	6
1.2.1 Kantorovich Duality Theorem	7
1.2.2 Existence of Optimal Solutions	11
1.3 Kantorovich-Rubinstein Duality Theorem	18
1.3.1 Existence of Optimal measures and functions	25
2 The Kantorovich-Rubinstein Distance	33
2.1 The Wasserstein Distance	34
2.1.1 Convergence in Wasserstein sense	35
2.1.2 Topological properties of the Wasserstein space	36
2.2 The Kantorovich-Rubinstein distance	37
2.2.1 The Kantorovich-Rubinstein distance between product mea- sures	41
2.2.2 The Kantorovich-Rubinstein distance for atomic measures	44
2.3 The Kantorovich-Rubinstein distance for a bounded distance on \mathcal{X}	44

2.4	The Kantorovich-Rubinstein distance for the discrete distance on \mathcal{X}	48
2.4.1	The Kantorovich-Rubinstein distance of atomic measures for the discrete distance on \mathcal{X}	51
2.4.2	The Kantorovich-Rubinstein distance of atomic measures on the product space \mathcal{X}^n where each \mathcal{X} is equipped with the discrete distance	53
2.5	The Kantorovich-Rubinstein distance on the line equipped with the Euclidean metric	55
2.5.1	The Kantorovich-Rubinstein distance for atomic measures on the line	61
2.6	The Kantorovich-Rubinstein distance on the circle \mathbb{S}_1	62
2.6.1	The Kantorovich-Rubinstein distance on the circle \mathbb{S}^1 for atomic measures	68
3	The Kantorovich-Rubinstein Distance and Statistical Trend Tests	70
3.1	The Cochran-Armitage Trend Test and the Pearson's Chi-square Test of Homogeneity	71
3.1.1	The Pearson's Chi-square Test of Homogeneity	71
3.1.2	The Cochran-Armitage Trend Test	73
3.1.3	Relationship between the Pearson and Cochran-Armitage statistics	77
3.2	The function T	78
3.3	Generalization of the Pearson and Cochran-Armitage Statistics	91
3.3.1	Standard Statistics: a Special Case of the Generalized Statistics	92
3.4	Relationship between the Kantorovich-Rubinstein Distance and Trend Tests	94

4	K-Lipschitz Maps and the Kantorovich-Rubinstein distance	97
4.1	Functorial properties of the Kantorovich-Rubinstein distance . . .	98
4.2	Projections and the Kantorovich-Rubinstein distance	105
5	An Introduction to Classification problems	109
5.1	Statistical Machine Learning	109
5.2	Classification	110
5.3	Bayes Error and the Estimation Error	111
5.4	Representational Capacity of a Model and Generalization Ability of a Classifier	113
5.5	Underfitting, Overfitting and Regularisation	114
5.6	Loss Functions and Confidence Margins	117
5.7	Rademacher Averages and Generalization Bounds	123
6	Classifiers and the Kantorovich-Rubinstein Distance	133
6.1	Definition of measurable classifiers	134
6.2	Lipschitz classifiers and the Kantorovich-Rubinstein distance . . .	143
6.3	A more general notion of error	149
7	Applications of the Kantorovich-Rubinstein Distance in Machine Learning	152
7.1	The Kantorovich-Rubinstein distance as a descriptor of sample complexity - 1.9	153
7.2	The Kantorovich-Rubinstein distance and feature-selection algo- rithms - 1.64	157
7.2.1	Branch & Bound algorithms - 1.73	158
7.2.2	The use of the Kantorovich-Rubinstein distance to construct an evaluation criterion function - 1.86	160
8	Conclusion	163

A Elementary Definitions and Results	167
A.1 Disintegration Theorem	167
A.2 Total Variation Norm	168
B A Quick Overview of Feature Selection Algorithms	174
B.0.1 Subset Generation	174
B.0.2 Subset Evaluation	175
B.0.3 Stopping Criteria	176
B.0.4 Result Validation	176
C Biological and Technical Background	177
C.1 Coronary Artery Disease	177
C.2 DNA and SNPs	178
C.3 DNA Microarrays	180
C.4 Chiamo: a Genotype Calling Algorithm	182
C.5 Quality Control	187
C.6 Information on OHGS Dataset	188
D A geometrical perspective of the Kantorovich-Rubinstein Dis- tance	191
Bibliography	202

List of Tables

3.1 Nomenclature for Contingency Table with one SNP	72
C.1 Information on the number of SNPs, from each of Chromosomes 1 to 22, in the OHGS-2 dataset.	190

Preface

If one ever needs an example to illustrate the proverb “Necessity is the mother of invention”, one should look no further than the field of *optimal transport*. The field’s birth can be traced to the publication, by the French geometer Gaspard Monge, of his famous work *Mémoire sur la théorie des déblais et des remblais* in 1781. Monge considered the following problem: suppose you have crushed stones that you need to extract from quarries and transport to construction sites. The location of the quarries and the construction sites are known. The volume extracted from each quarry and the volume needed at each construction site are also known. The problem is to decide for each quarry how to dispatch the crushed stones in such a way as to minimize the total transport cost. Monge assumed that the transport cost of one unit of mass along a certain distance was given by the product of the mass by the distance. It is to construct that optimal transportation strategy that Monge created the field of *Optimal Transport*. Most of Monge’s result were flawed (by current mathematical standards) and the field of optimal transport remained dormant for more than a century. The revival of Optimal Transport came in the 1930’s when the Russian mathematician Kantorovich realised that an optimal transport problem was in fact a particular case of an optimal coupling problem. In his quest to solve optimal coupling problems, Kantorovich stated and proved a fundamental duality theorem. He also defined a very useful notion of distance between two measures: the total cost of the optimal coupling when the cost is chosen as a distance function on the underlying product space. This distance is called the Kantorovich-Rubinstein distance. Throughout the second half of the twentieth century, statisticians and probabilists used the

Kantorovich-Rubinstein distance in the study of many different fields of Mathematics and Physics. In this thesis, we study the use of the Kantorovich-Rubinstein distance to analyse the difficulty of a classification problem. This type of analysis tries to predict under which scenarios a given classifier succeeds or fails without proceeding with the classification per se. We then focus on constructing this Kantorovich-Rubinstein based complexity descriptor for a Genome Wide Association Study (GWAS) dataset. This thesis will be structured as follows:

Chapter 1 introduces the Kantorovich Minimisation problem, its associated duality theorem and its particular case, the Kantorovich-Rubinstein theorem. It finishes with a proof of the existence of optimal solutions for the Kantorovich Minimisation problem. The results presented in this chapter are known.

Chapter 2 focuses on the Kantorovich-Rubinstein distance as a distance between measures. Once the Kantorovich-Rubinstein distance is defined, we study the convergence of measures and its topological properties. Then, we construct explicit formulas for the Kantorovich-Rubinstein distance for particular Polish spaces. The vast majority of results stated in this chapter are known. There are nonetheless a few new corollaries and new proofs for the old results.

Chapter 3 is new material. It studies the link between commonly used test statistics for genetic association and the Kantorovich-Rubinstein distance.

Chapter 4 is also new material. It is a technical chapter that studies the functional properties of the Kantorovich-Rubinstein distance in order to compare the Kantorovich-Rubinstein distance between two measures and between their respective push-forward measures.

Chapter 5 introduces the basic notions of classification problems as a particular case of statistical learning problems. We thus formalize the notions of loss functions and real-valued classification functions, overfitting and its link to generalization bound with the Rademacher averages as particular case of representational capacity. We end this overview of learning theory for classification problems with an introduction of margin theory.

Chapter 6 is the central chapter of the thesis and only contains new material. It studies thoroughly the association between the Kantorovich-Rubinstein distance and the risk functional for particular choices of loss functions.

Chapter 7 introduces the notion of sample complexity and explains why the Kantorovich-Rubinstein distance can be used as a descriptor of a sample complexity. Then, using the fact that the Kantorovich-Rubinstein distance can be used as a descriptor of a sample complexity, we show that the Kantorovich-Rubinstein distance is a good choice of evaluation criterion function in feature selection algorithms that could be of great interest to reduce the dimensionality of GWAS datasets such as the Ottawa Heart Genomics Study dataset (OHGS).

Finally, chapter 8, the conclusion chapter, addresses limitations of the Kantorovich-Rubinstein distance as a sample complexity and gives precise modifications to (we hope) improve the descriptive capability of the Kantorovich-Rubinstein distance.

Lastly, there are four appendices. Appendix A features some known definitions and results used in the thesis but not directly related to optimal transport or classification problems. Appendix B gives a brief overview of feature selection algorithms and groups them in different categories. Appendix C gives the biological, genetical and technical information necessary to understand how GWAS datasets, and more particularly the OHGS dataset, are constructed. Appendix D gives a description of the Kantorovich-Rubinstein distance from a geometrical perspective. It is given in the Appendix because it opens up (we hope) to a generalisation of the results in this thesis.

Chapter 1

Kantorovich minimisation problem and Duality Theorems

Chapter 1 starts with the description and the characterisation of the optimisation problem called the *Kantorovich Minimisation problem*. Then, we establish a theorem of major importance to the field, the Kantorovich Duality Theorem and study its particular case of interest, the Kantorovich-Rubinstein Theorem. The chapter ends with the a proof of the existence of optimal solutions for the optimisation problem under certain regularity assumptions.

The important results presented in this chapter are found in two books by Villani: *Topics in Optimal Transportation* [53] and *Optimal transport, old and new* [54].

Throughout this chapter (and more generally, throughout this thesis), we will often omit to specify the σ -algebras associated to a space if no specific σ -algebra is required for the result to stand or if the σ -algebra used is clear from the context. Note that, generally, in the case of a topological space, the σ -algebra is the Borel σ -algebra.

1.1 Couplings

Couplings are very well-known in probability theory. Since they are central to the definition of the Kantorovich minimisation problem, we recall both the measure theoretical and the probabilistic definitions. We first start with the definition of a *marginal* as it appears in the definition of a *coupling*.

Proposition 1.1.1. *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) be two probability spaces and let ϑ be probability measure on the product space $\mathcal{X} \times \mathcal{Y}$. Then, the following three statements are equivalent:*

- (i) $\mu(A) = \vartheta(A \times \mathcal{Y})$ and $\nu(B) = \vartheta(\mathcal{X} \times B)$ for all measurable subsets $A, B \subset \mathcal{X}$.
- (ii) $\pi_{\mathcal{X}}(\vartheta) = \mu$ and $\pi_{\mathcal{Y}}(\vartheta) = \nu$ where $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ denote the natural projections of $\mathcal{X} \times \mathcal{Y}$ onto \mathcal{X} and \mathcal{Y} respectively.
- (iii) For all integrable measurable functions φ , respectively ψ , on \mathcal{X} , respectively \mathcal{Y} ,

$$\int_{\mathcal{X} \times \mathcal{Y}} (\varphi(x) + \psi(y)) \, d\vartheta(x, y) = \int_{\mathcal{X}} \varphi(x) \, d\mu(x) + \int_{\mathcal{Y}} \psi(y) \, d\nu(y).$$

Note that formally, one should write

$$\int_{\mathcal{X} \times \mathcal{Y}} ((\varphi \circ \pi_{\mathcal{X}})(x, y) + (\psi \circ \pi_{\mathcal{Y}})(x, y)) \, d\vartheta(x, y) \quad \text{instead of} \quad \int_{\mathcal{X} \times \mathcal{Y}} (\varphi(x) + \psi(y)) \, d\vartheta(x, y).$$

Definition 1.1.2 (Marginals). *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) be two probability spaces and let ϑ be a probability measure on the product space $\mathcal{X} \times \mathcal{Y}$. Then, the probability measure ϑ admits the measure μ and ν as marginals on \mathcal{X} and \mathcal{Y} respectively, if μ and ν satisfy any of the conditions in Proposition 1.1.1.*

Proof: (Property 1.1.1)

The proposition will be established in the following way: (i) \Leftrightarrow (ii), (ii) \Rightarrow (iii) and (iii) \Rightarrow (i).

(i) \Leftrightarrow (ii): For a measurable subset $A \subset \mathcal{X}$ we have:

$$\pi_{\mathcal{X}}(\vartheta)(A) = \vartheta(\pi_{\mathcal{X}}^{-1}(A)) = \vartheta(A \times \mathcal{Y}) = \mu(A).$$

Likewise, for a measurable subset $B \subset \mathcal{Y}$ we obtain $\pi_{\mathcal{Y}}(\vartheta)(B) = \nu(B)$.

Conversely, we have $\mu(A) = \pi_{\mathcal{X}}(\vartheta)(A) = \vartheta(\pi_{\mathcal{X}}^{-1}(A)) = \vartheta(A \times \mathcal{Y})$.

Likewise, $\nu(B) = \vartheta(\mathcal{X} \times B)$.

(ii) \Rightarrow (iii): Let φ (respectively ψ) be measurable functions on \mathcal{X} (respectively \mathcal{Y}).

We have

$$\int_{\mathcal{X} \times \mathcal{Y}} (\varphi \circ \pi_{\mathcal{X}})(x, y) + (\psi \circ \pi_{\mathcal{Y}})(x, y) \, d\vartheta(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} (\varphi \circ \pi_{\mathcal{X}})(x, y) + \int_{\mathcal{X} \times \mathcal{Y}} (\psi \circ \pi_{\mathcal{Y}})(x, y) \, d\vartheta(x, y).$$

The change of variable formula yields:

$$\int_{\mathcal{X} \times \mathcal{Y}} (\varphi \circ \pi_{\mathcal{X}})(x, y) + \int_{\mathcal{X} \times \mathcal{Y}} (\psi \circ \pi_{\mathcal{Y}})(x, y) \, d\vartheta(x, y) = \int_{\mathcal{X}} \varphi(x) \, d\pi_{\mathcal{X}}(\vartheta)(x) + \int_{\mathcal{Y}} \psi(y) \, d\pi_{\mathcal{Y}}(\vartheta)(y).$$

Now, by (ii), we have

$$\int_{\mathcal{X}} \varphi(x) \, d\pi_{\mathcal{X}}(\vartheta)(x) + \int_{\mathcal{Y}} \psi(y) \, d\pi_{\mathcal{Y}}(\vartheta)(y) = \int_{\mathcal{X}} \varphi(x) \, d\mu(x) + \int_{\mathcal{Y}} \psi(y) \, d\nu(y).$$

(iii) \Rightarrow (i): For any $A \subset \mathcal{X}$,

$$\begin{aligned} \mu(A) &= \int_{\mathcal{X}} 1_A(x) \, d\mu(x) \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \pi_{\mathcal{X}}(1_A)(x, y) \, d\vartheta(x, y) \quad \text{by (iii)} \\ &= \int_{A \times \mathcal{Y}} (x, y) \, d\vartheta(x, y) \\ &= \vartheta(A \times \mathcal{Y}). \end{aligned}$$

Hence we have $\mu(A) = \vartheta(A \times \mathcal{Y})$.

Likewise, for any $B \subset \mathcal{Y}$, we have $\nu(B) = \vartheta(\mathcal{X} \times B)$. ■

Definition 1.1.3 (Couplings). *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) be two probability spaces. A coupling of μ and ν is a probability measure ϑ on the product space $\mathcal{X} \times \mathcal{Y}$ such that ϑ admits μ and ν as marginals on \mathcal{X} and \mathcal{Y} , respectively.*

Remark 1.1.4. We denote by $\mathcal{N}(\mu, \nu)$ the set of all couplings of μ and ν . Note that $\mathcal{N}(\mu, \nu) \neq \emptyset$, because the product measure $\mu \times \nu$ belongs to $\mathcal{N}(\mu, \nu)$.

Remark 1.1.5. The Definition 1.1.3 of a coupling can be generalized to the case of any two bounded measures μ and ν such that $\mu(\mathcal{X}) = \nu(\mathcal{Y})$. Indeed, suppose that $\mu(\mathcal{X}) = \nu(\mathcal{Y}) = k$ with $k < \infty$. Since $\mu(\mathcal{X}) = \vartheta(\mathcal{X} \times \mathcal{Y})$ then $\vartheta(\mathcal{X} \times \mathcal{Y}) = k$. Since a probability measure is a normalized bounded measure, and since all publications in the field of Optimal Transport Theory consider probability spaces, all results in this thesis will be written for probability spaces \mathcal{X} and \mathcal{Y} . Please keep in mind that these results remain true for measurable spaces equipped with bounded measures μ and ν such that $\mu(\mathcal{X}) = \nu(\mathcal{Y})$.

There also exists a classical probabilistic definition of a coupling. This definition relies on probability theory terminology. We thus first recall some basic definitions:

Definition 1.1.6 (Random Variable). *Let \mathcal{X} be a measurable space and let $(\Omega_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$ be a probability space.*

Then, a measurable map $X: \Omega_{\mathcal{X}} \rightarrow \mathcal{X}$ is called a random variable.

Definition 1.1.7 (Push-forward measures). *Let \mathcal{X} be a measurable space and let $(\Omega_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$ be a probability space. Let $X: \Omega_{\mathcal{X}} \rightarrow \mathcal{X}$ be a random variable. The push-forward measure $X(\mathbb{P}_{\mathcal{X}})$ of $\mathbb{P}_{\mathcal{X}}$ by X is the measure on \mathcal{X} defined by*

$$X(\mathbb{P}_{\mathcal{X}})(A) = \mathbb{P}_{\mathcal{X}}(\{\omega \in \Omega_{\mathcal{X}}; X(\omega) \in A\}),$$

for all a measurable set A in \mathcal{X} .

The push-forward measure $X(\mathbb{P}_{\mathcal{X}})$ is also called the law of X . One writes $\text{law}(X)$.

Remark 1.1.8. Let φ be a map from a measured space (\mathcal{X}, μ) to a space \mathcal{Y} . Several notations are used to write the push-forward of μ by φ , for example $\varphi\#\mu$ or $\varphi * (\mu)$. In this thesis, we will use the notation $\varphi(\mu)$.

We can now give the probabilistic definition of a *coupling*:

Definition 1.1.9 (Couplings in probabilistic formulation). *Let \mathcal{X} (respectively \mathcal{Y}) be a measurable space and $(\Omega_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$ (respectively $(\Omega_{\mathcal{Y}}, \mathbb{P}_{\mathcal{Y}})$) be a probability space.*

Let $X : \Omega_{\mathcal{X}} \rightarrow \mathcal{X}$ and $Y : \Omega_{\mathcal{Y}} \rightarrow \mathcal{Y}$ be two random variables.

A couple (\tilde{X}, \tilde{Y}) of random variables on the probability space $(\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}, \tilde{\mathbb{P}})$ is a coupling of X and Y if $\text{law}(\tilde{X}) = \text{law}(X)$ on \mathcal{X} and $\text{law}(\tilde{Y}) = \text{law}(Y)$ on \mathcal{Y} .

Hence, the push-forward measure $(\tilde{X}, \tilde{Y})(\tilde{\mathbb{P}})$ on $\mathcal{X} \times \mathcal{Y}$ is a coupling of $X(\mathbb{P}_{\mathcal{X}})$ and $Y(\mathbb{P}_{\mathcal{Y}})$ in the sense of definition 1.1.3.

In the literature, many authors are content with a reduced terminology and refer to the pair of random variables (\tilde{X}, \tilde{Y}) as a coupling of the probability measures $X(\mathbb{P}_{\mathcal{X}})$ and $Y(\mathbb{P}_{\mathcal{Y}})$. Hence, we can often find the following definition:

Definition 1.1.10 (Couplings in probabilistic formulation, version 2). *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) be two probability spaces and let $X : \Omega_{\mathcal{X}} \rightarrow \mathcal{X}$ and $Y : \Omega_{\mathcal{Y}} \rightarrow \mathcal{Y}$ be two random variables.*

Coupling μ and ν means constructing two random variables X and Y on some probability space (Ω, \mathbb{P}) such that $\text{law}(X) = \mu$, $\text{law}(Y) = \nu$. The couple (X, Y) is called a coupling of (μ, ν) .

Example 1.1.11 (Coupling for atomic measures). Let \mathcal{X} and \mathcal{Y} be two spaces and define two atomic probability measures

$$\mu = \sum_{i=1}^n \mu_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{i=1}^m \nu_i \delta_{y_i},$$

supported on $\{x_1, \dots, x_n\} \in \mathcal{X}$ and $\{y_1, \dots, y_m\} \in \mathcal{Y}$, respectively. Then,

$$\mathcal{N}(\mu, \nu) = \left\{ \sum_{i=1}^n \sum_{j=1}^m \vartheta_{ij} \delta_{(x_i, y_j)} : \forall i, \sum_{j=1}^m \vartheta_{ij} = \mu_i \quad \text{and} \quad \forall j, \sum_{i=1}^n \vartheta_{ij} = \nu_j \quad \text{with} \quad \vartheta_{ij} \geq 0 \right\}.$$

1.2 Kantorovich Minimisation Problem

We now have the required mathematical definitions to define the *Kantorovich minimisation problem*.

Definition 1.2.1 (Kantorovich minimisation problem). *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) be two probability spaces and let $\mathcal{N}(\mu, \nu)$ be the set of all couplings of μ and ν . Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{\infty\}$ be a nonnegative measurable function, called the cost function. For a coupling $\vartheta \in \mathcal{N}(\mu, \nu)$, consider the functional I defined by*

$$I(\vartheta) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\vartheta(x, y).$$

Then the Kantorovich minimisation problem is to find:

$$\inf\{I(\vartheta); \vartheta \in \mathcal{N}(\mu, \nu)\}.$$

In probability theory, one would write:

Find $\inf \mathbb{E} c(X, Y)$, where the pair of random variables (X, Y) runs over all possible couplings of (μ, ν) .

Of course, the solution of the Kantorovich minimisation problem depends on the cost function c . The cost function and the probability spaces here can be very general. In his books [53], [54], Villani obtains nontrivial results as soon as c is lower semi-continuous and \mathcal{X}, \mathcal{Y} are Polish spaces.

Definition 1.2.2 (Polish Space). *A Polish space \mathcal{X} is a topological space which is separable and completely metrizable.*

Recall that a completely metrizable space is a topological space (\mathcal{X}, T) for which there exists at least one metric d on \mathcal{X} such that (\mathcal{X}, d) is a complete metric space and d induces the topology T .

In this thesis, we consider Polish spaces with their Borel σ -algebra.

Definition 1.2.3 (Lower Semi-continuous Function). *Let \mathcal{X} be a topological space. A function $f : \mathcal{X} \rightarrow [-\infty, \infty]$ is lower semi-continuous if, for any $\alpha \in \mathbb{R}$, $f^{-1}((\alpha, +\infty]) =$*

$\{x \in \mathcal{X}; f(x) > \alpha\}$ is open in \mathcal{X} .

Remark 1.2.4. If \mathcal{X} is a Polish space, a function $f : \mathcal{X} \rightarrow [-\infty, \infty]$ is lower semi-continuous if, for all $x_o \in \mathcal{X}$,

$$f(x_o) \leq \liminf_{x \rightarrow x_o} f(x).$$

Moreover, each lower semi-continuous function $f : \mathcal{X} \rightarrow \mathbb{R}_+$ is the (pointwise) supremum of an increasing sequence of uniformly continuous nonnegative functions (see Villani p.26 [53])

Historical Note 1.2.5. The Kantorovich minimisation problem is named after the Russian mathematician Leonid Vitaliyevich Kantorovich. Born in 1912, Kantorovich was a very gifted mathematician who made his reputation as a first-class researcher at the age of 18, and earned a position of professor at just 22 at the University of Leningrad. He worked in many areas of mathematics, with a strong taste for applications in economics, and theoretical computer science. In 1938 a laboratory consulted him for the solution of a production optimization problem, which he found out was representative of a whole class of linear problems arising in various areas of economics. He received the Nobel Prize in Economic Sciences in 1975. It was shared with Tjalling Koopmans, and it was given "for their contributions to the theory of optimum allocation of resources."

1.2.1 Kantorovich Duality Theorem

In 1942, Kantorovich stated and proved, by means of functional analytical tools, a duality theorem that is both well-known and widely used. It is stated below. Note that, for a probability space (\mathcal{X}, μ) , the notation $L_1(\mu)$ represents all the integrable functions on \mathcal{X} with respect to μ .

Theorem 1.2.6 (Kantorovich Duality Theorem - Part 1). *Let μ and ν be two probability measures on the Polish spaces \mathcal{X} and \mathcal{Y} respectively. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$*

be a nonnegative lower semi-continuous cost function.

Let Φ_c be the set of all pairs of measurable functions $(\varphi, \psi) \in L_1(\mu) \times L_1(\nu)$ satisfying

$$\varphi(x) + \psi(y) \leq c(x, y), \quad (1.2.1)$$

for μ -almost all $x \in \mathcal{X}$, ν -almost all $y \in \mathcal{Y}$.

Then, for $(\varphi, \psi) \in \Phi_c$, we have

$$\inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta] = \sup_{\Phi_c} J(\varphi, \psi), \quad (1.2.2)$$

where

$$J(\varphi, \psi) = \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) \quad \text{and} \quad I[\vartheta] = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\vartheta(x, y).$$

The right hand side of equation (1.2.2) is known as the dual formulation of the Monge-Kantorovich minimisation problem, which is itself often called the primal problem.

In the case of atomic measures with finite support, the Monge-Kantorovich theorem is equivalent to the well known linear programming duality:

Corollary 1.2.7 (Kantorovich Duality Theorem for atomic measures). *Let \mathcal{X} and \mathcal{Y} be two Polish spaces. Let us define two atomic probability measures*

$$\mu = \sum_{i=1}^n \mu_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{i=1}^m \nu_i \delta_{y_i},$$

supported on $\{x_1, \dots, x_n\} \in \mathcal{X}$ and $\{y_1, \dots, y_m\} \in \mathcal{Y}$, respectively.

Recall that any coupling ϑ of μ and ν is an atomic measure on $\mathcal{X} \times \mathcal{Y}$ written as $\sum \sum \vartheta_{ij} \delta_{(x_i, y_j)}$, for $i = 1, \dots, n$ and $j = 1, \dots, m$.

Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a nonnegative lower semi-continuous cost function.

Then, the functional

$$\min \left\{ \sum_{i=1}^n \sum_{j=1}^m c(x_i, y_j) \vartheta_{ij} : \forall i, \sum_{j=1}^m \vartheta_{ij} = \mu_i \quad \text{and} \quad \forall j, \sum_{i=1}^n \vartheta_{ij} = \nu_j \quad \text{with} \quad \vartheta_{ij} \geq 0 \right\}$$

admits the dual representation

$$\max \left\{ \sum_{i=1}^n \varphi_i \mu_i + \sum_{j=1}^m \psi_j \nu_j; \varphi_i + \psi_j \leq c(x_i, y_j), \forall i, j \right\}.$$

Remark 1.2.8 (Regarding theorem 1.2.6). Notice that inequality (1.2.1) in Theorem 1.2.6 holds ϑ -a.e. For any $x_o \in \mathcal{X}$ and $y_o \in \mathcal{Y}$, let S_{x_o} and S_{y_o} be measurable sets of measure zero on \mathcal{Y} and \mathcal{X} , respectively, on which inequality (1.2.1) is not satisfied.

We have

$$S_{x_o}^c = \{y \in \mathcal{Y} : \varphi(x_o) + \psi(y) \leq c(x_o, y)\} \quad \text{and} \quad S_{y_o}^c = \{x \in \mathcal{X} : \varphi(x) + \psi(y_o) \leq c(x, y_o)\}.$$

Then, equation (1.2.1) holds for all $(x, y) \in S_{y_o}^c \times S_{x_o}^c$. Thus, for equation (1.2.1) to hold ϑ -a.e., we need to show that $\vartheta((S_{y_o}^c \times S_{x_o}^c)^c) = 0$. Now, $(S_{y_o}^c \times S_{x_o}^c)^c = (S_{y_o} \times \mathcal{Y}) \cup (\mathcal{X} \times S_{x_o})$ and ϑ has marginals μ and ν . Hence $\vartheta(S_{y_o} \times \mathcal{Y}) = 0$ and $\vartheta(\mathcal{X} \times S_{x_o}) = 0$. Therefore, $\vartheta((S_{y_o}^c \times S_{x_o}^c)^c) = 0$.

One can show that the value of the supremum of J on Φ_c is the same as the value of the supremum of J if one restricts Φ_c to the functions (φ, ψ) that are bounded and continuous. It is not obvious that pairs of L_1 functions satisfying equation (1.2.1) can be approximated by pairs of continuous function also satisfying (1.2.1). More details are in Proposition 1.2.9:

Proposition 1.2.9. *Let μ and ν be probability measures on the Polish spaces \mathcal{X} and \mathcal{Y} respectively. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a nonnegative lower semi-continuous function.*

Let Φ_c be defined as in 1.2.6 and Φ'_c be defined as Φ_c but restricted to the functions

(φ', ψ') which are bounded and continuous. Then,

$$\sup_{\Phi'_c} J(\varphi', \psi') \leq \sup_{\Phi_c} J(\varphi, \psi) \leq \inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta]. \quad (1.2.3)$$

where $J(\varphi, \psi) = \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y)$ and $I[\vartheta] = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\vartheta$.

Proof: (Proposition 1.2.9)

The inequality on the left of 1.2.3 is trivial since $C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \subset L_1(\mu) \times L_1(\nu)$. Hence, we only have to consider the inequality on the right. Let $(\varphi, \psi) \in \Phi_c$ and let $\vartheta \in \mathcal{N}(\mu, \nu)$. By Proposition 1.1.1, we have

$$J(\varphi, \psi) = \int_{\mathcal{X} \times \mathcal{Y}} (\varphi(x) + \psi(y)) d\vartheta(x, y).$$

Moreover, since $\varphi(x) + \psi(y) \leq c(x, y)$, for ϑ -almost all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we obtain

$$\int_{\mathcal{X} \times \mathcal{Y}} (\varphi(x) + \psi(y)) d\vartheta(x, y) \leq \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\vartheta(x, y).$$

As the supremum is the lowest upper bound, we have $\sup_{\Phi_c} J(\varphi, \psi) \leq \inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta]$, which completes the proof. ■

Now, it follows from Proposition 1.2.9 that the duality

$$\sup_{\Phi'_c} J(\varphi', \psi') = \inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta] \quad \text{implies that} \quad \sup_{\Phi'_c} J(\varphi', \psi') = \sup_{\Phi_c} J(\varphi, \psi).$$

We can now give the general idea of the proof of the Monge-Kantorovich Theorem:

Proof: (Theorem 1.2.6)

To prove that the infimum is at least as large as the supremum is easy. It has been proved in Proposition 1.2.9.

The proof for the reverse inequality is much more complicated. It is separated

in 3 steps by increasing order of generality. The first step assumes that \mathcal{X} and \mathcal{Y} are compact and c is continuous. It uses a minimax argument which is an argument of the form

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \varphi(x, y) = \sup_{a \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \varphi(x, y).$$

The minimax argument allows to show that

$$\inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta] \leq \sup_{\Phi'_c} J(\varphi', \psi'). \tag{1.2.4}$$

Putting together inequalities (1.2.3) and (1.2.4), one obtains

$$\sup_{\Phi'_c} J(\varphi', \psi') = \inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta] \quad \text{and thus} \quad \sup_{\Phi_c} J(\varphi, \psi) = \inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta],$$

which proves the duality (1.2.2) in this special case.

The last two steps consist in showing that the duality (1.2.2) holds with relaxed assumptions, using approximation arguments. The second step relaxes the assumption of compactness while making the assumption that c is bounded and uniformly continuous. Finally, in the third step, the condition of continuity on c is relaxed to obtain the general result. To do so, one has to write $c = \sup c_n$, where c_n is a nondecreasing sequence of bounded, nonnegative, uniformly continuous cost functions.

A detailed version of the proof is available in *Topics in Optimal Transportation* [54] and *Optimal Transport: Old and New* [53], both by Villani. ■

1.2.2 Existence of Optimal Solutions

As its name indicates, the Kantorovich Minimisation Problem requires to find an infimum. A natural question is therefore to ask whether this infimum is realised. That is to say, do there exist, over all couplings, minimisers of the Kantorovich Minimisation Problem?

In light of the Kantorovich Duality Theorem 1.2.6, it is also natural to ask whether the dual formulation achieves its supremum. That is, do there exist, over all pairs of functions in Φ_c , maximisers for J ?

In fact, the answer to both questions is yes. It is stated in Theorem 1.2.15 (Kantorovich Duality Theorem Part 2. - Existence of optimal solutions). In order to write Theorem 1.2.15 formally, we need to introduce the notions of c -concavity, c -transform, and conjugate c -transform functions.

Definition 1.2.10 (c -transform). *Let \mathcal{X} and \mathcal{Y} be two nonempty sets and let $c(x, y)$ be defined on $\mathcal{X} \times \mathcal{Y}$ with values in $\mathbb{R} \cup \{+\infty\}$.*

For any function $\varphi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$, $\varphi \not\equiv -\infty$, one can define its c -transform $\varphi^c : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$ by

$$\varphi^c(y) = \inf_{x \in \mathcal{X}} [c(x, y) - \varphi(x)].$$

The functions φ and φ^c are said to be c -conjugate.

Proposition 1.2.11. *Let \mathcal{X} and \mathcal{Y} be two nonempty sets and let $c(x, y)$ be defined on $\mathcal{X} \times \mathcal{Y}$ with values in $\mathbb{R} \cup \{+\infty\}$. For any function $\varphi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$, $\varphi \not\equiv -\infty$, one has the identity $\varphi^{ccc} = \varphi^c$ where $\varphi^{cc} = (\varphi^c)^c$.*

Note that both φ^{ccc} and φ^c are functions defined on \mathcal{X} while φ and φ^{cc} are defined on \mathcal{Y} .

Proof: (Proposition 1.2.11)

Consider a function $\varphi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ and its c -transform φ^c . Then, we have

$$\begin{aligned} \varphi^{cc} &= \inf_{y \in \mathcal{Y}} [c(x, y) - \varphi^c(y)] \\ &= \inf_{y \in \mathcal{Y}} \left[c(x, y) - \inf_{\tilde{x} \in \mathcal{X}} [c(\tilde{x}, y) - \varphi(\tilde{x})] \right] \\ &= \inf_{y \in \mathcal{Y}} \left[c(x, y) + \sup_{\tilde{x} \in \mathcal{X}} [\varphi(\tilde{x}) - c(\tilde{x}, y)] \right] \\ &= \inf_{y \in \mathcal{Y}} \sup_{\tilde{x} \in \mathcal{X}} [\varphi(\tilde{x}) + c(x, y) - c(\tilde{x}, y)]. \end{aligned}$$

and therefore,

$$\begin{aligned}
 \varphi^{ccc}(y) &= \inf_{x \in \mathcal{X}} [c(x, y) - \varphi^{cc}(x)] \\
 &= \inf_{x \in \mathcal{X}} \left[c(x, y) - \inf_{\tilde{y} \in \mathcal{Y}} \sup_{\tilde{x} \in \mathcal{X}} [\varphi(\tilde{x}) + c(x, \tilde{y}) - c(\tilde{x}, \tilde{y})] \right] \\
 &= \inf_{x \in \mathcal{X}} \left[c(x, y) + \sup_{\tilde{y} \in \mathcal{Y}} [-\sup_{\tilde{x} \in \mathcal{X}} [\varphi(\tilde{x}) + c(x, \tilde{y}) - c(\tilde{x}, \tilde{y})]] \right] \\
 &= \inf_{x \in \mathcal{X}} \left[c(x, y) + \sup_{\tilde{y} \in \mathcal{Y}} \inf_{\tilde{x} \in \mathcal{X}} [c(\tilde{x}, \tilde{y}) - c(x, \tilde{y}) - \varphi(\tilde{x})] \right] \\
 &= \inf_{x \in \mathcal{X}} \sup_{\tilde{y} \in \mathcal{Y}} \inf_{\tilde{x} \in \mathcal{X}} [c(x, y) + c(\tilde{x}, \tilde{y}) - c(x, \tilde{y}) - \varphi(\tilde{x})].
 \end{aligned}$$

If we set $\tilde{x} = x$ we obtain

$$\varphi^{ccc}(y) \leq \inf_{x \in \mathcal{X}} \sup_{\tilde{y} \in \mathcal{Y}} [c(x, y) - \varphi(x)]. \text{ That is } \varphi^{ccc}(y) \leq \inf_{x \in \mathcal{X}} [c(x, y) - \varphi(x)] = \varphi^c(y).$$

On the other hand, by setting $\tilde{y} = y$ we obtain

$$\varphi^{ccc}(y) \geq \inf_{x \in \mathcal{X}} \inf_{\tilde{x} \in \mathcal{X}} [c(\tilde{x}, y) - \varphi(\tilde{x})]. \text{ That is } \varphi^{ccc}(y) \geq \inf_{\tilde{x} \in \mathcal{X}} [c(\tilde{x}, y) - \varphi(\tilde{x})] = \varphi^c(y).$$

■

Definition 1.2.12. [*c-concavity*] Let \mathcal{X} and \mathcal{Y} be two nonempty sets and let $c(x, y)$ be defined on $\mathcal{X} \times \mathcal{Y}$ with values in $\mathbb{R} \cup \{+\infty\}$.

A function $\rho : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ is said to be *c-concave* if there exists $\psi : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$, $\psi \not\equiv -\infty$, such that $\rho(x) = \psi^c(x)$, where

$$\psi^c(x) = \inf_{y \in \mathcal{Y}} [c(x, y) - \psi(y)].$$

That is ρ is *c-concave* if there exists ψ such that ρ is the *c-transform* of ψ .

Proposition 1.2.13 (Alternative characterization of *c-concavity*). Let \mathcal{X} and \mathcal{Y} be two nonempty sets and let $c(x, y)$ be defined on $\mathcal{X} \times \mathcal{Y}$ with values in $\mathbb{R} \cup \{+\infty\}$.

Consider a function $\rho : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$, $\rho \not\equiv -\infty$.

Then ρ is c -concave if and only if $\rho^{cc} = \rho$, where $\rho^{cc} = (\rho^c)^c$.

Proof: (Proposition 1.2.13)

Suppose that ρ is c -concave. Then, there exists $\psi : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$, $\psi \not\equiv -\infty$, such that $\rho(x) = \psi^c(x)$, for all $x \in \mathcal{X}$. By Proposition 1.2.11, $\psi^c = \psi^{ccc}$. Thus $\rho = (\psi^c)^{cc} = \rho^{cc}$.

Conversely, suppose that $\rho(x) = \rho^{cc}(x)$ for all $x \in \mathcal{X}$. Let $\rho^c(y) = \psi(y)$. Then, $\psi^c(x) = \rho^{cc}(x) = \rho(x)$. Hence, there exists ψ (namely ρ^c) such that ρ is the c -transform of ψ^c (namely ρ^{cc}). ■

Remark 1.2.14. Note that for any not necessarily c -concave function φ , $\varphi \not\equiv -\infty$, its c -transform φ^c is always c -concave. Indeed, by Proposition 1.2.11, we have $\varphi^c = \varphi^{ccc} = (\varphi^c)^{cc}$. Thus, by Proposition 1.2.13, φ^c is c -concave.

We can now state the Kantorovich Duality Theorem Part 2 - Existence of optimal solutions:

Theorem 1.2.15 (Kantorovich Duality Theorem Part 2 - Existence of optimal solutions). *Let μ and ν be probability measures on the Polish spaces \mathcal{X} and \mathcal{Y} respectively. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a nonnegative lower semi-continuous cost function.*

Let Φ_c be the set of all measurable functions $(\varphi, \psi) \in L_1(\mu) \times L_1(\nu)$ satisfying

$$\varphi(x) + \psi(y) \leq c(x, y) \tag{1.2.5}$$

for μ -almost all $x \in \mathcal{X}$, ν -almost all $y \in \mathcal{Y}$.

As defined in Theorem 1.2.6, let

$$J(\varphi, \psi) = \int_{\mathcal{X}} \varphi(x) \, d\mu(x) + \int_{\mathcal{Y}} \psi(y) \, d\nu(y) \quad \text{and} \quad I[\vartheta] = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\vartheta(x, y).$$

(i) Then, the infimum of $I[\vartheta]$ over all $\vartheta \in \mathcal{N}(\mu, \nu)$ is attained. That is, there exists $\vartheta^* \in \mathcal{N}(\mu, \nu)$ such that

$$\inf_{\vartheta \in \mathcal{N}(\mu, \nu)} I[\vartheta] = I[\vartheta^*].$$

(ii) Assume, moreover, that there exist nonnegative measurable functions $c_{\mathcal{X}} \in L_1(\mu)$ and $c_{\mathcal{Y}} \in L_1(\nu)$ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$, $c(x, y) \leq c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y)$. Then, the supremum of $J(\varphi, \psi)$ over all $(\varphi, \psi) \in \Phi_c$ is attained. Indeed, the dual Kantorovich problem admits a maximiser in the form of a pair of conjugate c -concave functions (ρ, ρ^c) . Hence,

$$\sup_{\Phi_c} J(\varphi, \psi) = \max_{\rho \in L_1(\mu)} \left(\int_{\mathcal{X}} \rho(x) d\mu(x) + \int_{\mathcal{Y}} \rho^c(y) d\nu(y) \right).$$

The proof of existence of an optimal coupling (Theorem 1.2.15, Part (i)) requires the notion of tightness of measures, basic results associated to the notion of tightness and two theorems: Prokhorov theorem and Portmanteau theorem.

For the proof of the existence of a maximiser for the dual Kantorovich problem (Theorem 1.2.15, Part (ii)), one can consult p.86 of Villani's *Optimal Transport: Old and New* [54].

To give the definition of the tightness of a measure, one needs to define a topology (and thus a convergence) on the space of Borel probability measures $P(\mathcal{X})$.

Definition 1.2.16. [Weak Convergence] Let (\mathcal{X}, d) be a metric space and $P(\mathcal{X})$ be its space of Borel probability measures. A sequence of probability measures $(\mu_k)_{k \in \mathbb{N}} \in P(\mathcal{X})$ is said to converge weakly to $\mu \in P(\mathcal{X})$ (denoted by $\mu_k \rightarrow \mu$) if, for every bounded continuous function $\varphi : \mathcal{X} \rightarrow \mathbb{R}$,

$$\int \varphi d\mu_k \rightarrow \int \varphi d\mu.$$

The topology induced by the weak convergence on $M(\mathcal{X})$, the space of probability measures on \mathcal{X} , is called the weak topology.

The following theorem gives useful equivalent definitions of weak convergence:

Theorem 1.2.17 (Portmanteau Theorem). *Let (\mathcal{X}, d) be a metric space and $P(\mathcal{X})$ be its space of Borel probability measures. Then, these five conditions are equivalent to the definition 1.2.16 of weak convergence:*

$$(i) \int \varphi d\mu_k \rightarrow \int \varphi d\mu \text{ for any bounded, real function } \varphi, \text{ continuous } \mu\text{-a.e.};$$

$$(ii) \int \varphi d\mu_k \rightarrow \int \varphi d\mu \text{ for any bounded, uniformly continuous function } \varphi;$$

$$(iii) \limsup_{k \rightarrow \infty} \mu_k(F) \leq \mu(F) \text{ for all closed } F;$$

$$(iv) \mu(G) \leq \liminf_{k \rightarrow \infty} \mu_k(G) \text{ for all open } G;$$

$$(v) \mu_k(A) \rightarrow \mu(A) \text{ for all Borel sets } A \text{ for which } \mu(\partial A) = 0, \text{ where } \partial A := \bar{A} \cap \bar{A}^c.$$

Proof: For a proof of Theorem 1.2.17, one can consult Billingsley ([10], Theorem 2.1). ■

We recall the definition of *tightness* for measures and family of measures (see for example Billingsley [10]):

Definition 1.2.18. *Let (\mathcal{X}, d) be a metric space and $P(\mathcal{X})$ be its space of Borel probability measures.*

A probability measure $\mu \in P(\mathcal{X})$ is tight if, for any $\epsilon > 0$, there is a compact set K_ϵ such that $\mu(\mathcal{X} \setminus K_\epsilon) \leq \epsilon$.

A subset $S \subset P(\mathcal{X})$ of probability measures is tight if, for any $\epsilon > 0$, there is a compact set $K_\epsilon \subset \mathcal{X}$ such that $\mu(\mathcal{X} \setminus K_\epsilon) \leq \epsilon$, for all $\mu \in S$.

Theorem 1.2.19. *Let \mathcal{X} be a Polish space and $P(\mathcal{X})$ be its space of Borel probability measures.*

Then, any probability measure $\mu \in P(\mathcal{X})$ is tight.

Proof: (Theorem 1.2.19)

For a proof of Theorem 1.2.19, one can consult Billingsley ([10], Theorem 1.3). ■

Lemma 1.2.20 (Tightness of Couplings). *Let (\mathcal{X}, d_X) and (\mathcal{Y}, d_Y) be two metric spaces. Let S_x and S_y be tight subsets of $P(\mathcal{X})$ and $P(\mathcal{Y})$, respectively. Then the set $\mathcal{N}(S_x, S_y)$ of all couplings whose marginals lie in S_x and S_y respectively, is itself tight in $P(\mathcal{X} \times \mathcal{Y})$.*

Proof: (Lemma 1.2.20)

Let $\mu \in S_x, \nu \in S_y$. Since S_x is tight, for any $\epsilon > 0$, there is a compact set $K_\epsilon \subset \mathcal{X}$, independent of the choice of μ in S_x , such that $\mu(\mathcal{X} \setminus K_\epsilon) \leq \epsilon$. Similarly, since S_y is tight, there is a compact set $L_\epsilon \subset \mathcal{Y}$, independent of the choice of ν in S_y , such that $\nu(\mathcal{Y} \setminus L_\epsilon) \leq \epsilon$. Then for any coupling $\vartheta \in \mathcal{N}(\mu, \nu)$, one have

$$\mathcal{N}[(\mathcal{X} \times \mathcal{Y}) \setminus (K_\epsilon \times L_\epsilon)] \leq \mathcal{N}[(\mathcal{X} \setminus K_\epsilon) \times \mathcal{Y}] + \mathcal{N}[\mathcal{X} \times (\mathcal{Y} \setminus L_\epsilon)] \mu[\mathcal{X} \setminus K_\epsilon] + \nu[\mathcal{Y} \setminus L_\epsilon] \leq 2\epsilon$$

Now, since this bound does not depend on the choice of the coupling ϑ and since $K_\epsilon \times L_\epsilon$ is compact in $\mathcal{X} \times \mathcal{Y}$, the proof is complete. ■

Theorem 1.2.21 (Prokhorov Theorem). *Let (\mathcal{X}, d) be a metric space and $P(\mathcal{X})$ be its space of Borel probability measures.*

If $S \subset P(\mathcal{X})$ is tight, then S is relatively compact (ie. \bar{S} is compact in $P(\mathcal{X})$).

Let \mathcal{X} be a Polish space and $P(\mathcal{X})$ be its space of Borel probability measures.

For any subset $S \subseteq P(\mathcal{X})$, the following two statements are equivalent:

(i) \bar{S} is relatively compact

(ii) S is tight.

Proof: For a proof of Theorem 1.2.21, one can consult Billingsley ([10], Theorem 5.1 and Theorem 5.2). ■

We can now prove the existence of an optimal coupling (Theorem 1.2.15).

Proof: (Theorem 1.2.15 - part (i))

Since \mathcal{X} and \mathcal{Y} are Polish spaces, both $\{\mu\} \subset P(\mathcal{X})$ and $\{\nu\} \subset P(\mathcal{Y})$ are tight. By Lemma 1.2.20, $\mathcal{N}(\mu, \nu)$ is a tight subset of $P(\mathcal{X} \times \mathcal{Y})$ and by Prokhorov's theorem (1.2.21), it is relatively compact for the weak topology.

Let I^* denote the infimum of $I(\vartheta) = \int_{\mathcal{X} \times \mathcal{Y}} c \, d\vartheta$, for $\vartheta \in \mathcal{N}(\mu, \nu)$, and let $(\vartheta_k)_{k \geq 1}$ be a minimizing sequence for I . Let us show that if ϑ^* is any weak unit point of $\vartheta \in \mathcal{N}(\mu, \nu)$, then $I(\vartheta^*) = I^*$. Since c is a nonnegative lower semi-continuous function, it is the pointwise limit of an increasing sequence $(c_n)_{n \geq 1}$ of continuous bounded functions. Then, by the monotone convergence theorem, we have for all $k \geq 1$,

$$\lim_{n \rightarrow \infty} \int c_n \, d\vartheta_k = \int c \, d\vartheta_k \quad \text{and} \quad \lim_{n \rightarrow \infty} \int c_n \, d\vartheta^* = \int c \, d\vartheta^*.$$

then, by definition of ϑ^* , we have

$$I(\vartheta^*) = \lim_{n \rightarrow \infty} \int c_n \, d\vartheta^* \leq \lim_{n \rightarrow \infty} \limsup_{k \rightarrow \infty} \int c_n \, d\vartheta_k \leq \limsup_{k \rightarrow \infty} \int c \, d\vartheta_k = I.$$

As $\vartheta^* \in \mathcal{N}(\mu, \nu)$, $I(\vartheta^*) = I$. ■

1.3 Kantorovich-Rubinstein Duality Theorem

When the cost function c is in fact a metric d on some Polish space \mathcal{X} , and both μ and ν are probability Borel measures on \mathcal{X} , we obtain a particular case of the Kantorovich Duality Theorem 1.2.6, the so-called Kantorovich-Rubinstein theorem. It first appeared in a paper by Kantorovich in 1942. In Kantorovich original paper, the result is proved for a compact metric space equipped with Borel probability measures.

The version of the Kantorovich-Rubinstein Theorem 1.3.2, given below is found in *Topics in Optimal Transportation* by Villani [53].

Notation 1.3.1. For clarity, we first introduce the following notations:

- (i) Let $\mathcal{D}_{\mathcal{X}}$ denote the set of all lower semi-continuous metrics d on \mathcal{X} .
- (ii) Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$. Let μ_1 and μ_2 be two Borel probability measures on \mathcal{X} . Let $I_d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be the function defined by

$$I_d(\vartheta) = \inf_{\vartheta \in \mathcal{N}(\mu_1, \mu_2)} \int_{\mathcal{X} \times \mathcal{X}} d(x_1, x_2) d\vartheta(x_1, x_2).$$

We can now state the Kantorovich-Rubinstein theorem:

Theorem 1.3.2 (Kantorovich-Rubinstein theorem). *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$. Let μ_1, μ_2 be two Borel probability measures on \mathcal{X} and I_d be defined as in Notation 1.3.1.*

Let $\text{Lip}(\mathcal{X})$ denote the space of all Lipschitz functions g on \mathcal{X} , and define

$$\|g\|_{\text{Lip}} \equiv \sup_{x \neq y} \frac{|g(x_1) - g(x_2)|}{d(x_1, x_2)}.$$

Then

$$I_d(\vartheta) = \sup \left\{ \int_{\mathcal{X}} g(x) d(\mu_1 - \mu_2)(x); g \in L_1(|\mu_1 - \mu_2|), \|g\|_{\text{Lip}} \leq 1 \right\}.$$

Moreover, it does not change the value of the supremum above to impose the additional condition that g be bounded.

Remark 1.3.3. There are two facts worthy of attention:

- (i) The metric $d \in \mathcal{D}_{\mathcal{X}}$ is not necessarily the distance defining the topology on \mathcal{X} . More often than not it is the same distance, but it need not be.
- (ii) If the distance $d \in \mathcal{D}_{\mathcal{X}}$ and the distance defining the topology on \mathcal{X} are different, the Lipschitz property for functions on \mathcal{X} is defined with respect to the distance d .

Notation 1.3.4. We introduce these two notations that we will use for the remaining of the thesis:

- (i) Let \mathcal{X} be a Polish space. Any metric in $\mathcal{D}_{\mathcal{X}}$ that defines the topology on \mathcal{X} will be denoted d^* . Hence, we denote by (\mathcal{X}, d^*) a Polish space \mathcal{X} whose topology is endowed by the metric $d^* \in \mathcal{D}_{\mathcal{X}}$.
- (ii) Let $\mathcal{D}_{\mathcal{X}}^*$ denote the space of all metrics $d^* \in \mathcal{D}_{\mathcal{X}}$ defining the topology on \mathcal{X} . Hence, $\mathcal{D}_{\mathcal{X}}^* \subset \mathcal{D}_{\mathcal{X}}$.

Corollary 1.3.5 (Kantorovich-Rubinstein theorem for atomic measures). *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$. Let us define two atomic probability measures*

$$\mu_1 = \sum_{i=1}^n \mu_i^{(1)} \delta_{x_i} \quad \text{and} \quad \mu_2 = \sum_{i=1}^m \mu_i^{(2)} \delta_{x_i},$$

supported on the same finite number of ordered points $\{x_1, \dots, x_n\} \in \mathcal{X}$.

Let I_d be given by

$$I_d[\vartheta] = \min_{\vartheta \in \mathcal{N}(\mu_1, \mu_2)} \left\{ \sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) \vartheta_{ij} \right\}$$

where

$$\mathcal{N}(\mu, \nu) = \left\{ \sum_{i=1}^n \sum_{j=1}^m \vartheta_{ij} \delta_{(x_i, y_j)} : \forall i, \sum_{j=1}^m \vartheta_{ij} = \mu_i^{(1)} \quad \text{and} \quad \forall j, \sum_{i=1}^n \vartheta_{ij} = \mu_j^{(2)} \quad \text{with} \quad \vartheta_{ij} \geq 0 \right\}.$$

Let $\text{Lip}(\mathcal{X})$ denote the space of all Lipschitz functions on \mathcal{X} , and $\|g\|_{\text{Lip}}$ be defined as in Theorem 1.3.2. Then,

$$I_d[\vartheta] = \sup \left\{ \sum_{i=1}^n g(x_i) (\mu_i^{(1)} - \mu_i^{(2)}); \|g\|_{\text{Lip}} \leq 1 \right\}.$$

Proof: (Corollary 1.3.5)

A direct application of Theorem 1.3.2 for two atomic measures μ_1 and μ_2 supported

on an ordered set $\{x_1, \dots, x_n\} \in \mathcal{X}$ yields the following dual representation formula:

$$I_d[\vartheta] = \sup \left\{ \sum_{i=1}^n g(x_i)(\mu_i^{(1)} - \mu_i^{(2)}) : \|g\|_{\text{Lip}} \leq 1 \right\}.$$

Since g is only evaluated on the finite set $\{x_1, \dots, x_n\}$, it is bounded. Thus, as shown in the proof of Lemma 1.3.6 (given below), the condition $g \in L_1(|\mu_1 - \mu_2|)$ is not necessary. ■

The following Lemma 1.3.6 is required for the proof of the Kantorovich-Rubinstein Theorem 1.3.2.

Proposition 1.3.6. *Let \mathcal{X} and \mathcal{Y} be Polish spaces and let $c \in \mathcal{D}_{\mathcal{X} \times \mathcal{Y}}$ be a bounded lower semi-continuous cost function on $\mathcal{X} \times \mathcal{Y}$. Let μ, ν be two probability measures on \mathcal{X} and \mathcal{Y} .*

Let Φ_c'' be the set of all measurable and bounded (bnd) functions $(\varphi'', \psi'') \in L_1(\mu) \times L_1(\nu)$ satisfying

$$\varphi''(x) + \psi''(y) \leq c(x, y),$$

for μ -almost all $x \in \mathcal{X}$, ν -almost all $y \in \mathcal{Y}$.

Let $J : L_1(\mu) \times L_1(\nu) \rightarrow \mathbb{R}$ be defined by

$$J(\varphi, \psi) = \int \varphi(x) d\mu(x) + \int \psi(y) d\nu(y).$$

Then,

$$\sup_{\Phi_c''} J(\varphi, \psi) = \sup_{\rho^{bnd}} J(\rho^{cc}, \rho^c),$$

$$\text{where } \rho^c(y) = \inf_{x \in \mathcal{X}} [c(x, y) - \rho(x)] \quad \text{and} \quad \rho^{cc}(x) = \inf_{y \in \mathcal{Y}} [c(x, y) - \rho^c(y)].$$

Note that, in order to be consistent with the notations in Proposition 1.2.9, one should write $(\varphi'', \psi'') \in \Phi_c''$ and not $(\varphi, \psi) \in \Phi_c''$. We have chosen to use the latter instead of the former in order to avoid writing $(\varphi'')^c$ and $(\varphi'')^{cc}$.

Proof: (Proposition 1.3.6)

Let us first note that ρ^c and similarly ρ^{cc} are measurable. As recalled in Villani [53], p.26, c is a point wise limit of an increasing sequence $(c_l)_{l \leq 1}$ of bounded uniformly continuous functions (to see this, it suffice to write $c_l(w) = \inf[c(z) + ld(w, z)]$, where the infimum is taken over all $z \in \mathcal{X} \times \mathcal{Y}$). Then,

$$\rho^c = \lim_n \rho_n^c, \text{ where } \rho_n^c(y) = \inf_{x \in \mathcal{X}} [c_n(x, y) - \rho(x)].$$

The functions ρ_n^c are uniformly continuous and therefore ρ^c is measurable. The same argument shows that ρ^{cc} is measurable. Moreover, ρ^c and ρ^{cc} are bounded since ρ and c are bounded.

By construction, $\rho^{cc}(x) + \rho^c(y) \leq c(x, y)$ and therefore $\{(\rho^{cc}, \rho^c); \rho \text{ bnd}\} \subset \Phi_c''$. Hence,

$$\sup_{\rho \text{ bnd}} J(\rho^{cc}, \rho^c) \leq \sup_{\Phi_c''} J(\varphi, \psi'')$$

To prove the converse, we first observe that if $(\varphi, \psi) \in \Phi_c''$, then $\psi(y) \leq c(x, y) - \varphi(x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Hence, $\psi \leq \varphi^c$ since $\varphi^c(y) = \inf_x [c(x, y) - \varphi(x)]$.

Likewise, for any pair of functions $(\varphi, \varphi^c) \in \Phi_c''$, we have $\varphi \leq \varphi^{cc}$ since the function $\varphi^{cc}(x) = \inf_y [c(x, y) - \varphi^c(y)]$.

As $\rho^{cc}(x) = \inf_y [c(x, y) - \rho^c(y)]$, we obtain, for any $x_0 \in \mathcal{X}$, that

$$\begin{aligned} \rho^{cc}(x_0) &= \inf_{y \in \mathcal{Y}} (c(x_0, y) - \inf_{x \in \mathcal{X}} (c(x, y) - \rho(x))) \\ &\geq \inf_y (c(x_0, y) - (c(x_0, y) - \rho(x_0))) = \rho(x_0). \end{aligned}$$

Hence,

$$J(\varphi, \psi) \leq J(\varphi, \varphi^c) \leq J(\varphi^{cc}, \varphi^c) \text{ and } \sup_{\Phi_c''} J(\varphi, \psi) \leq \sup_{\rho \text{ bnd}} J(\rho^{cc}, \rho^c).$$

■

Remark 1.3.7. If $\text{diam}(\mathcal{X}) < \infty$, then a 1-Lipschitz function g is such that g is bounded and therefore is $L^1(\mu)$. Indeed, for any $x_o \in \mathcal{X}$ and $x \in \mathcal{X}$,

$$|g(x)| \leq |g(x_o)| + |g(x) - g(x_o)| \leq |g(x_o)| + d(x, x_o) \leq |g(x_o)| + \text{diam}(\mathcal{X}).$$

Notation 1.3.8. For simplicity, we will use the following notations in the proof Theorem 1.3.2:

- (i) $E_d = \sup \{J(\rho^{dd}, \rho^d); \rho \in L_1(\mu)\}$ and $E'_d = \sup \{J(\rho^{dd}, \rho^d); \rho \in C_b(\mathcal{X})\}$.
- (ii) $L_d = \sup \{J(g, -g); g \in L_1(|\mu_1 - \mu_2|), \|g\|_{\text{Lip}} \leq 1\}$ and $L'_d = \sup \{J(h, -h); h \in C_b(\mathcal{X}), \|h\|_{\text{Lip}} \leq 1\}$.

We can now give the proof of the Kantorovich-Rubinstein Theorem 1.3.2:

Proof: (Theorem 1.3.2)

We define $d_n = d/(1+n^{-1}d)$ for $n \in \mathbb{N}$. For each n , d_n is a bounded distance satisfying $d_n \leq d$ and $d_n(x_1, x_2)$ converges monotonically to $d(x_1, x_2)$, for all (x_1, x_2) .

The proof is separated in 2 sections. For the first section, we assume that the equality $S_{d_n} = L_{d_n}$ holds for any $n \in \mathbb{N}$. Given that assumption, we show that $S_d = L_d$. The second section focuses on proving that, $S_{d_n} = L_{d_n}$ for any bounded distance d_n .

1. We need to prove that $S_d \leq L_d$. We write $d = \sup d_n$, where d_n is a nondecreasing sequence of nonnegative, uniformly continuous functions. We assume that $S_{d_n} = L_{d_n}$ holds for any $n \in \mathbb{N}$. Also, it is clear that $L_{d_n} \leq L_d$, for all $n \geq 1$ since $\Phi_{d_n} \subset \Phi_d$. Hence, we have the inequality: $S_{d_n} \leq L_d$, for all n and by the property of the supremum, we obtain $\sup_n S_{d_n} \leq L_d$.

It is left to show that $\sup_n S_{d_n} = S_d$. In the proof of the Monge-Kantorovich duality theorem 1.2.6 presented in *Topics in Optimal Transportation* [54], Villani

proves that $\sup I_{d_n} = I_d$, where

$$I_d = \inf \left\{ \int_{\mathcal{X} \times \mathcal{Y}} d(x, y) \, d\vartheta(x, y); \vartheta \in \mathcal{N}(\mu, \nu) \right\}.$$

Moreover, by Kantorovich Duality Theorem 1.2.6, we know that $I_d = S_d$. Therefore, we have $\sup_n S_{d_n} = S_d$. Since $\sup_n S_{d_n} \leq L_d$ and $\sup_n S_{d_n} = S_d$, we obtain the inequality $S_d \leq L_d$.

2. We shall now prove that $S_{d_n} = L_{d_n}$ for any bounded distance d_n . For clarity, we shall drop the index n thus writing d instead of d_n . First of all, note that, by Remark 1.3.7, if ρ is a 1-Lipschitz function and d is bounded, then ρ is bounded, and therefore is in $L_1(|\mu - \nu|)$, for any pair of probability measures μ and ν . Hence $(\rho, -\rho) \in \Phi_d$. Thus, it is clear that $L_d \leq S_d$.

To prove that $S_d \leq L_d$, it is enough to show, by Remark 1.3.7 that

$$\sup_{\Phi_d} J(\varphi, \psi) = \sup \left\{ \int_{\mathcal{X}} \rho \, d(\mu - \nu); \|\rho\|_{Lip} \leq 1 \right\} \quad (1.3.1)$$

Consider $S_d = \sup_{\Phi_d} J(\varphi, \psi)$. By Proposition 1.3.6, we know that $S_d = E_d$ where

$$E_d = \sup_{\rho \in L_1(\mu)} J(\rho^{dd}, \rho^d) \quad \text{with} \quad \rho^d(x_2) = \inf_{x_1 \in \mathcal{X}} [d(x_1, x_2) - \rho(x_1)].$$

Since ρ^d is the infimum of two 1-Lipschitz and bounded functions, ρ^d is bounded and 1-Lipschitz. Hence, $\rho^d(x_2)$ is finite for any $x_2 \in \mathcal{X}$.

Therefore, we obtain two inequalities: First, $\rho^{dd} \leq -\rho^d$. That is because $\inf_{x_2 \in \mathcal{X}} [d(x_1, x_2) - \rho^d(x_2)] \leq d(x_1, x_1) - \rho^d(x_1)$. Secondly, $-\rho^d \leq \rho^{dd}$. Indeed, since $|\rho^d(x_1) - \rho^d(x_2)| \leq d(x_1, x_2)$, we have $-\rho^d(x_1) \leq d(x_1, x_2) - \rho^d(x_2)$, for all $x_1, x_2 \in \mathcal{X}$. The infimum being the largest lower bound, we get $-\rho^d(x_1) \leq \rho^{dd}(x_1)$, for all $x_1 \in \mathcal{X}$, with $\rho^{dd}(x_1) = \inf_{x_2 \in \mathcal{X}} [d(x_1, x_2) - \rho^d(x_2)]$.

Together, these two inequalities yield that $\rho^{dd} = -\rho^d$.

Finally, one notes that, since ρ^d is bounded and 1-Lipschitz, $\{(-\rho^d, \rho^d); \rho \text{ bounded}\} \subset$

$\{(-\rho, \rho); \|\rho\|_{Lip} \leq 1\}$. Thus,

$$\sup_{\rho \in L_1(\mu)} J(-\rho^d, \rho^d) \leq L'_d \quad \text{where} \quad L'_d = \sup_{\|\rho\|_{Lip} \leq 1} J(-\rho^d, \rho^d).$$

Therefore we have, for ρ bounded,

$$S_d = S'_d = E_d = \sup_{\rho \in L_1(\mu)} J(-\rho^d, \rho^d) \leq L'_d = L_d.$$

Since $L_d \leq S_d$ and $S_d \leq L_d$, we have shown that S_d and L_d are equal. ■

1.3.1 Existence of Optimal measures and functions

Since the Kantorovich-Rubinstein theorem is a particular case of the Kantorovich Duality theorem 1.2.6, it is natural to suppose that both the infimum over all couplings is realized and that the dual formulation achieves its supremum over all 1-Lipschitz functions.

Definition 1.3.9. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$. Let μ_1, μ_2 be two probability measures on \mathcal{X} .*

(i) *By an optimal measure on $\mathcal{X} \times \mathcal{X}$ we shall understand a measure $\rho \in \mathcal{N}(\mu_1, \mu_2)$ such that*

$$\int_{\mathcal{X}} d(x_1, x_2) d\rho(x_1, x_2) = \inf_{\vartheta \in \mathcal{N}(\mu_1, \mu_2)} \int_{\mathcal{X} \times \mathcal{X}} d(x_1, x_2) d\vartheta(x_1, x_2).$$

(ii) *By an optimal function on \mathcal{X} , we shall understand a 1-Lipschitz function f , $f \in L_1(|\mu_1 - \mu_2|)$ such that*

$$\int_{\mathcal{X}} f(x) d(\mu_1 - \mu_2)(x) = \sup \left\{ \int_{\mathcal{X}} g d(\mu_1 - \mu_2); g \in L_1(|\mu_1 - \mu_2|), \|g\|_{Lip} \leq 1 \right\}.$$

Remark 1.3.10. If f is an optimal function, then $f + c$, for all $c \in \mathbb{R}$ is also an optimal function. Indeed, as any constant function is 1-Lipschitz, and belongs to $L_1(|\mu_1 - \mu_2|)$, we only have to note that

$$\begin{aligned} \int_{\mathcal{X}} (f + c)(x) d(\mu_1 - \mu_2)(x) &= \int_{\mathcal{X}} f(x) d(\mu_1 - \mu_2)(x) + c \int_{\mathcal{X}} d(\mu_1 - \mu_2)(x) \\ &= \int_{\mathcal{X}} f(x) d(\mu_1 - \mu_2)(x). \end{aligned}$$

Theorem 1.3.11 (Existence of optimal solutions). *Let \mathcal{X} be a Polish space and $d \in D_{\mathcal{X}}$ be a bounded lower semi-continuous metric on \mathcal{X} . Let μ_1, μ_2 be two probability measures on \mathcal{X} .*

- (i) *Then, there exists an optimal measure ρ on $\mathcal{X} \times \mathcal{X}$. That is, the infimum of $I[\vartheta]$ over all $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$ is attained.*
- (ii) *Then, there exists an optimal 1-Lipschitz function f on \mathcal{X} . That is, the supremum of $J(g, -g)$ over all 1-Lipschitz functions g is attained by a 1-Lipschitz function.*

To write the proof of Theorem 1.3.11, we first need Proposition 1.3.14 and Proposition 1.3.15. We start by recalling the definitions and basic results of c -concavity seen in subsection 1.2.2:

Definition 1.3.12. [d -concavity] *Let (\mathcal{X}, d) be a metric space. A function $\rho : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ is said to be d -concave if there exists $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$, $\psi \not\equiv -\infty$, such that $\rho(x) = \psi^d(x)$, where*

$$\psi^d(x_1) = \inf_{x_2 \in \mathcal{X}} [d(x_1, x_2) - \psi(x_2)].$$

That is ρ is d -concave if there exists ψ such that ρ is the d -transform of ψ .

Proposition 1.3.13 (Alternative characterization of d -concavity). *Let (\mathcal{X}, d) be a metric space. Consider a function $\rho : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$, $\rho \not\equiv -\infty$.*

Then, ρ is d -concave if and only if $\rho^{dd} = \rho$, where $\rho^{dd} = (\rho^d)^d$.

We can now give propositions 1.3.14 and 1.3.15:

Proposition 1.3.14. *Let (\mathcal{X}, d) be a metric space and let $\rho : \mathcal{X} \rightarrow \mathbb{R}$ a real function defined on \mathcal{X} .*

The function ρ is d -concave if and only if

$$\rho(x_2) - \rho(x_1) \leq d(x_1, x_2), \forall x_1, x_2 \in \mathcal{X}.$$

Proof: (Proposition 1.3.14)

Suppose that $\rho(x_2) - \rho(x_1) \leq d(x_1, x_2), \forall x_1, x_2 \in \mathcal{X}$. To show that ρ is d -concave, one needs to find ψ such that

$$\rho(x_2) = \inf_{x_1 \in \mathcal{X}} (d(x_1, x_2) - \psi(x_1)).$$

Let $\psi(x_1) = -\rho(x_1)$. We have $\rho(x_2) + \psi(x_1) \leq d(x_1, x_2)$ and thus

$\rho(x_2) \leq d(x_1, x_2) - \psi(x_1), \forall x_1, x_2 \in \mathcal{X}$. Since the infimum is the greatest lower bound,

$$\rho(x_2) \leq \inf_{x_1 \in \mathcal{X}} (d(x_1, x_2) - \psi(x_1)), \forall x_2 \in \mathcal{X}.$$

Now, for $x_1 = x_2$, we have $\rho(x_2) = d(x_1, x_2) - \psi(x_1)$. Hence

$$\rho(x_2) \geq \inf_{x_1 \in \mathcal{X}} (d(x_1, x_2) - \psi(x_1)).$$

Thus, $\rho(x_2) = \inf_{x_1 \in \mathcal{X}} (d(x_1, x_2) - (-\rho(x_1)))$, for all $x_2 \in \mathcal{X}$, and therefore ρ is d -concave.

Conversely, suppose that ρ is d -concave. Hence, there exists ψ such that, for all $x_2 \in \mathcal{X}$,

$$\begin{aligned} \rho(x_2) &= \inf_{x_3 \in \mathcal{X}} (d(x_3, x_2) - \psi(x_3)) \\ &= \inf_{x_3 \in \mathcal{X}} (d(x_3, x_1) - \psi(x_3) - d(x_3, x_1) + d(x_3, x_2)). \end{aligned}$$

The triangle inequality gives $d(x_3, x_2) - d(x_3, x_1) \leq d(x_1, x_2)$, hence

$$\rho(x_2) \leq \inf_{x_3 \in \mathcal{X}} (d(x_3, x_1) - \psi(x_3) + d(x_1, x_2)) \text{ for all } x_2 \in \mathcal{X}.$$

For $x_3 = x_1$, we obtain $\rho(x_2) \leq d(x_1, x_2) - \psi(x_1)$. As $\rho(x_1) = \inf_{x_3 \in \mathcal{X}} (d(x_3, x_1) - \psi(x_3))$, we get $\rho(x_1) \geq d(x_1, x_1) - \psi(x_1) = -\psi(x_1)$. Hence $\rho(x_2) - \rho(x_1) \leq d(x_1, x_2)$. ■

Proposition 1.3.15. *Let (\mathcal{X}, d) and ρ be defined as in Proposition 1.3.14.*

Then, the function ρ is d -concave if and only if ρ is 1-Lipschitz. Also, $-\rho$ is the d -transform of ρ .

Proof: (Proposition 1.3.15)

If ρ is d -concave, then, by Proposition 1.3.14, $\rho(x_2) - \rho(x_1) \leq d(x_1, x_2)$, $\forall x_1, x_2 \in \mathcal{X}$. As d is symmetric, we get $|\rho(x_2) - \rho(x_1)| \leq d(x_1, x_2)$, $\forall x_1, x_2 \in \mathcal{X}$. Hence ρ is 1-Lipschitz.

Similarly, if ρ is 1-Lipschitz, then $\rho(x_1) - \rho(x_2) \leq d(x_1, x_2)$, $\forall x_1, x_2 \in \mathcal{X}$, and by Proposition 1.3.14, ρ is d -concave.

Regarding the d -transform, we need to show that $\rho^d \leq \rho$ and $\rho \leq \rho^d$. Recall that ρ^d of ρ is given by

$$\rho^d(x_2) = \inf_{x_1 \in \mathcal{X}} [d(x_1, x_2) - \rho(x_1)].$$

For $x_1 = x_2$, we have $\rho^d(x_2) \leq -\rho(x_2)$.

Now, since ρ is 1-Lipschitz, $\inf_{x_1 \in \mathcal{X}} [-(\rho(x_2) - \rho(x_1)) - \rho(x_1)] \leq \rho^d(x_2)$. That is $\rho \leq \rho^d$. ■

We can now give the proof of Theorem 1.3.11:

Proof: (Theorem 1.3.11)

- (i) Since d is a nonnegative lower semi-continuous metric, a direct application of Theorem 1.2.15 (Kantorovich Duality Theorem - Part 2.) yields the conclusion.
- (ii) Since d is bounded, the conditions the Kantorovich Duality Theorem 1.2.15 are clearly satisfied. Therefore, we know that the supremum of $J(\varphi, \psi)$ over all $(\varphi, \psi) \in \Phi_d$ is realized by a pair of conjugate d -concave functions (ρ, ρ^d) . By Proposition 1.3.15, a function is d -concave if and only if it is 1-Lipschitz. Moreover, $\rho^d = -\rho$. Hence, the supremum of $J(\varphi, \psi)$ over all $(\varphi, \psi) \in \Phi_d$ is equivalent to the supremum over all 1-Lipschitz functions ρ .

■

Remark 1.3.16. More general versions of Theorem 1.3.11 have been proved. For example, let (\mathcal{X}, d) be a Polish space and μ_1, μ_2 be two probability measures in $P_{\mathcal{X}}$ (ie. for any fixed $x_o \in \mathcal{X}$, $d(\cdot, x_o) \in L_1(\mu_i)$), then there exists a function $f : \mathcal{X} \rightarrow \mathbb{R}$, 1-Lipschitz and in $L_1(|\mu_1 - \mu_2|)$ which is optimal.

We end this section with the following important characterization of an optimal function (see, for example, [23], Theorem 8.1):

Proposition 1.3.17 (Kantorovich Optimality Criterion). *Let (\mathcal{X}, d) be a metric space whose diameter is bounded and μ_1, μ_2 be two probability measures on \mathcal{X} . For $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$ and $f : \mathcal{X} \rightarrow \mathbb{R}$ a 1-Lipschitz function, the following statements are equivalent:*

- (i) ϑ is an optimal measure and f is an optimal function.
- (ii) $d(x_1, x_2) = f(x_1) - f(x_2), \quad \forall (x_1, x_2) \in \text{supp } \vartheta$.

Proof: (Proposition 1.3.17)

First of all, notice that, as d is bounded, $I(\vartheta) = \int_{\mathcal{X} \times \mathcal{X}} d(x, y) d\vartheta(x, y) < \infty$ and that

$f \in L_1(\mu_i)$ for $i = 1, 2$. Hence,

$$f \in L_1(|\mu_1 - \mu_2|), \text{ as } \int_{\mathcal{X}} |f(x)| d|\mu_1 - \mu_2|(x) \leq \int |f(x)|(d\mu_1(x) + d\mu_2(x)) < \infty.$$

Suppose that ρ is an optimal measure on $\mathcal{X} \times \mathcal{X}$ and f is an optimal function on \mathcal{X} . Then,

$$\begin{aligned} \mathcal{T}_d(\mu_1, \mu_2) &= \int_{\mathcal{X} \times \mathcal{X}} d(x_1, x_2) d\rho(x_1, x_2) && \text{by the optimality of } \rho. \\ &\geq \int_{\mathcal{X} \times \mathcal{X}} |f(x_1) - f(x_2)| d\rho(x_1, x_2) && \text{since } f \text{ is 1-Lipschitz.} \\ &\geq \int_{\mathcal{X} \times \mathcal{X}} (f(x_1) - f(x_2)) d\rho(x_1, x_2) \\ &= \int_{\mathcal{X}} f(x_1) d\rho(x_1, x_2) - \int_{\mathcal{X}} f(x_2) d\rho(x_1, x_2) \\ &= \int_{\mathcal{X}} f(x_1) d\mu_1(x_1) - \int_{\mathcal{X}} f(x_2) d\mu_2(x_2) && \text{since } \vartheta \in \mathcal{N}(\mu_1, \mu_2), \\ &= \int_{\mathcal{X}} f(x) d(\mu_1 - \mu_2)(x) \\ &= \mathcal{T}_d(\mu_1, \mu_2) && \text{since } f \text{ is optimal.} \end{aligned}$$

Therefore we obtain:

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{X}} d(x_1, x_2) d\rho(x_1, x_2) &= \int_{\mathcal{X} \times \mathcal{X}} (f(x_1) - f(x_2)) d\rho(x_1, x_2), \\ \text{hence } \int_{\mathcal{X}} \left(d(x_1, x_2) - (f(x_1) - f(x_2)) \right) d\rho(x_1, x_2) &= 0. \end{aligned}$$

Since f is a 1-Lipschitz function, $d(x_1, x_2) - (f(x_1) - f(x_2)) \geq 0$. Thus $d(x_1, x_2) = f(x_1) - f(x_2)$ for all (x_1, x_2) , ρ -a.e. In fact, $d(x_1, x_2) = f(x_1) - f(x_2)$ for all $(x_1, x_2) \in \text{supp } \rho$. Indeed suppose there exists $(x_1^*, x_2^*) \in \text{supp } \rho$ such that $g(x_1^*, x_2^*) = d(x_1, x_2) - (f(x_1) - f(x_2)) > 0$. Then, there exists an open neighbourhood \mathcal{U} of (x_1^*, x_2^*) such that

$g(x_1, x_2) > g(x_1^*, x_2^*)/2, \forall (x_1, x_2) \in \mathcal{U}$. Thus,

$$\int_{\mathcal{U}} g(x_1, x_2) d\rho(x_1, x_2) > \frac{g(x_1^*, x_2^*)}{2} \rho(\mathcal{U}) > 0,$$

which is a contradiction.

Conversely, suppose that f is a 1-Lipschitz function such that, for all $(x_1, x_2) \in \text{supp } \rho, d(x_1, x_2) = f(x_1) - f(x_2)$. Then, for $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$,

$$\begin{aligned} \mathcal{T}_d(\mu_1, \mu_2) &\leq \int_{\mathcal{X} \times \mathcal{X}} d(x_1, x_2) d\vartheta(x_1, x_2) && \text{by definition of } \mathcal{T}_d, \\ &= \int_{\mathcal{X} \times \mathcal{X}} (f(x_1) - f(x_2)) d\vartheta(x_1, x_2) \\ &= \int_{\mathcal{X} \times \mathcal{X}} f(x_1) d\vartheta(x_1, x_2) - \int_{\mathcal{X} \times \mathcal{X}} f(x_2) d\vartheta(x_1, x_2) \\ &= \int_{\mathcal{X} \times \mathcal{X}} f(x_1) d\mu_1(x_1) - \int_{\mathcal{X} \times \mathcal{X}} f(x_2) d\mu_2(x_2) \\ &= \int_{\mathcal{X}} f(x) d(\mu_1 - \mu_2)(x) \\ &\leq \mathcal{T}_d(\mu_1, \mu_2) && \text{by the K-R Theorem 1.3.2.} \end{aligned}$$

Hence we obtain:

$$\mathcal{T}_d(\mu_1, \mu_2) = \int_{\mathcal{X} \times \mathcal{X}} d(x_1, x_2) d\vartheta(x_1, x_2) \quad \text{and} \quad \mathcal{T}_d(\mu_1, \mu_2) = \int_{\mathcal{X}} f(x) d(\mu_1 - \mu_2)(x).$$

Therefore, ϑ is an optimal measure on $\mathcal{X} \times \mathcal{X}$ and f is an optimal function on \mathcal{X} . ■

Remark 1.3.18. Proposition 1.3.17 is still valid if d is not a bounded distance, but if $\mu_1, \mu_2 \in P_{\mathcal{X}}$, ie. for some $x_0 \in \mathcal{X}$ (and therefore for any x_0), $\int d(x, x_0) d\mu_i(x) < \infty$.

Indeed, keeping the notation of 1.3.17, we have in this case:

$$\int d(x, y) d\vartheta(x, y) \leq \int d(x, x_0) d\mu_1(x) + \int d(x_0, y) d\mu_2(x) < \infty$$

Since f is 1-Lipschitz, f therefore belongs to $L_1(\mu_i)$ as

$$\int |f(x)| d\mu_i(x) \leq \int |f(x) - f(x_0)| d\mu_i(x) + |f(x_0)| \leq \int d(x, x_0) d\mu_i(x) + |f(x_0)| < \infty.$$

Then, $f \in L_1(|\mu_1 - \mu_2|)$.

Chapter 2

The Kantorovich-Rubinstein Distance

In the previous chapter, we introduced the functional I_c defined by

$$I_c(\vartheta) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\vartheta(x, y),$$

where c is the cost function. When the cost function is defined in terms of a distance on \mathcal{X} , it allows to define a family of distances between the measures μ and ν , called the Wasserstein distances. We briefly introduce the Wasserstein distances and their convergence and topological properties before focusing on a particular case called the Kantorovich-Rubinstein distance. Using the Kantorovich-Rubinstein theorem 1.3.2, it is possible for some particular pairs of Polish spaces and distances on that space to construct explicit formulas for the Kantorovich-Rubinstein distance. In this chapter, we construct the explicit formula for the Kantorovich-Rubinstein distance for three different pairs of Polish space and distance: any Polish space equipped with the discrete distance, the real line equipped with the Euclidean distance and the circle \mathbb{S}_1 .

2.1 The Wasserstein Distance

Definition 2.1.1 (Wasserstein space). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and $P(\mathcal{X})$ be the space of Borel probability measures on \mathcal{X} . The Wasserstein space of order p ($p > 1$) is defined as*

$$P_p(\mathcal{X}) := \left\{ \mu \in P(\mathcal{X}); \int_{\mathcal{X}} d(x_0, x)^p d\mu < +\infty \right\}, \quad (2.1.1)$$

where $x_0 \in \mathcal{X}$ is arbitrary. This space does not depend on the choice of the point x_0 .

Definition 2.1.2. *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and let $P_p(\mathcal{X})$ be the Wasserstein space associated to \mathcal{X} . For any two probability measures $\mu_1, \mu_2 \in P_p(\mathcal{X})$, the Wasserstein function of order p ($p > 1$) between μ_1 and μ_2 is defined by the formula*

$$W_p(\mu_1, \mu_2) = \left(\inf_{\nu \in \mathcal{N}(\mu_1, \mu_2)} \int_{\mathcal{X}} d(x, y)^p d\nu(x, y) \right)^{\frac{1}{p}}.$$

In probability theory, one would write:

$$W_p(\mu_1, \mu_2) = \inf \left\{ [\mathbb{E} d(X, Y)^p]^{\frac{1}{p}}, \text{law}(X) = \mu_1, \text{law}(Y) = \mu_2 \right\}.$$

where the infimum is taken over all possible coupling (X, Y) of μ and ν .

Proposition 2.1.3 (Wasserstein distance). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and $P(\mathcal{X})$ be the space of Borel probability measures on \mathcal{X} . Then, the Wasserstein function W_p defines a finite distance on the Wasserstein space $P_p(\mathcal{X})$. The distance W_p is called the Wasserstein distance.*

Proof: (Proposition 2.1.3)

Let $\mu_1, \mu_2 \in P_p(\mathcal{X})$ and $\nu \in \mathcal{N}(\mu_1, \mu_2)$. We need to show that W_p satisfies the axioms of a distance and that W_p is finite on $P_p(\mathcal{X})$. To prove that W_p satisfies the axioms of a distance, one can consult Villani ([54], p.106). We give the proof that W_p is finite on $P_p(\mathcal{X})$.

For $p \in [1, \infty)$, the function $f(x) = x^p$ is convex. By the definition of convexity, the inequality $(tx_1 + (1-t)x_2)^t \leq tx_1^p + (1-t)x_2^p$ holds for $x_1, x_2 \in \mathbb{R}$ and $t \in [0, 1]$. Thus, for $x_0 \in \mathbb{R}$, and $t = 1/2$, we obtain:

$$d(x, y)^p \leq 2^{p-1}[d(x, x_0)^p + d(x_0, y)^p].$$

Since $\mu_1, \mu_2 \in P_p(\mathcal{X})$, we therefore have

$$\int_{\mathcal{X}} d(x, y)^p d\nu(x, y) \leq 2^{p-1} \int_{\mathcal{X}} d(x, x_0)^p d\mu_1(x) + \int_{\mathcal{X}} d(x_0, y)^p d\mu_2(y) < \infty,$$

which completes the proof. ■

2.1.1 Convergence in Wasserstein sense

Now we shall define a type of convergence on the Wasserstein space.

Recall from Definition 1.2.16 that, for Borel probability measures, $\mu_k \rightarrow \mu$ means that μ_k converges weakly to μ , i.e. $\int \varphi d\mu_k \rightarrow \int \varphi d\mu$ for any bounded continuous φ on \mathcal{X} .

Definition 2.1.4. [*Weak Convergence on $P_p(\mathcal{X})$*] Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and $P_p(\mathcal{X})$ its Wasserstein space of order p . A sequence of probability measures $(\mu_k)_{k \in \mathbb{N}} \in P_p(\mathcal{X})$, is said to converge weakly to $\mu \in P_p(\mathcal{X})$ (denoted by $\mu_k \Rightarrow \mu$) if

$$\mu_k \longrightarrow \mu \quad \text{and} \quad \int d(x_0, x)^p d\mu_k \rightarrow \int d(x_0, x)^p d\mu,$$

for some (and then any) $x_0 \in \mathcal{X}$.

Theorem 2.1.5. Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and $p \in [1, \infty)$. Let $(\mu_k)_{k \in \mathbb{N}}$ be a sequence of probability measures in $P_p(\mathcal{X})$ and let μ be another element of $P_p(\mathcal{X})$. Then, for any $x_0 \in \mathcal{X}$, the following three statements are equivalent to the definition of weak convergence 2.1.4:

$$(i) \mu_k \longrightarrow \mu \text{ and } \limsup_{k \rightarrow \infty} \int d(x_0, x)^p d\mu_k(x) \leq \int d(x_0, x)^p d\mu(x);$$

$$(ii) \mu_k \longrightarrow \mu \text{ and } \lim_{R \rightarrow \infty} \limsup_{k \rightarrow \infty} \int_{d(x_0, x) \geq R} d(x_0, x)^p d\mu_k(x) = 0;$$

(iii) For all continuous functions φ with $|\varphi(x)| \leq C(1 + d(x_0, x)^p)$, $C \in \mathbb{R}$, one has

$$\int \varphi(x) d\mu_k(x) \rightarrow \int \varphi(x) d\mu(x).$$

Theorem 2.1.6 (W_p metrizes P_p). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and $p \in [1, \infty)$. Then the Wasserstein distance W_p metrizes the weak convergence in $P_p(\mathcal{X})$. In other words, if $(\mu_k)_{k \in \mathbb{N}}$ is a sequence of measures in $P_p(\mathcal{X})$ and μ is another measure in $P(\mathcal{X})$, then the following two statements are equivalent:*

(i) μ_k converges weakly in $P_p(\mathcal{X})$ to μ ;

(ii) $W_p(\mu_k, \mu) \rightarrow 0$.

Proof: For a proof of the W_p metrizes P_p theorem, one can consult Villani ([54], p.113). ■

Corollary 2.1.7 (Continuity of W_p). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and $p \in [1, \infty)$. Then the Wasserstein distance W_p is continuous on $P_p(\mathcal{X})$. More explicitly, if the sequence $(\mu_k)_{k \in \mathbb{N}}$ (respectively $(\nu_k)_{k \in \mathbb{N}}$) converges weakly to μ (respectively ν) in $P_p(\mathcal{X})$ then $W_p(\mu_k, \nu_k) \rightarrow W_p(\mu, \nu)$.*

2.1.2 Topological properties of the Wasserstein space

The Wasserstein space $P_p(\mathcal{X})$ inherits several properties of the base space \mathcal{X} . Here is a first illustration:

Theorem 2.1.8 (Topology of the Wasserstein space). *Let \mathcal{X} be a Polish space and $p \in [1, \infty)$. Then the Wasserstein space $P_p(\mathcal{X})$, equipped with the Wasserstein distance*

W_p , is also a Polish space. In short, the Wasserstein space over a Polish space is itself a Polish space. Moreover, any probability measure can be approximated by a sequence of probability measures with finite support.

Remark 2.1.9. If \mathcal{X} is compact, then $P_p(\mathcal{X})$ is also compact.

Proof: (Theorem 2.1.8)

For a proof of the *Topology of the Wasserstein space* theorem, one can consult Villani ([54], p.117). ■

2.2 The Kantorovich-Rubinstein distance

For the particular case with $p = 1$, the Wasserstein distance is commonly called the Kantorovich-Rubinstein distance. In this thesis, we will use the Kantorovich-Rubinstein distance extensively. To make it easier to read, the notation $W_{\mathcal{X}}$ will be used instead of W_1 for the Kantorovich-Rubinstein distance while the notation $P_{\mathcal{X}}$ will be used instead of $P_1(\mathcal{X})$ for the Kantorovich-Rubinstein space. Recall that $P(\mathcal{X})$ represents the set of all Borel measures on \mathcal{X} . We obtain the following definitions for the Kantorovich-Rubinstein distance and space:

Definition 2.2.1 (Kantorovich-Rubinstein space). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and $P(\mathcal{X})$ be the space of Borel probability measures on \mathcal{X} . The Kantorovich-Rubinstein space is defined as*

$$P_{\mathcal{X}} = \left\{ \mu \in P(\mathcal{X}); \int_{\mathcal{X}} d(x_0, x) \, d\mu(x) < +\infty \right\},$$

where $x_0 \in \mathcal{X}$ is arbitrary. This space does not depend on the choice of the point x_0 .

Definition 2.2.2 (Kantorovich-Rubinstein Distance). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$ and let $P_{\mathcal{X}}$ be the Kantorovich-Rubinstein space associated to \mathcal{X} . For any*

two probability measures $\mu_1, \mu_2 \in P_{\mathcal{X}}$, the Kantorovich-Rubinstein distance between μ_1 and μ_2 is defined by the formula

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \inf_{\vartheta \in \mathcal{N}(\mu_1, \mu_2)} \int_{\mathcal{X}} d(x, y) \, d\vartheta(x, y),$$

where $\mathcal{N}(\mu_1, \mu_2)$ denotes the set of couplings of μ_1 and μ_2 (see Definition 1.1.3).

Remark 2.2.3. There are two remarks worth making:

- (i) The Kantorovich-Rubinstein theorem leads to the following useful duality formula for the Kantorovich-Rubinstein distance:

For any $\mu_1, \mu_2 \in P_{\mathcal{X}}$, we have

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \sup \left\{ \int_{\mathcal{X}} \psi \, d(\mu_1 - \mu_2); \psi \in L_1(|\mu_1 - \mu_2|), \psi \text{ 1-Lipschitz} \right\}. \quad (2.2.1)$$

- (ii) Using the classical probability definition of the Monge-Kantorovich minimization problem, we can write the Kantorovich-Rubinstein distance as follow:

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \inf \{ \mathbb{E} d(X, Y), \text{law}(X) = \mu_1, \text{law}(Y) = \mu_2 \}.$$

where the infimum is taken over all possible coupling (X, Y) of μ and ν .

- (iii) Using the same notations as in Definition 2.2.1, we define

$$M_{\mathcal{X}} = \{ \mu \in M_+(\mathcal{X}); \int_{\mathcal{X}} d(x_0, x) \, d\mu(x) < +\infty \},$$

where $M_+(\mathcal{X})$ is the space of finite positive Borel measures on \mathcal{X} .

The Definition 2.2.2 of $W_{\mathcal{X}}$ can be generalized to the case of $\mu, \nu \in M_{\mathcal{X}}$ such that $\mu(\mathcal{X}) = \nu(\mathcal{X})$, since, as mentioned in Remark 1.1.5, the definition of a coupling is generalizable to two bounded measures.

An important property of the Kantorovich-Rubinstein distance is its invariance under mass subtraction. This property appears in Corollary 1.16 of Villani [53] and

is given below.

Proposition 2.2.4 (Invariance of Kantorovich-Rubinstein distance under mass subtraction). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$. Let μ_1 , μ_2 and ν be three measures in $M_{\mathcal{X}}$ such that $\mu_1(\mathcal{X}) = \mu_2(\mathcal{X})$. Then,*

$$W_{\mathcal{X}}(\mu_1 + \nu, \mu_2 + \nu) = W_{\mathcal{X}}(\mu_1, \mu_2).$$

A Corollary of the invariance of the Kantorovich-Rubinstein distance under mass subtraction is presented below:

Corollary 2.2.5. *Let $(\mathcal{X}, d_{\mathcal{X}}^*)$ be a Polish space such that $\mathcal{X} = \mathcal{X}_1 \amalg \mathcal{X}_2$. Let μ_1 and μ_2 be two probability measures on \mathcal{X} such that $\mu_1(\mathcal{X}_i) = \mu_2(\mathcal{X}_i)$, for $i = 1, 2$. Define by $\mu_i^{(1)}$ and $\mu_i^{(2)}$ the respective measures of μ_1 and μ_2 supported on \mathcal{X}_i , for $i = 1, 2$. That is, $\mu_i^{(1)} = \mu_1|_{\mathcal{X}_i}$ and $\mu_i^{(2)} = \mu_2|_{\mathcal{X}_i}$. Then,*

$$|W_{\mathcal{X}}(\mu_1, \mu_2) - W_{\mathcal{X}}(\mu_1^{(1)}, \mu_1^{(2)})| \leq W_{\mathcal{X}}(\mu_2^{(1)}, \mu_2^{(2)}).$$

To prove Corollary 2.2.5, we will need the following Lemma 2.2.6:

Lemma 2.2.6. *Let $(\mathcal{X}, d_{\mathcal{X}}^*)$ be a Polish space such that $\mathcal{X} = \amalg_{i=1}^n \mathcal{X}_i$. Let μ_1 and μ_2 be two probability measures on \mathcal{X} such that $\mu_1(\mathcal{X}_i) = \mu_2(\mathcal{X}_i)$, for $i = 1, \dots, n$. Define by $\mu_i^{(1)}$ and $\mu_i^{(2)}$ the respective measures of μ_1 and μ_2 supported on \mathcal{X}_i , for $i = 1, \dots, n$. That is, $\mu_i^{(1)} = \mu_1|_{\mathcal{X}_i}$ and $\mu_i^{(2)} = \mu_2|_{\mathcal{X}_i}$. Then,*

$$W_{\mathcal{X}}(\mu_1, \mu_2) \leq \sum_{i=1}^n W_{\mathcal{X}}(\mu_i^{(1)}, \mu_i^{(2)}).$$

Proof: (Lemma 2.2.6)

The proof is done by induction. For $1 \leq k \leq n$ fixed,

$$W_{\mathcal{X}}\left(\sum_{i=1}^{k-1} \mu_i^{(1)} + \mu_k^{(1)}, \sum_{i=1}^{k-1} \mu_i^{(2)} + \mu_k^{(2)}\right)$$

$$\leq W_{\mathcal{X}}\left(\sum_{i=1}^{k-1} \mu_i^{(1)} + \mu_k^{(1)}, \sum_{i=1}^{k-1} \mu_i^{(2)} + \mu_k^{(1)}\right) + W_{\mathcal{X}}\left(\sum_{i=1}^{k-1} \mu_i^{(2)} + \mu_k^{(1)}, \sum_{i=1}^{k-1} \mu_i^{(2)} + \mu_k^{(2)}\right) \quad (2.2.2)$$

$$\leq W_{\mathcal{X}}\left(\sum_{i=1}^{k-1} \mu_i^{(1)}, \sum_{i=1}^{k-1} \mu_i^{(2)}\right) + W_{\mathcal{X}}(\mu_k^{(1)}, \mu_k^{(2)}). \quad (2.2.3)$$

Inequality (2.2.2) is obtained by the triangle inequality and inequality (2.2.3) is obtained by direct application of Corollary 2.2.4. ■

We can now write the proof of Corollary 2.2.5

Proof: (Corollary 2.2.5)

Applying Lemma 2.2.6, for $n = 2$, we obtain

$$W_{\mathcal{X}}(\mu_1, \mu_2) \leq W_{\mathcal{X}}(\mu_1^{(1)}, \mu_1^{(2)}) + W_{\mathcal{X}}(\mu_2^{(1)}, \mu_2^{(2)})$$

and thus $W_{\mathcal{X}}(\mu_1, \mu_2) - W_{\mathcal{X}}(\mu_1^{(1)}, \mu_1^{(2)}) \leq W_{\mathcal{X}}(\mu_2^{(1)}, \mu_2^{(2)})$.

It is therefore left to show that $W_{\mathcal{X}}(\mu_1^{(1)}, \mu_1^{(2)}) - W_{\mathcal{X}}(\mu_1, \mu_2) \leq W_{\mathcal{X}}(\mu_2^{(1)}, \mu_2^{(2)})$.

By Proposition 2.2.4, we obtain:

$$\begin{aligned} 2W_{\mathcal{X}}(\mu_1^{(1)}, \mu_1^{(2)}) &= W_{\mathcal{X}}(\mu_1^{(1)} + \mu_2^{(1)}, \mu_1^{(2)} + \mu_2^{(1)}) + W_{\mathcal{X}}(\mu_1^{(1)} + \mu_2^{(2)}, \mu_1^{(2)} + \mu_2^{(2)}) \\ &= W_{\mathcal{X}}(\mu^{(1)}, \mu_1^{(2)} + \mu_2^{(1)}) + W_{\mathcal{X}}(\mu_1^{(1)} + \mu_2^{(2)}, \mu^{(2)}). \end{aligned}$$

Using the triangle inequality, we have the following two inequalities:

$$\begin{aligned} W_{\mathcal{X}}(\mu^{(1)}, \mu_1^{(2)} + \mu_2^{(1)}) &\leq W_{\mathcal{X}}(\mu^{(1)}, \mu^{(2)}) + W_{\mathcal{X}}(\mu_1^{(2)} + \mu_2^{(2)}, \mu_1^{(2)} + \mu_2^{(1)}) \\ W_{\mathcal{X}}(\mu_1^{(1)} + \mu_2^{(2)}, \mu^{(2)}) &\leq W_{\mathcal{X}}(\mu_1^{(1)} + \mu_2^{(2)}, \mu_1^{(1)} + \mu_2^{(1)}) + W_{\mathcal{X}}(\mu^{(1)}, \mu^{(2)}). \end{aligned}$$

By Proposition 2.2.4, we can write:

$$W_{\mathcal{X}}(\mu_1^{(2)} + \mu_2^{(2)}, \mu_1^{(2)} + \mu_2^{(1)}) = W_{\mathcal{X}}(\mu_2^{(2)}, \mu_2^{(1)})$$

and $W_{\mathcal{X}}(\mu_1^{(1)} + \mu_2^{(2)}, \mu_1^{(1)} + \mu_2^{(1)}) = W_{\mathcal{X}}(\mu_2^{(2)}, \mu_2^{(1)})$.

Therefore we obtain

$$W_{\mathcal{X}}(\mu_1^{(1)}, \mu_1^{(2)}) \leq W_{\mathcal{X}}(\mu^{(1)}, \mu^{(2)}) + W_{\mathcal{X}}(\mu_2^{(2)}, \mu_2^{(1)}).$$

■

2.2.1 The Kantorovich-Rubinstein distance between product measures

In this section, we study the additivity of the Kantorovich-Rubinstein distance for product measures.

Theorem 2.2.7. *Let \mathcal{X} and \mathcal{Y} be two Polish spaces and $d_{\mathcal{X}} \in \mathcal{D}_{\mathcal{X}}$ and $d_{\mathcal{Y}} \in \mathcal{D}_{\mathcal{Y}}$ be bounded lower semi-continuous distances. Let $\mathcal{X} \times \mathcal{Y}$ be the Polish cartesian product space, equipped with the distance d , given by*

$$d((x_1, y_1), (x_2, y_2)) = d_{\mathcal{X}}(x_1, x_2) + d_{\mathcal{Y}}(y_1, y_2), \text{ for } (x_1, y_1), (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}.$$

Let μ_1, μ_2 be two probability measures in $P_{\mathcal{X}}$ and ν_1, ν_2 two probability measures in $P_{\mathcal{Y}}$. Then, $\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_2$ are probability measures in $P_{\mathcal{X} \times \mathcal{Y}}$ and

$$W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_2) = W_{\mathcal{X}}(\mu_1, \mu_2) + W_{\mathcal{Y}}(\nu_1, \nu_2).$$

Proof: (Theorem 2.2.7)

Let $(x_o, y_o) \in \mathcal{X} \times \mathcal{Y}$ be fixed. Then, by definition of d ,

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{Y}} d((x_o, y_o), (x, y)) d(\mu_i \otimes \nu_i)(x, y) &= \int_{\mathcal{X} \times \mathcal{Y}} (d_{\mathcal{X}}(x_o, x) + d_{\mathcal{Y}}(y_o, y)) d\mu_i(x) d\nu_i(y) \\ &= \int_{\mathcal{X}} d(x_o, x) d\mu_i(x) + \int_{\mathcal{Y}} d(y_o, y) d\nu_i(y) < \infty, \end{aligned}$$

as $\mu_i \in P_{\mathcal{X}}$, $\nu_i \in P_{\mathcal{Y}}$. Hence, $\mu_i \otimes \nu_i \in P_{\mathcal{X} \times \mathcal{Y}}$, for $i = 1, 2$.

Now, let us show that $W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_2) \geq W_{\mathcal{X}}(\mu_1, \mu_2) + W_{\mathcal{Y}}(\nu_1, \nu_2)$. There exist, by the Kantorovich-Rubinstein theorem 1.3.2, two bounded measurable, 1-Lipschitz functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{Y} \rightarrow \mathbb{R}$ such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) - \frac{\epsilon}{2} < \int_{\mathcal{X}} f d(\mu_1 - \mu_2) \quad \text{and} \quad W_{\mathcal{Y}}(\nu_1, \nu_2) - \frac{\epsilon}{2} < \int_{\mathcal{Y}} g d(\nu_1 - \nu_2).$$

Denote by $k : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ the function defined by $k(x, y) = f(x) + g(y)$. The function k is bounded and 1-Lipschitz as

$$\begin{aligned} |k(x_1, y_1) - k(x_2, y_2)| &\leq |f(x_1) - f(x_2)| + |g(y_1) - g(y_2)| \\ &\leq d_{\mathcal{X}}(x_1, x_2) + d_{\mathcal{Y}}(y_1, y_2) = d((x_1, y_1), (x_2, y_2)), \quad \text{for } (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}. \end{aligned}$$

By the Kantorovich-Rubinstein theorem 1.3.2,

$$\int_{\mathcal{X} \times \mathcal{Y}} k(x, y) d(\mu_1 \otimes \nu_1 - \mu_2 \otimes \nu_2)(x, y) \leq W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_2).$$

By the definition of k , we have:

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{Y}} k(x, y) d(\mu_1 \otimes \nu_1 - \mu_2 \otimes \nu_2)(x, y) &= \int_{\mathcal{X} \times \mathcal{Y}} (f(x) + g(y)) d\mu_1(x) d\nu_1(y) - d\mu_2(x) d\nu_2(y) \\ &= \int_{\mathcal{X}} f(x)(d\mu_1 - d\mu_2)(x) + \int_{\mathcal{Y}} g(y)(d\nu_1 - d\nu_2)(y) \\ &\geq W_{\mathcal{X}}(\mu_1, \mu_2) + W_{\mathcal{Y}}(\nu_1, \nu_2) - \epsilon. \end{aligned}$$

As ϵ is arbitrary, we have $W_{\mathcal{X}}(\mu_1, \mu_2) + W_{\mathcal{Y}}(\nu_1, \nu_2) \leq W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_2)$.

For the converse inequality, we show first that $W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_1) \geq W_{\mathcal{X}}(\mu_1, \mu_2)$. For every $\epsilon > 0$, there exists $h : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, a bounded 1-Lipschitz function such that

$$W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_1) - \epsilon < \int_{\mathcal{X} \times \mathcal{Y}} h \, d(\mu_1 \otimes \nu_1 - \mu_2 \otimes \nu_1).$$

Using Fubini Theorem, we have

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{Y}} h \, d(\mu_1 \otimes \nu_1 - \mu_2 \otimes \nu_1) &= \int_{\mathcal{X} \times \mathcal{Y}} h(x, y) (d\mu_1(x) - d\mu_2(x)) \, d\nu_1(y) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} h(x, y) \, d\nu_1(y) \right) d(\mu_1(x) - \mu_2(x)). \end{aligned}$$

Set $t(x) = \int_{\mathcal{Y}} h(x, y) \, d\nu_1(y)$. Then, t is bounded and 1-Lipschitz as, for all $x_1, x_2 \in \mathcal{X}$,

$$|t(x_1) - t(x_2)| \leq \int_{\mathcal{Y}} |h(x_1, y) - h(x_2, y)| \, d\nu_1(y) \leq d((x_1, y), (x_2, y)) = d_{\mathcal{X}}(x_1, x_2).$$

Thus, by Kantorovich-Rubinstein theorem 1.3.2: $W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_1) \leq W_{\mathcal{X}}(\mu_1, \mu_2)$.

Similarly, we obtain $W_{\mathcal{X} \times \mathcal{Y}}(\mu_2 \otimes \nu_1, \mu_2 \otimes \nu_2) \leq W_{\mathcal{Y}}(\nu_1, \nu_2)$. Hence,

$$\begin{aligned} W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_2) &\leq W_{\mathcal{X} \times \mathcal{Y}}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_1) + W_{\mathcal{X} \times \mathcal{Y}}(\mu_2 \otimes \nu_1, \mu_2 \otimes \nu_2) \\ &\leq W_{\mathcal{X}}(\mu_1, \mu_2) + W_{\mathcal{Y}}(\nu_1, \nu_2). \end{aligned}$$

Thus completing the proof. ■

2.2.2 The Kantorovich-Rubinstein distance for atomic measures

In the case of atomic measures supported on a finite number of points, we can use the Kantorovich-Rubinstein Theorem 1.3.5 for atomic measures. We obtain the following definition:

Definition 2.2.8 (Kantorovich-Rubinstein distance for atomic measures). *Let \mathcal{X} be a Polish space with $d \in \mathcal{D}_{\mathcal{X}}$. Let us define two atomic probability measures*

$$\mu_1 = \sum_{i=1}^n \mu_i^{(1)} \delta_{x_i} \quad \text{and} \quad \mu_2 = \sum_{i=1}^n \mu_i^{(2)} \delta_{x_i},$$

supported on the same finite number of points $\{x_1, \dots, x_n\} \in \mathcal{X}$.

The Kantorovich-Rubinstein distance between μ_1 and μ_2 is then given by

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \sup \left\{ \sum_{i=1}^n g(x_i) (\mu_i^{(1)} - \mu_i^{(2)}) : |g(x_i) - g(x_j)| \leq d(x_i, x_j), 1 \leq i, j \leq n \right\}.$$

2.3 The Kantorovich-Rubinstein distance for a bounded distance on \mathcal{X}

Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . We will denote by $\Delta = \sup\{d(x, y); x, y \in \mathcal{X}\}$ the d -diameter of \mathcal{X} .

We state and, for completeness, give the proof of Remark 1.15 in [53]:

Theorem 2.3.1. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} . Let Δ be the d -diameter of \mathcal{X} and define \mathcal{F}_{Δ} by*

$$\mathcal{F}_{\Delta} = \{f : \mathcal{X} \rightarrow [0, \Delta]; f \text{ measurable and } 1\text{-Lipschitz}\}.$$

Then, the Kantorovich-Rubinstein distance can be written as

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \sup \left\{ \int f \, d(\mu_1 - \mu_2); f \in \mathcal{F}_{\Delta} \right\}.$$

Remark 2.3.2. There are two facts worth mentioning:

- (i) Recall from Remark 1.3.3 that the distance d on \mathcal{X} need not be the distance defining the topology on \mathcal{X} . In Theorem 2.4.5, the 1-discrete distance does not define the topology on \mathcal{X} . If it was, \mathcal{X} would not be a Polish space as it is not separable with the 1-discrete distance.
- (ii) The Lipschitz property for functions on \mathcal{X} is defined with respect to the k -discrete distance.

Before giving the proof of Theorem 2.3.1, let us state the following short result we will use often in the rest of the thesis.

Lemma 2.3.3. *Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two metric spaces and $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$ be a 1-Lipschitz map. Then, $\text{diam}(\varphi(\mathcal{X})) \leq \text{diam}(\mathcal{X})$.*

Proof: (Lemma 2.3.3)

Without loss of generality, we can assume that $\text{diam}(\mathcal{X}) < \infty$. Then,

$$\text{diam}(\varphi(\mathcal{X})) = \sup_{x_1, x_2 \in \mathcal{X}} d_{\mathcal{Y}}(\varphi(x_1), \varphi(x_2)) \leq \sup_{x_1, x_2 \in \mathcal{X}} d_{\mathcal{X}}(x_1, x_2) = \text{diam} \mathcal{X}. \quad \blacksquare$$

Proof: (Theorem 2.3.1)

Let us denote by \mathcal{F} the set of all measurable, 1-Lipschitz functions $f : \mathcal{X} \rightarrow \mathbb{R}$.

Since $\Delta < \infty$, by Lemma 2.3.3, if $f \in \mathcal{F}$, then f is bounded and therefore belongs to $L_1(|\mu_1 - \mu_2|)$.

By Theorem 1.3.2, we have

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \sup \left\{ \int f \, d(\mu_1 - \mu_2); f \in \mathcal{F} \right\}$$

To finish the proof, it is therefore enough to show that for any $f \in \mathcal{F}$, there exists $g \in \mathcal{F}_\Delta$ such that

$$\int f \, d(\mu_1 - \mu_2) = \int g \, d(\mu_1 - \mu_2).$$

Let $f \in \mathcal{F}$. Since f is 1-Lipschitz, and $\Delta < \infty$, then

$$-\infty < a = \inf(f(\mathcal{X})) \leq \sup f(\mathcal{X}) = b < \infty,$$

and $b - a \leq \Delta$. Set $g = f - a$, then $g \in \mathcal{F}_\Delta$ and as

$$\int g \, d(\mu_1 - \mu_2) = \int (f - a) \, d(\mu_1 - \mu_2) = \int f \, d(\mu_1 - \mu_2),$$

the proof is complete. ■

Keeping the assumptions and notations of Theorem 2.3.1, there is, by Theorem 1.3.11, an optimal function $g \in \mathcal{F}$ such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} g \, d(\mu_1 - \mu_2).$$

As seen in the proof of Theorem 2.3.1, the optimal function g is invariant by translation. Hence we obtain the following result:

Theorem 2.3.4. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} . If Δ denotes the d -diameter of \mathcal{X} , then there exists a function $f : \mathcal{X} \rightarrow [0, \Delta]$, measurable and 1-Lipschitz such that*

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} f \, d(\mu_1 - \mu_2).$$

Theorem 2.3.4 yields the following corollary:

Corollary 2.3.5. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-*

continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} . If Δ denotes the d -diameter of \mathcal{X} , then there exist measurable and 1-Lipschitz functions f and g from \mathcal{X} to $[0, \Delta]$ such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} f \, d(\mu_1 - \mu_2) = \int_{\mathcal{X}} g \, d(\mu_2 - \mu_1).$$

Proof: (Corollary 2.3.5)

By Theorem 2.3.4, there exists $f : \mathcal{X} \rightarrow [0, \Delta]$, measurable and 1-Lipschitz such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} f \, d(\mu_1 - \mu_2).$$

Then, let $g : \mathcal{X} \rightarrow [0, \Delta]$ be defined as $g = \Delta - f$. The function g is measurable, 1-Lipschitz and such that

$$\int_{\mathcal{X}} g \, d(\mu_2 - \mu_1) = \int_{\mathcal{X}} (\Delta - f) \, d(\mu_2 - \mu_1) = \int_{\mathcal{X}} (-f) \, d(\mu_2 - \mu_1) = \int_{\mathcal{X}} f \, d(\mu_1 - \mu_2).$$

■

Corollary 2.3.6. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} . If Δ denotes the d -diameter of \mathcal{X} , then there exist measurable and 1-Lipschitz functions f and g from \mathcal{X} to $[0, \Delta]$ such that*

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_0^{\Delta} F_2(y) - F_1(y) \, d(y) = \int_0^{\Delta} G_1(y) - G_2(y) \, d(y),$$

where F_i and G_i are the respective distribution functions of $f(\mu_i)$ and $g(\mu_i)$.

Proof: (Corollary 2.3.6)

By Theorem 2.3.4, there exist a function $f : \mathcal{X} \rightarrow [0, \Delta]$ such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} f \, d(\mu_1 - \mu_2).$$

As written on p.162 of *Measure Theory* [15], a direct application of Fubini Theorem allows to write (with λ denoting the Lebesgue measure on \mathbb{R} and $E = \{(x, y) \in \mathcal{X} \times \mathbb{R} : 0 \leq y < f(x)\}$):

$$\begin{aligned} (\mu_i \times \lambda)(E) &= \int_{\mathcal{X}} f(x) \, d\mu_i(x) \\ &= \int_0^\infty \mu_i(\{x \in \mathcal{X}; f(x) > y\}) \, dy \\ &= \int_0^\infty (1 - F_i(y)) \, dy. \end{aligned}$$

Since $f(\mathcal{X}) \in [0, \Delta]$ and $F_i(y) = 1$ for $y \geq \Delta$, the integrals of $1 - F_i(y)$ over $[0, \infty]$ and $[0, \Delta]$ are equal. Therefore, we can write:

$$\begin{aligned} W_{\mathcal{X}}(\mu_1, \mu_2) &= \int_0^\infty (1 - F_1(y)) - (1 - F_2(y)) \, dy \\ &= \int_0^\infty (F_2(y) - F_1(y)) \, dy \\ &= \int_0^\Delta (F_2(y) - F_1(y)) \, dy. \end{aligned}$$

Likewise, using Corollary 2.3.5, we have a function $g : \mathcal{X} \rightarrow [0, \Delta]$ such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} g \, d(\mu_2 - \mu_1).$$

Repeating the same steps as above, we obtain $\int_0^\Delta G_1(y) - G_2(y) \, d(y)$. ■

2.4 The Kantorovich-Rubinstein distance for the discrete distance on \mathcal{X}

In this section, we show that for two probability measures the Kantorovich-Rubinstein distance associated to the discrete distance is equal to the total variation distance of

these measures.

Let us first recall the definition of the total variation distance, induced by the total variation norm. The definition and properties of the total variation norm are in the Appendix (see A.2.4).

Definition 2.4.1. *[Total Variation distance] Let $(\mathcal{X}, \mathcal{F})$ be a probability space and \mathcal{F} be a σ -algebra of subsets of \mathcal{X} .*

For any two probability measures μ_1 and μ_2 on \mathcal{X} , the total variation distance between μ_1 and μ_2 is defined by the formula:

$$\|\mu_1 - \mu_2\|_{TV} = \frac{1}{2}|\mu_1 - \mu_2|(\mathcal{X}) = \frac{1}{2}(\mu_1 - \mu_2)_+(\mathcal{X}) + \frac{1}{2}(\mu_1 - \mu_2)_-(\mathcal{X}),$$

where $(\mu_1 - \mu_2)_+$ and $(\mu_1 - \mu_2)_-$ are the positive and negative part, respectively, of the signed measure $\mu_1 - \mu_2$.

By definition of the total variation norm of a signed measure (see Definition A.2.4) we have that for two probability measures μ_1 and μ_2 ,

$$\|\mu_1 - \mu_2\|_{TV} = \frac{1}{2}\|\mu_1 - \mu_2\|_1.$$

Therefore, by Proposition A.2.7, we obtain:

Proposition 2.4.2. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and μ_1, μ_2 be two probability measures on \mathcal{X} . Then*

$$\|\mu_1 - \mu_2\|_{TV} = \sup \left\{ \left| \int f d(\mu_1 - \mu_2) \right| ; f : \mathcal{X} \rightarrow \left[-\frac{1}{2}; \frac{1}{2}\right], \text{measurable} \right\}$$

Moreover, there exists $M \in \mathcal{B}$ such that

$$\|\mu_1 - \mu_2\|_{TV} = \int \left(\frac{1}{2}\chi_M - \frac{1}{2}\chi_{M^c} \right) d(\mu_1 - \mu_2).$$

Let us now recall that the discrete distance $1_{x \neq y}$ is such that $1_{x \neq y} = 0$ if $x = y$ and 1 otherwise. Formally, we have:

Definition 2.4.3 (discrete distance). *The discrete distance $1_{x \neq y} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is the function defined by*

$$1_{x \neq y}(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y. \end{cases}$$

To consider the Kantorovich-Rubinstein distance with respect to the discrete distance, notice the following:

Lemma 2.4.4. *Let \mathcal{X} be a Polish space. The 1-discrete distance $1_{x \neq y}$ on \mathcal{X} is lower semi-continuous.*

To prove Lemma 2.4.4, we first recall two topological results:

- (i) On a topological space \mathcal{Y} , a function $f : \mathcal{Y} \rightarrow \mathbb{R}$ is lower semi-continuous if the set $\{y \in \mathcal{Y}; f(y) \leq \alpha\}$ is closed in \mathcal{Y} , for all $\alpha \in \mathbb{R}$.
- (ii) The topological space \mathcal{Y} is separable if and only if the diagonal Δ , defined by $\Delta = \{(x, y) \in \mathcal{Y} \times \mathcal{Y}; x = y\}$, is closed in $\mathcal{Y} \times \mathcal{Y}$. See the proof in [45].

Proof: (Lemma 2.4.4)

Let $\alpha \in \mathbb{R}$ and define the set F_α by $F_\alpha = \{(x, y) \in \mathcal{X} \times \mathcal{X}; 1_{x \neq y}(x, y) \leq \alpha\}$.

If $\alpha < 0$, $F_\alpha = \emptyset$. For $0 \leq \alpha < 1$, $F_\alpha = \{(x, y) \in \mathcal{X} \times \mathcal{X}; 1_{x \neq y}(x, y) = 0\}$. Finally, for $\alpha \geq 1$, $F_\alpha = \mathcal{X} \times \mathcal{X}$.

Since F_α is closed $\forall \alpha$, $1_{x \neq y}$ is a lower semi-continuous distance. ■

Theorem 2.4.5. *Let \mathcal{X} be a Polish space and let μ_1 and μ_2 be two Borel probability measures on \mathcal{X} . If $W_{\mathcal{X}}(\mu_1, \mu_2)$ denotes the Kantorovich-Rubinstein with respect to the discrete distance $1_{x \neq y}$ on \mathcal{X} , then*

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|_{TV}.$$

Proof: (Theorem 2.4.5)

We first note that any function $f : \mathcal{X} \rightarrow [0, 1]$ is 1-Lipschitz with respect to the

discrete distance $1_{x \neq y}$. Therefore, as shown in the proof of Theorem 2.3.1,

$$\begin{aligned} W_{\mathcal{X}}(\mu_1, \mu_2) &= \sup \left\{ \int f \, d(\mu_1 - \mu_2); f : \mathcal{X} \rightarrow [0, 1] \text{ measurable} \right\} \\ &= \sup \left\{ \int f \, d(\mu_1 - \mu_2); f : \mathcal{X} \rightarrow \left[-\frac{1}{2}, \frac{1}{2}\right] \text{ measurable} \right\}. \end{aligned}$$

Using Proposition 2.4.2 we then obtain $W_{\mathcal{X}}(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|_{TV}$. ■

Remark 2.4.6. For $k > 0$, let $k_{x \neq y} = k \cdot 1_{x \neq y}$ denote the k -discrete distance on \mathcal{X} . If \mathcal{X} is a Polish space and μ_1, μ_2 are two Borel probability measures on \mathcal{X} , then the Kantorovich-Rubinstein distance (with respect to $k_{x \neq y}$) of μ_1, μ_2 is equal to $k\|\mu_1 - \mu_2\|_{TV}$.

2.4.1 The Kantorovich-Rubinstein distance of atomic measures for the discrete distance on \mathcal{X}

In the case of atomic measures supported on a finite number of points, the total variation distance can be written as an analytical and computationally friendly expression. To obtain this analytical expression for the total variation distance, we first need the following result:

By direct application of Theorem 2.4.5 and Proposition A.2.9 we obtain the theorem:

Theorem 2.4.7. *Let $(\mathcal{X}, \mathcal{F})$ be a probability space and \mathcal{F} be a σ -algebra of subsets of \mathcal{X} . Let us define two atomic probability measures on $S \in \mathcal{F}$:*

$$\mu_1 = \sum_{x \in S} \mu_x^{(1)} \delta_x \quad \text{and} \quad \mu_2 = \sum_{x \in S} \mu_x^{(2)} \delta_x,$$

Let $S_1 = \{x \in S; \mu_x^{(1)} \geq \mu_x^{(2)}\}$ and $S_2 = \{x \in S; \mu_x^{(2)} > \mu_x^{(1)}\}$ partition S . That is,

$S = S_1 \amalg S_2$. Then,

$$\|\mu_1 - \mu_2\|_{TV} = \sum_{i=1}^n |\mu_1(x_i) - \mu_2(x_i)|.$$

Remark 2.4.8. There are two remarks worth making:

- (i) It is interesting to note that, informally, the total variation distance between two probability measures can be seen as the maximum difference between the two probabilities assigned to a single event by the two distributions.
- (ii) As said above, total variation is a classical notion of distance between probability measures. There is also a classical probabilistic representation formula of the total variation:

For two given probability measure μ and ν on a measurable space \mathcal{X} , the total variation formula can be defined as

$$\|\mu - \nu\|_{TV} = 2 \inf \mathbb{P}[X \neq Y],$$

where the infimum is over all couplings (X, Y) of (μ, ν) ; this identity can be seen as a very particular case of duality for the cost function $c(x, y) = 1_{x \neq y}$. For a proof of this result, please see Lindvall ([34], Theorem 5.2).

From Theorem 2.4.7, we can deduce a very useful corollary:

Lemma 2.4.9. *Let (\mathcal{X}, d) be a Polish metric space and let $S = \{x_1, \dots, x_n\}$ be a finite set of mutually equidistant points in \mathcal{X} . Let us define two atomic probability measures*

$$\mu_1 = \sum_{i=1}^n \mu_i^{(1)} \delta_{x_i} \quad \text{and} \quad \mu_2 = \sum_{i=1}^n \mu_i^{(2)} \delta_{x_i},$$

supported on the set S . Then,

$$W_{\mathcal{X}}(\mu_1, \mu_2) = k \sum_{i=1}^n |\mu_1(x_i) - \mu_2(x_i)|,$$

where k is the distance between each points.

Proof: (Lemma 2.4.9)

Since the n points in S are mutually equidistant, it is clear that the distance d restricted to S is equal to the k -discrete distance. Then, a direct application of theorem 2.4.7 finishes the proof. ■

2.4.2 The Kantorovich-Rubinstein distance of atomic measures on the product space \mathcal{X}^n where each \mathcal{X} is equipped with the discrete distance

Definition 2.4.10. Let (\mathcal{X}_i, d_i) be m metric spaces and let $\mathcal{Y} = \mathcal{X}_1 \times \dots \times \mathcal{X}_m$ be the Cartesian product of these m metric spaces. For $p \in [1, +\infty)$, the p product metric d_p is defined as the p norm of the m -vector of the distances d_m . That is :

$$d_p(x, y) = \left(\sum_{i=1}^m d_i(x_i, y_i)^p \right)^{1/p}, \text{ for } x, y \in \mathcal{Y},$$

where $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_m)$.

For $p = \infty$, the p product metric is also called the sup metric and is defined as

$$d_{\infty}(x, y) = \max_{i \leq m} d_i(x_i, y_i).$$

Proposition 2.4.11. Let $(\mathcal{X}_i, k_{x \neq y})$ be m metric spaces each equipped with the k -discrete distance and let $\mathcal{Y} = \mathcal{X}_1 \times \dots \times \mathcal{X}_m$ be the Cartesian product of these m metric spaces.

Then, the sup metric d_∞ on \mathcal{Y} is the k -discrete distance on \mathcal{Y} . That is, $d_\infty(x, y) = k_{x \neq y}(x, y)$, for all $(x, y) \in \mathcal{Y} \times \mathcal{Y}$.

Proof: (Proposition 2.4.11)

As written in Definition 4.2.1, $d_\infty(x, y) = \max_{i \leq m} d_i(x_i, y_i)$ for $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_m)$. When $d_i(x, y) = k_{x \neq y}(x, y)$, $\forall i$, it is clear that

$$d_\infty(x, y) = \begin{cases} 0 & \text{if } x_i = y_i, \forall i \\ k & \text{otherwise.} \end{cases}$$

Thus, we can deduce that $d_\infty = k_{x \neq y}$ on \mathcal{Y} . Indeed, if $x = y$ we have $x_i = y_i, \forall i$ and therefore $d_\infty(x, y) = 0$. On the contrary, if $x \neq y$, then $x_{i_r} = y_{i_r}$ for some $r \in \{1, \dots, s\}$, $s \leq m$. Thus, $d_\infty(x, y) = k$. ■

A direct application of Theorem 2.4.7 on the cartesian product \mathcal{Y} yields the following result:

Theorem 2.4.12. [l.636] *Let $(\mathcal{X}_j, k_{x \neq y})$ be m metric spaces each equipped with the 1-discrete distance and let $\mathcal{Y} = \mathcal{X}_1 \times \dots \times \mathcal{X}_m$ be the Cartesian product of these m metric spaces.*

For each \mathcal{X}_j , let us define two atomic probability measures,

$$\mu_{1,j} = \sum_{i=1}^n \mu_{i,j}^{(1)} \delta_{x_{i,j}} \quad \text{and} \quad \mu_{2,j} = \sum_{i=1}^n \mu_{i,j}^{(2)} \delta_{x_{i,j}},$$

both supported on n_j finite number of points $\{x_{1j}, \dots, x_{n_j}\} \in \mathcal{X}_j$.

On \mathcal{Y} equipped with the sup metric, the two atomic probability measures are given by

$$\mu_1 = \sum_{i=1}^{n^m} \mu_i^{(1)} \delta_{y_i} \quad \text{and} \quad \mu_2 = \sum_{i=1}^{n^m} \mu_i^{(2)} \delta_{y_i},$$

supported on the same finite number of points $\{y_1, \dots, y_{n^m}\} \in \mathcal{Y}$. Then,

$$W_{\mathcal{Y}}(\mu_1, \mu_2) = \sum_{i=1}^{n^m} |\mu_1(y_i) - \mu_2(y_i)|.$$

2.5 The Kantorovich-Rubinstein distance on the line equipped with the Euclidean metric

There is a useful way to describe the collection of all finite measures on \mathbb{R} . If μ is such a measure, one defines the real function F by

$$F(x) = \mu((-\infty, x]).$$

Then, F is non-decreasing, right-continuous and satisfies $\lim_{x \rightarrow -\infty} F(t) = 0$ and $\lim_{x \rightarrow \infty} F(t) = \mu(\mathbb{R})$. Finally, for any bounded interval $(a, b]$, the following equality holds:

$$\mu((a, b]) = F(b) - F(a). \quad (2.5.1)$$

If μ is a probability measure, the function F is called the distribution function of μ . It is also often called the cumulative distribution function of μ .

The measure μ is completely determined by its distribution function F . Indeed, the following theorem (see[9] p.for a proof) ensures that to each F , there exist a μ :

Theorem 2.5.1. *Let F be a non-decreasing, right-continuous, real function on \mathbb{R} . Then there exists on the Borel σ -algebra of \mathbb{R} a unique measure μ satisfying equation 2.5.1, for all $a, b \in \mathbb{R}$.*

An immediate consequence of the previous theorem is that such an F is the distribution function of some random variable:

Lemma 2.5.2. *If F is non-decreasing, right-continuous, real function on \mathbb{R} , satisfying $\lim_{x \rightarrow -\infty} F(t) = 0$ and $\lim_{x \rightarrow \infty} F(t) = \mu(\mathbb{R})$, then there exists on \mathbb{R} a Borel random variable \mathcal{X} such that $F(x) = \mu[X \leq x]$.*

The two propositions 2.5.3 and 2.5.4 are well known and necessary to prove Theorem 2.5.5:

Proposition 2.5.3. *Let $F : \mathbb{R} \rightarrow [0, 1]$ be non-decreasing, right-continuous, real function on \mathbb{R} , satisfying $\lim_{x \rightarrow -\infty} F(t) = 0$ and $\lim_{x \rightarrow \infty} F(t) = 1$. Then, there exist a measurable, left-continuous function $G : [0, 1] \rightarrow \mathbb{R} \cup \{\pm\infty\}$ defined in the following way:*

(i) *for a given $y \in [0, 1]$, if there is an x such that $F(x) = y$,*

$$G(y) = \inf\{x; F(x) = y\}$$

(ii) *for a given $y \in [0, 1]$, if there is no x such that $F(x) = y$,*

$$G(y) = \inf_{z > y} G(z)$$
 where all z are such that $G(z)$ exists.

The function G is called the generalized inverse of F and is denoted by F^{-1} .

Proposition 2.5.4 (Probability Integral Transformation). *Let (Ω, μ) be a probability measured space and let $X : \Omega \rightarrow [0, 1]$ be a uniformly distributed random variable.*

Consider $F : \mathbb{R} \rightarrow [0, 1]$, a non-decreasing, right-continuous function satisfying $\lim_{x \rightarrow -\infty} F(t) = 0$ and $\lim_{x \rightarrow \infty} F(t) = 1$.

Then, $F^{-1}(X)$ is a random variable with distribution function F .

Proof: (Proposition 2.5.4)

The distribution function of the random variable $F^{-1}(X)$ will be denoted by G while the distribution function of X will be denoted F_X . Hence, we need to show that $G(t) = F(t), \forall t$.

By definition, $G(t) = \mu(\{\omega \in \Omega; F^{-1}(X)(\omega) \leq t\})$. Since F is non-decreasing, we obtain $G(t) = \mu(\{\omega \in \Omega; X(\omega) \leq F(t)\})$. But $\mu(\{\omega \in \Omega; X(\omega) \leq F(t)\}) = F_X(F(t))$ and X is uniformly distributed hence $\mu(\{\omega \in \Omega; X(\omega) \leq F(t)\}) = F(t)$. ■

We can now state a very important result. A proof was published by Vallender [49] in 1974. The proof has been revisited for clarity and to use the concept of couplings.

Theorem 2.5.5. *Let us consider \mathbb{R} equipped with the usual Euclidean metric. Let μ_1, μ_2 be two probability measures on \mathbb{R} and F_{μ_1}, F_{μ_2} be their respective cumulative distribution functions. Then,*

$$W_{\mathbb{R}}(\mu_1, \mu_2) = \int_{-\infty}^{+\infty} |F_{\mu_1}(x) - F_{\mu_2}(x)| dx.$$

Proof: (Theorem 2.5.5)

Suppose that $\int_{-\infty}^{+\infty} |F_{\mu_1}(x) - F_{\mu_2}(x)| dx < \infty$. Let ϑ be a coupling of μ_1 and μ_2 on $\mathbb{R} \times \mathbb{R}$ such that $\int |x - y| d\vartheta(x, y) \leq \infty$. Then, let us first show that

$$\int_{-\infty}^{+\infty} |F_{\mu_1}(x) - F_{\mu_2}(x)| dx \leq \int |x - y| d\vartheta(x, y).$$

Let us set $A = \{(x, y) \in \mathbb{R}^2; x > y\}$ and $B = \{(x, y) \in \mathbb{R}^2; y \geq x\}$. Then we can write

$$\int |x - y| d\vartheta(x, y) = \int_A (x - y) d\vartheta(x, y) + \int_B (y - x) d\vartheta(x, y).$$

Now, we need to prove that

$$\int_A (x - y) d\vartheta(x, y) = \int_{\mathbb{R}} \vartheta(D_s) ds \quad \text{and} \quad \int_B (y - x) d\vartheta(x, y) = \int_{\mathbb{R}} \vartheta(C_s) ds,$$

where $C_s = \{(x, y); x \leq s \text{ and } y > s\}$ and $D_s = \{(x, y); x > s \text{ and } y \leq s\}$.

We note that, by the construction of B ,

$$\int_B (y - x) d\vartheta(x, y) = \int_0^{+\infty} \vartheta(\{(x, y); y - x \geq t\}) dt$$

By the Disintegration Theorem A.1.1, we obtain

$$\vartheta(\{(x, y); y - x \geq t\}) = \int_{\mathbb{R}} \vartheta_x(\{(x, y); y \geq x + t\}) d\mu_1(x)$$

Then, by Fubini Theorem, we obtain

$$\begin{aligned} \int_B (y-x) d\vartheta(x,y) &= \int_0^{+\infty} \int_{\mathbb{R}} \vartheta_x(\{(x,y); y \geq x+t\}) d\mu_1(x) dt \\ &= \int_{\mathbb{R}} \int_0^{+\infty} \vartheta_x(\{(x,y); y \geq x+t\}) dt d\mu_1(x) \end{aligned}$$

But with the change of variable $s = x + t$, we have

$$\begin{aligned} \int_0^{+\infty} \vartheta_x(\{(x,y); y \geq x+t\}) dt &= \int_x^{+\infty} \vartheta(\{(x,y); y \geq s\}) ds \\ &= \int_{\mathbb{R}} \chi_{[x,+\infty)}(s) \vartheta(\{(x,y); y \geq s\}) ds \end{aligned}$$

For s fixed, we have $\{x \in \mathbb{R}; \chi_{[x,+\infty)}(s) = 1\} = \{x \in \mathbb{R}; x \leq s\}$ and $\{x; \chi_{[x,+\infty)}(s) = 0\} = \{x \in \mathbb{R}; x > s\}$. That is, $\chi_{[x,+\infty)}(s) = \chi_{(-\infty,s]}(x)$. Hence,

$$\begin{aligned} \int_B (y-x) d\vartheta(x,y) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \chi_{[x,+\infty)}(s) \vartheta_x(\{(x,y); y \geq s\}) ds d\mu_1(x) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \chi_{(-\infty,s]}(x) \vartheta_x(\{(x,y); y \geq s\}) d\mu_1(x) ds \end{aligned}$$

For a given s , $\int_{\mathbb{R}} \chi_{(-\infty,s]}(x) \vartheta_s(\{(x,y); y \geq s\}) d\mu_1(x) = \vartheta(\{(x,y); x \leq s \text{ and } y \geq s\})$.

Hence, we have $\int_B (y-x) d\vartheta(x,y) = \int_{\mathbb{R}} \vartheta(C_s) ds$.

Using a similar argument, we obtain $\int_A (x-y) d\vartheta(x,y) = \int_{\mathbb{R}} \vartheta(D_s) ds$. Therefore,

$$\int |x-y| d\vartheta(x,y) = \int_{\mathbb{R}} (\vartheta(C_s) + \vartheta(D_s)) ds.$$

Geometrically, it is easy to see that

$$C_s = \{(x,y); x \leq s\} \setminus \{(x,y); x \leq s \text{ and } y \leq s\}$$

$$D_s = \{(x,y); y \leq s\} \setminus \{(x,y); x \leq s \text{ and } y \leq s\}.$$

To simplify notation for the rest of the proof we will denote the set $\{(x, y); x \leq s \text{ and } y \leq s\}$ by E_s . We thus obtain:

$$\begin{aligned} \vartheta(C_s) + \vartheta(D_s) &= \vartheta(\{(x, y); x \leq s\}) + \vartheta(\{(x, y); y \leq s\}) - 2\vartheta(E_s) \\ &\geq \mu_1(x \leq s) + \mu_2(y \leq s) \\ &\quad - 2 \min\{\vartheta(\{(x, y); x \leq s\}), \vartheta(\{(x, y); y \leq s\})\} \\ &\geq F_{\mu_1}(s) + F_{\mu_2}(s) - 2 \min\{F_{\mu_1}(s), F_{\mu_2}(s)\} \end{aligned}$$

For $F_{\mu_1}(s) \leq F_{\mu_2}(s)$, we have $F_{\mu_1}(s) + F_{\mu_2}(s) - 2 \min(F_{\mu_1}(s), F_{\mu_2}(s)) = F_{\mu_2}(s) - F_{\mu_1}(s)$. For $F_{\mu_1}(s) \geq F_{\mu_2}(s)$, we have $F_{\mu_1}(s) + F_{\mu_2}(s) - 2 \min(F_{\mu_1}(s), F_{\mu_2}(s)) = F_{\mu_1}(s) - F_{\mu_2}(s)$. Thus, $F_{\mu_1}(s) + F_{\mu_2}(s) - 2 \min(F_{\mu_1}(s), F_{\mu_2}(s)) = |F_{\mu_1}(s) - F_{\mu_2}(s)|$. Hence we have

$$\int |x - y| d\vartheta(x, y) \geq \int_{\mathbb{R}} |F_{\mu_1}(s) - F_{\mu_2}(s)| ds.$$

As $W_{\mathbb{R}}(\mu_1, \mu_2)$ is the infimum over all couplings of μ_1 and μ_2 , it is the greatest lower bound and thus

$$W_{\mathbb{R}}(\mu_1, \mu_2) \geq \int_{\mathbb{R}} |F_{\mu_1}(s) - F_{\mu_2}(s)| ds.$$

Now, it is left to show that there exist a coupling ϑ^* such that

$$\int_{\mathbb{R}} |F_{\mu_1}(x) - F_{\mu_2}(x)| dx = \int |x - y| d\vartheta^*(x, y). \quad (2.5.2)$$

Let $X : \mathbb{R} \rightarrow [0, 1]$ be a uniformly distributed random variable. Then, by Proposition 2.5.4, the real random variable $F_{\mu_1}^{-1}(X)$, respectively $F_{\mu_2}^{-1}(X)$ has distribution function F_{μ_1} , respectively F_{μ_2} . We choose, as a candidate for ϑ^* , the product measure on \mathbb{R}^2 of the push-forward measures $F_{\mu_1}^{-1}(X)(\mu_1)$ and $F_{\mu_2}^{-1}(X)(\mu_1)$ of μ_1 from \mathbb{R} .

First, we verify that the candidate product measure for ϑ^* is a coupling of the measures μ_1 and μ_2 . That is, one needs to show that, for $i = 1, 2$, $\pi_i(\vartheta^*) = \mu_i$, where π_i is the natural projection from \mathbb{R}^2 to \mathbb{R} . If A_s denotes $\{(x_1, x_2) \in \mathbb{R}^2; x_i \in (-\infty, s]\}$,

we have:

$$\begin{aligned}
\pi_i(\vartheta^*)((-\infty, s]) &= \vartheta^* (\{(x_1, x_2) \in \mathbb{R}^2; \pi_i(x_1, x_2) \in (-\infty, s]\}) \\
&= (F_{\mu_1}^{-1} \times F_{\mu_2}^{-1})(X)(\mu_1)(A_s) \quad \text{where} \\
&= \mu_1(\{x \in \mathbb{R}; (F_{\mu_1}^{-1} \times F_{\mu_2}^{-1})(X)(x) \in A_s\}) \\
&= \mu_1(\{x \in \mathbb{R}; (F_{\mu_1}^{-1}(X)(x); F_{\mu_2}^{-1}(X)(x)) \in A_s\}) \\
&= \mu_1(\{x \in \mathbb{R}; F_{\mu_i}^{-1}(X)(x) \in (-\infty, s]\}) \\
&= \mu_1(\{x \in \mathbb{R}; X(x) \in (-\infty, F_{\mu_i}(s)]\}) \\
&= F_{\mu_i}(s) \quad (\text{since } X \text{ is uniformly distributed}), \\
&= \mu_i((-\infty, s]).
\end{aligned}$$

It is worth noting that one could have used the push-forward measures of μ_2 or of any other probability measure on \mathbb{R} .

To complete the proof, we show that the equality (2.5.2) holds for our choice of ϑ^* .

We know that:

$$\int_{\mathbb{R}} \vartheta^*(C_s) + \vartheta^*(D_s) \, ds = \vartheta^* (\{(x, y); x \leq s\}) + \vartheta^* (\{(x, y); y \leq s\}) - 2\vartheta^*(E_s)$$

where $\vartheta^* (\{(x, y); x \leq s\}) = F_{\mu_1}(s)$ and $\vartheta^* (\{(x, y); y \leq s\}) = F_{\mu_2}(s)$. Let us now focus on $\vartheta^*(E_s)$:

$$\begin{aligned}
\vartheta^*(E_s) &= (F_{\mu_1}^{-1} \times F_{\mu_2}^{-1})(X)(\mu_1)(E_s) \\
&= \mu_1(\{x \in \mathbb{R}; (F_{\mu_1}^{-1}(X)(x); F_{\mu_2}^{-1}(X)(x)) \in E_s\}) \\
&= \mu_1(\{x \in \mathbb{R}; X(x) \in (-\infty, F_{\mu_1}(s)] \text{ and } X(x) \in (-\infty, F_{\mu_2}(s)]\}) \\
&= \min\{\mu_1(\{x \in \mathbb{R}; X(x) \leq F_{\mu_1}(s)\}), \mu_1(\{x \in \mathbb{R}; X(x) \leq F_{\mu_2}(s)\})\} \\
&= \min\{F_{\mu_1}(s), F_{\mu_2}(s)\} \quad (\text{since } X \text{ is uniformly distributed}).
\end{aligned}$$

As shown above, we therefore obtain $\vartheta^*(C_s) + \vartheta^*(D_s) = |F_{\mu_1} - F_{\mu_2}|$ and thus Equation

(2.5.2) is true and the proof is complete. ■

2.5.1 The Kantorovich-Rubinstein distance for atomic measures on the line

Proposition 2.5.6. *Let us consider \mathbb{R} equipped with the usual euclidean metric. Let us define, on \mathbb{R} , two atomic probability measures*

$$\mu_1 = \sum_{i=1}^n \mu_i^{(1)} \delta_{x_i} \quad \text{and} \quad \mu_2 = \sum_{i=1}^n \mu_i^{(2)} \delta_{x_i},$$

supported on the same finite number of ordered points $\{x_1, \dots, x_n\} \in \mathbb{R}$.

The Kantorovich-Rubinstein distance between μ_1 and μ_2 is then given by

$$W_{\mathbb{R}}(\mu_1, \mu_2) = \sum_{i=1}^{n-1} |F_i|(x_{i+1} - x_i),$$

where $F_i = \sum_{k=1}^i \mu_k^{(1)} - \mu_k^{(2)}$.

Proof: (Proposition 2.5.6)

By construction, the distribution functions F_{μ_1} and F_{μ_2} of the atomic measures μ_1 and μ_2 supported on the ordered set $\{x_1, \dots, x_n\} \in \mathbb{R}$ are defined as

$$F_{\mu_1}(t) = \sum_{i=1}^n \mu_i^{(1)} 1_{x_i \leq t} \quad \text{and} \quad F_{\mu_2}(t) = \sum_{i=1}^n \mu_i^{(2)} 1_{x_i \leq t}.$$

Since the distribution function of an atomic measure is a simple function, $|F_{\mu_1} - F_{\mu_2}|$ is also a simple function. Integrals of simple functions are well known (see Billingsley [10]). Thus, a direct application of Theorem 2.5.5 for μ_1 and μ_2 yields the

following Kantorovich-Rubinstein distance:

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \sum_{i=1}^{n-1} |F_i|(x_{i+1} - x_i),$$

where $F_i = \sum_{k=1}^i \mu_k^{(1)} - \mu_k^{(2)}$. ■

2.6 The Kantorovich-Rubinstein distance on the circle \mathbb{S}_1

In this section we consider the Kantorovich-Rubinstein distance where the underlying Polish space \mathcal{X} is the unit circle. We use the following notations from the paper on the Wasserstein distance on the circle by Cabrelli and Molter[14]:

We identify the circle $K = \{z \in \mathbb{C} : |z| = 1\}$ with $T = [0, 1)$ as a fundamental domain for \mathbb{R}/\mathbb{Z} , via the transformation $t \mapsto e^{i2\pi t}$, and use the natural metric on T given by $\rho(s_1, s_2) = \min(|s_1 - s_2|, 1 - |s_1 - s_2|)$. It corresponds to the minimum arc length on the circle.

We identify the functions on T with the periodic functions on \mathbb{R} of period 1; using this identification we will use $f(t)$, with $t \in \mathbb{R}$, for functions on T .

We denote by $(\mathcal{X}, |\cdot|)$ the unit interval $[0, 1]$ on the line with the Euclidean distance. For measures with bounded support on \mathcal{X} , the distribution function of μ is defined by $F_{\mu}(t) = \mu(\{x \in [0, 1] : x \in [0, t]\}) = \mu([0, t])$.

An analytic expression for the Kantorovich-Rubinstein distance is between probability measures on the circle is given in Theorem 2.6.5. We first need the following definition 2.6.1 and three lemmas from [14].

Definition 2.6.1. *For an arbitrary measurable function $\gamma : \mathcal{X} \rightarrow \mathbb{R}$, consider the*

following three measurable sets:

$$A^+(\gamma) = \{x \in \mathcal{X} : \gamma(x) > 0\}, \quad A^-(\gamma) = \{x \in \mathcal{X} : \gamma(x) < 0\}, \quad A^0(\gamma) = \{x \in \mathcal{X} : \gamma(x) = 0\}.$$

Then, a function $\gamma : \mathcal{X} \rightarrow \mathbb{R}$ is said to be balanced if

$$|\lambda(A^+(\gamma)) - \lambda(A^-(\gamma))| \leq \lambda(A^0(\gamma)).$$

We also denote by D the set defined by $D(\gamma) = \{c : A^0(\gamma) \rightarrow \mathbb{R}, \text{ measurable: } |c(t)| \leq 1 \text{ a.e.}\}$.

We now relate the distance between measures on the circle with the distance between measures on \mathcal{X} obtained from the former by "cutting" the circle. First, we identify the measures on T with the appropriate measures on \mathcal{X} .

For $\mu \in M(T)$, we consider the function $G_\mu : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$G_\mu(x) = \mu([0, x]), \quad \text{for } 0 \leq x < 1,$$

and extended to \mathbb{R} by the equation $G_\mu(x + 1) = G_\mu(x) + 1$. By construction, G_μ is right-continuous.

For each $s \in T$, we associate to any measure $\mu \in M(T)$ the pair (μ_s^D, μ_s^T) of measures on T , determined by their respective distribution functions:

$$\begin{aligned} D_\mu^s : \mathcal{X} &\rightarrow \mathbb{R}, \quad \text{given by } D_\mu^s(x) = G_\mu(x + s) - G_\mu(s), \\ I_\mu^s : \mathcal{X} &\rightarrow \mathbb{R}, \quad \text{given by } I_\mu^s(x) = G_\mu(x + s) - G_\mu(s-). \end{aligned}$$

As distribution functions, D_μ^s and I_μ^s are defined on \mathbb{R} . Note that:

- (i) $D_\mu^s(x) = 0$ for $x \in (-\infty, 0]$, $D_\mu^s(x) = 1$ for $x \in [1, +\infty)$ and D_μ^s has a jump of height $G_\mu(s) - G_\mu(s^-)$ at $x = 1$ while $I_\mu^s(x) = 0$ for $x \in (-\infty, 0)$, $I_\mu^s(x) = 1$ for $x \in [1, +\infty)$ and has a jump of height $G_\mu(s) - G_\mu(s^-)$ at $x = 0$;
- (ii) if $\mu(s) = 0$ then G_μ is continuous at s and thus $\mu_s^D = \mu_s^I$ and $D_\mu^s = I_\mu^s$. Moreover, note that $\mu_s^D(\{0\}) = 0$ and that $\mu_s^I(\{1\}) = 0$.

(iii) $\mu_s^D(\{0\}) = 0$ and $\mu_s^I(\{1\}) = 0$.

Informally, one refers to the value $s \in T$ as a cut of the circle. One can picture μ_s^D, μ_s^I , as representing the measures on the line obtained by "cutting" the circle at s and taking $(s, s + 1]$ and $[s, s + 1)$ as fundamental domains in \mathbb{R}/\mathbb{Z} . Now that we have defined balanced functions, we can state the lemmas:

Lemma 2.6.2. *Let $\mu, \nu \in M(T)$ and $\eta = \mu - \nu$. For $r \in T$ fixed, we consider the function $D^r = D_\mu^r - D_\nu^r$, the difference of the distribution functions associated to μ_r^D and ν_r^D . Then,*

$$W_T(\mu, \nu) \leq \inf_{r \in T} W_{\mathcal{X}}(\mu_r^D, \nu_r^D). \quad (2.6.1)$$

Moreover if, for $s \in T$, the function D^s is balanced, the infimum on the right-hand side of the equation (2.6.1) is attained at s . Thus, for D^s balanced we can write

$$W_T(\mu, \nu) = W_{\mathcal{X}}(\mu_s^D, \nu_s^D).$$

Analogously, if for $s \in T$, the function $I^s = I_\mu^s - I_\nu^s$ is balanced, we have

$$W_T(\mu, \nu) = W_{\mathcal{X}}(\mu_s^I, \nu_s^I) = \inf_{r \in T} W_{\mathcal{X}}(\mu_r^I, \nu_r^I).$$

The next lemma show that there always exists a so-called optimal cut $s \in T$. that is, an s such that D^s or I^s is balanced:

Lemma 2.6.3. *Let μ and ν be two measures on T . Define $G = G_\mu - G_\nu$ on \mathbb{R} and let α be the restriction of G on \mathcal{X} (i.e. $\alpha(x) = G_\mu(x) - G_\nu(x)$, for $x \in [0, 1]$). Then, there exists an optimal cut $s \in T$ such that either D^s or I^s is balanced.*

Now, we need one more lemma to prove Theorem 2.6.5, on the existence of an analytic expression for the Kantorovich distance between probability measures on the circle.

Lemma 2.6.4. *Let $\gamma : \mathcal{X} \rightarrow \mathbb{R}$ be a right continuous function and λ be the Lebesgue measure on \mathbb{R} . Consider the function $m_\gamma : \mathbb{R} \rightarrow \mathcal{X}$ defined by*

$$m_\gamma(t) = \lambda(\{x \in \mathcal{X} : \alpha(x) \geq t\}).$$

Define the constant a_γ by

$$a_\gamma = \sup \left\{ t \in \mathbb{R} : m_\gamma(t) > \frac{1}{2} \right\}.$$

Then, for any neighbourhood V_{a_γ} of a_γ , $\lambda(V_{a_\gamma} \cap \gamma(\mathcal{X})) > 0$. In particular, a_γ is in the closure of $\gamma(\mathcal{X})$.

Likewise, define the constant b_γ by

$$b_\gamma = \inf \left\{ t \in \mathbb{R} : m_\gamma(t) < \frac{1}{2} \right\}.$$

Then, for any neighbourhood V_{b_γ} of b_γ , $\lambda(V_{b_\gamma} \cap \gamma(\mathcal{X})) > 0$. In particular, b_γ is in the closure of $\gamma(\mathcal{X})$.

We can now prove the theorem:

Theorem 2.6.5. *Let T be the unit circle equipped with the minimum arc length metric $\rho(s_1, s_2) = \min(|s_1 - s_2|, 1 - |s_1 - s_2|)$. Define two probability measures μ and ν on T . Then,*

$$W_T(\mu, \nu) = \int_T |\alpha(x) - a_\alpha| dx,$$

where $\alpha : \mathcal{X} \rightarrow \mathbb{R}$ is defined on $x \in [0, 1)$ by

$$\alpha(x) = \mu([0, x]) - \nu([0, x]) \text{ with } \alpha(1) = 0,$$

$$\text{and } a_\alpha = \sup \left\{ t \in \mathbb{R} : \lambda(\{x \in \mathcal{X} : \alpha(x) \geq t\}) > \frac{1}{2} \right\}$$

is a translation constant that depends on μ and ν .

Proof: (Theorem 2.6.5)

Recall that, by definition, α is the restriction to \mathcal{X} of the function G defined in Lemma 2.6.3. Moreover, by the same lemma, we know that there exists $s \in T$ such that either D^s or I^s is balanced.

Assume first that $D^s = D_\mu^s - D_\nu^s$ is balanced, where D_μ^s and D_ν^s are the distribution functions of μ_s^D and ν_s^D . By Lemma 2.6.2, $W_T(\mu, \nu) = W_{\mathcal{X}}(\mu_s^D, \nu_s^D)$.

Now, by Theorem 2.5.5, we know that

$$W_{\mathcal{X}}(\mu_s^D, \nu_s^D) = \int_0^1 |D^s(x)| dx \quad \text{since} \quad D^s = D_\mu^s - D_\nu^s.$$

Recall also that $D^s(x) = G(x + s) - G(s)$. But $s \in T$ hence $G(s) = \alpha(s)$. Also, the function $\alpha(s)$ is right-continuous. Hence, by Lemma 2.6.4, there exists a sequence $\{s_n\}$ in \mathcal{X} such that $\alpha(s_n) \rightarrow a_\alpha$. Let $s \in \mathcal{X}$ be the limit of point of $\{s_n\}$. Since $\{s_n\} \subset \mathcal{X}$, we can extract a decreasing sub-sequence $\{s_{n_k}\}$ such that $\{s_{n_k}\} \rightarrow s$. Since α is right-continuous, $\alpha(s_{n_k}) \rightarrow \alpha(s)$ and thus $\alpha(s) = a_\alpha$. Therefore we have $D^s(x) = G(x + s) - a_\alpha$.

By a basic change of variables we obtain:

$$\int_0^1 |G(x + s) - a_\alpha| dx = \int_s^1 |G(x) - a_\alpha| dx + \int_1^{1+s} |G(x) - a_\alpha| dx.$$

For $x \in [1, 1 + s]$, we can write $x = 1 + r$ for $r \in [0, s]$ and thus we have

$$G(x) = G(1 + r) = G_\mu(1 + r) - G_\nu(1 + r) = (1 + G_\mu(r)) - (1 + G_\nu(r)) = G(r) = \alpha(r).$$

Therefore,

$$\int_0^1 |G(x + s) - a_\alpha| dx = \int_s^1 |G(x) - a_\alpha| dx + \int_0^s |G(x) - a_\alpha| dx = \int_0^1 |\alpha(x) - a_\alpha| dx,$$

which completes the proof.

If, on the other hand, s is such that I^s is balanced, the same steps allow to show

that

$$W_T(\mu, \nu) = \int_0^1 |I^s(x)| dx = \int_0^1 |\alpha(x) - a_\alpha| dx.$$

■

We end this section, with the following corollary of Theorem 2.6.5 that we will use in the next section:

Corollary 2.6.6. *Under the same hypothesis as Theorem 2.6.5, we have*

$$W_T(\mu, \nu) = \min \left(\inf_{s \in T} \int_0^1 |\alpha(x) - \alpha(s)| dx, \inf_{s \in T} \int_0^1 |\alpha(x) - \alpha(s-)| dx \right).$$

Proof: (Corollary 2.6.6)

By Lemma 2.6.2, we know that

$$W_T(\mu, \nu) \leq \inf_{r \in T} W_{\mathcal{X}}(\mu_r^D, \nu_r^D) \quad \text{and} \quad W_T(\mu, \nu) \leq \inf_{r \in T} W_{\mathcal{X}}(\mu_r^I, \nu_r^I).$$

Therefore,

$$W_T(\mu, \nu) \leq \min \left(\inf_{r \in T} W_{\mathcal{X}}(\mu_r^I, \nu_r^I), \inf_{r \in T} d_X(\mu_r^D, \nu_r^D) \right).$$

Now, by Theorem 2.5.5, we know that

$$W_{\mathcal{X}}(\mu_s^D, \nu_s^D) = \int_0^1 |D^s(x)| dx \quad \text{since} \quad D^s = D_\mu^s - D_\nu^s,$$

where D_μ^s and D_ν^s are the distribution functions of μ_s^D and ν_s^D , respectively.

Moreover, by Lemma 2.6.3 and Lemma 2.6.2, we know that there exists an $s \in T$ such that

$$W_T(\mu, \nu) = W_{\mathcal{X}}(\mu_s^D, \nu_s^D) \quad \text{or} \quad W_T(\mu, \nu) = W_{\mathcal{X}}(\mu_s^I, \nu_s^I).$$

Therefore, since $D^s(x) = G(x + s) - \alpha(s)$ and $I^s(x) = G(x + s) - \alpha(s-)$ (see the

proof of Theorem 2.6.5),

$$d_T(\mu, \nu) = \min \left(\inf_{s \in T} \int_0^1 |\alpha(x) - \alpha(s)| dx, \inf_{s \in T} \int_0^1 |\alpha(x) - \alpha(s-)| dx \right).$$

■

2.6.1 The Kantorovich-Rubinstein distance on the circle \mathbb{S}^1 for atomic measures

Let us consider the case of atomic measures on the circle T . We have the following proposition:

Proposition 2.6.7. *Let the unit circle T be equipped with the minimum arc length metric $\rho(s_1, s_2) = \min(|s_1 - s_2|, 1 - |s_1 - s_2|)$. Let us define, on T , two atomic probability measures*

$$\mu_1 = \sum_{i=1}^n \mu_i^{(1)} \delta_{s_i} \quad \text{and} \quad \mu_2 = \sum_{i=1}^m \mu_i^{(2)} \delta_{s_i},$$

supported on the same finite number of ordered points $\{s_1, \dots, s_n\} \in T$.

The Kantorovich-Rubinstein distance between μ_1 and μ_2 is then given by

$$W_T(\mu_1, \mu_2) = \min_{1 \leq s \leq n} \sum_{i=1}^n \rho(s_{i+1}, s_i) |\alpha_i - \alpha_s|,$$

where $\alpha_j = \sum_{k=1}^j \mu_k^{(1)} - \mu_k^{(2)}$ for $1 \leq j \leq n$, and $s_{n+1} \equiv s_1$.

Proof: (Proposition 2.6.7)

By Corollary 2.6.6, we know that

$$W_T(\mu, \nu) = \min \left(\inf_{s \in T} \int_0^1 |\alpha(x) - \alpha(s)| dx, \inf_{s \in T} \int_0^1 |\alpha(x) - \alpha(s-)| dx \right).$$

where $\alpha(x) = F_{\mu_1}(x) - F_{\mu_2}(x)$ with $F_{\mu_1}(x) = \mu_1([0, x])$ and $F_{\mu_2}(x) = \mu_2([0, x])$, respectively.

Since both μ_1 and μ_2 are atomic measures, α is a simple function. Evaluating α at s and s^- , we have

$$\alpha(s) = \sum_{k=1}^i \mu_k^{(1)} - \mu_k^{(2)} \text{ for } s_i \leq s < s_{i+1} \quad \text{and} \quad \alpha(s^-) = \sum_{k=1}^i \mu_k^{(1)} - \mu_k^{(2)} \text{ for } s_i < s \leq s_{i+1},$$

with $\alpha(s) = 0$ for $s_n \leq s < s_1$ and $\alpha(s^-) = 0$ for $s_n < s \leq s_1$.

Note that the functions $\alpha(x) - \alpha(s)$ and $\alpha(x) - \alpha(s^-)$ remain simple since $\alpha(s)$ and $\alpha(s^-)$ are constants. Integrals of simple functions are well known (see Billingsley [10]) thus a direct application of Corollary 2.6.6 for μ_1 and μ_2 yields the following Kantorovich-Rubinstein distance:

$$W_T(\mu, \nu) = \min \left(\min_{1 \leq s \leq n} \sum_{i=1}^n \rho(s_{i+1}, s_i) |\alpha(x) - \alpha(s)|, \min_{1 \leq s \leq n} \sum_{i=1}^n \rho(s_{i+1}, s_i) |\alpha(x) - \alpha(s^-)| \right).$$

where $\alpha_j = \sum_{k=1}^j \mu_k^{(1)} - \mu_k^{(2)}$ for $1 \leq j \leq n$, and $s_{n+1} \equiv s_1$.

By construction of $\alpha(s)$ and $\alpha(s^-)$, $\min_{s_1 \leq s \leq s_n} \alpha(x) - \alpha(s) = \min_{s_1 \leq s \leq s_n} \alpha(x) - \alpha(s^-)$. Hence, we obtain

$$W_T(\mu_1, \mu_2) = \min_{1 \leq s \leq n} \sum_{i=1}^n \rho(s_{i+1}, s_i) |\alpha_i - \alpha_s|.$$

■

Chapter 3

The Kantorovich-Rubinstein Distance and Statistical Trend Tests

The most common design to test for association between a genetic marker and a disease is the case-control study. Basic test statistics for GWAS using case-control sample are reviewed by Balding [4]. In this chapter, we start by studying two of the most commonly used test statistics for genetic association: Pearson's Chi-square test and the Cochran-Armitage trend test which assumes a dose-response effect between the genotype and the disease: the risk of disease increases with the number of risk allele (Sasieni [46]).

In 2009, Zheng *et al.* [56] showed that, in fact, Pearson's test is a trend test with unrestricted data-driven scores. By re-writing both test statistics in a more general form, we are able to significantly simplify the proofs of Zheng *et al.* and build an upper and lower bound for the Pearson statistic that depends on the Kantorovich-Rubinstein distance.

The chapter is organised as follows: the first section introduces the two test statistics. In the second section, we define a particular application T and proceed to give the interesting properties of T . In section 3, we show how both the Pearson and

the Cochran-Armitage statistic can be defined as multiples of T . Finally, in section 4, we construct an upper and lower bound for the Pearson statistic as functions of the Kantorovich-Rubinstein distance.

3.1 The Cochran-Armitage Trend Test and the Pearson's Chi-square Test of Homogeneity

The purpose of this section is to introduce two single variable tests of association which are well-known and commonly used in genome-wide association studies (GWAS): the Pearson chi-square test of independence and the Cochran-Armitage test for trend.

3.1.1 The Pearson's Chi-square Test of Homogeneity

In the case of a multinomial distribution with joint probabilities π_{ij} , for $i = 1, \dots, k$ and $j = 1, \dots, l$, the sampling can be summarize by an $k \times l$ contingency table. The null hypothesis is that the joint probabilities π_{ij} are equal to the product of their marginals π_{i+} and π_{+j} . Hence the null hypothesis of the statistical test is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, for all (i, j) , where $\sum_i \sum_j \pi_{ij} = 1$. Since the marginal distributions π_{i+} and π_{+j} are unknown, the sample marginal proportions $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$ are used as estimates. Under the null hypothesis, one obtains the following statistic:

$$T_{\chi^2} = \sum_i^k \sum_j^l \frac{(o_{ij} - \hat{\varepsilon}_{ij})^2}{\hat{\varepsilon}_{ij}}, \tag{3.1.1}$$

where o_{ij} is the number of observations of type (i, j) and $\hat{\varepsilon}_{ij}$ is the estimated expected frequency of type (i, j) under the null hypothesis (hence $\hat{\varepsilon}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j}$).

Replacing ε_{ij} by the estimates $\hat{\varepsilon}_{ij}$ affects the distribution of the Pearson chi-squared statistic. Indeed, $\hat{\varepsilon}_{ij}$ require estimating both marginal distributions π_{i+} and π_{+j} thus the degrees of freedom of the statistic is $(k - 1)(l - 1)$.

	aa	aA	AA	
Case	r_0	r_1	r_2	r
Control	s_0	s_1	s_2	s
	n_0	n_1	n_2	n

Table 3.1: Nomenclature for Contingency Table with one SNP

The Pearson’s Chi-square Test of Homogeneity Applied to Genetics

To summarize the results of case-control samples for a single marker (in our case a single snp), one can build a 2×3 contingency table. The two rows summarize the outcome of controls and cases while the columns summarize the outcome of the genotype of homozygous major (AA), heterozygous(Aa) and homozygous minor (aa), respectively. Hence, we obtain categorical data with 2 variables. The row variable has two categories while the ordinal column variable has 3 categories. To build this contingency table, we will follow the notation used by Zheng *et al.* in [56]. Denote the genotype counts for cases by (r_0, r_1, r_2) and for controls by (s_0, s_1, s_2) . Let $r = r_0 + r_1 + r_2$ and $s = s_0 + s_1 + s_2$ be the number of cases and controls while $n_i = r_i + s_i$ (for $i = 0, 1, 2$) is the number of patients with each genotype. Let $n = r + s$ be the total number of patients and $p = r/n$, $q = s/n$ be the sample marginal proportions. Recall that $q = 1 - p$. With such notations, the estimated expected frequencies can be written as $\hat{\varepsilon}_{ij} = n\hat{\pi}_{r+}\hat{\pi}_{+j} = n\frac{r}{n}\frac{n_i}{n} = n_i\frac{r}{n} = n_i p$.

The Pearson-Chi square test of homogeneity works under the null hypothesis that the probability distribution of any genotype is the same between case and control. Hence, the null hypothesis is $H_o : \pi_{ri} = \pi_{si}$ for $i = 0, 1, 2$, where $\sum_i \pi_{ri} = \sum_i \pi_{si} = 1$. Under the null hypothesis, the statistic 3.1.1 is written in the following form :

$$T_{\chi^2} = \sum_{i=0}^2 \frac{(r_i - n_i p)^2}{n_i p} + \sum_{i=0}^2 \frac{(s_i - n_i q)^2}{n_i q}. \tag{3.1.2}$$

The statistic 3.1.2 can equivalently be written using the *Brandt-Snedecor for-*

mula:

$$T_{\chi^2} = \frac{1}{pq} \sum_{i=0}^3 n_i (p_i - p)^2, \quad \text{where } p_i = \frac{r_i}{n_i}. \quad (3.1.3)$$

Indeed,

$$\begin{aligned} T_{\chi^2} &= \sum_{i=0}^2 \frac{(r_i - n_i p)^2}{n_i p} + \sum_{i=0}^2 \frac{(s_i - n_i q)^2}{n_i q} \\ &= \sum_{i=0}^2 \frac{n_i^2 \left(\frac{r_i}{n_i} - p\right)^2}{n_i p} + \sum_{i=0}^2 \frac{n_i^2 \left(\frac{s_i}{n_i} - q\right)^2}{n_i q} \\ &= \sum_{i=0}^2 \frac{n_i}{p} (p_i - p)^2 + \sum_{i=0}^2 \frac{n_i}{q} (q_i - q)^2 \quad \text{where } q_i = \frac{s_i}{n_i}, \\ &= \sum_{i=0}^2 \frac{n_i}{p} (p_i - p)^2 + \sum_{i=0}^2 \frac{n_i}{q} (p - p_i)^2 \quad (\text{note that } p_i + q_i = 1), \\ &= \sum_{i=0}^2 n_i (p_i - p)^2 \left(\frac{1}{p} + \frac{1}{q}\right) \\ &= \frac{1}{pq} \sum_{i=0}^2 n_i (p_i - p)^2. \end{aligned}$$

3.1.2 The Cochran-Armitage Trend Test

The Cochran-Armitage Trend Test (CATT) is a method of directing the Pearson's Chi-square Test towards narrow alternatives. Recall that for the Pearson's Chi-square Test, Pearson assumes that an $2 \times l$ contingency table represents the sampling of a $2l$ -multinomial distribution. In the case of the CATT, Cochran and Armitage assume that a $2 \times l$ contingency table represents the sampling of J independent binomial random variables of parameter (ρ_i, n_i) where the probability of success ρ_i is the probability of either categories as a function of the ordinal categories.

As seen above, in the in the case of a case-control setting for a single snp, the ordinal column variable has 3 categories: the 3 genotypes. In genetics ρ_i is the

probability of being a case given the genotype. It is called the penetrance. CATT works under the null hypothesis that the probability of success ρ_i of each random variable is the same. Hence the null hypothesis is $H_o : \rho_i = \rho_o$ for $i = 0, 1, 2$. Under the null hypothesis, Cochran and Armitage proposed the following trend statistic:

$$T_{CA}(c) = \frac{\left(\sum_{i=0}^2 c_i (qr_i - ps_i) \right)^2}{npq \left(\sum_{i=0}^2 c_i^2 \frac{n_i}{n} - \left(\sum_{i=0}^2 c_i \frac{n_i}{n} \right)^2 \right)} \quad (3.1.4)$$

where $c = (c_0, c_1, c_2)$ is a chosen triple of increasing scores such that $c_0 \leq c_1 \leq c_2$.

For large samples, the statistic T_{CA} is asymptotically equal to T_{χ^2} with one degree of freedom.

The Cochran-Armitage statistic (3.1.4) can be rearranged to obtain the following expression:

$$T_{CA}(c) = \frac{b^2}{pq} \sum_{i=0}^2 n_i (c_i - \bar{c}). \quad (3.1.5)$$

where

$$b = \frac{\sum n_i (p_i - p)(c_i - \bar{c})}{\sum n_i (c_i - \bar{c})^2} \quad \text{and} \quad \bar{c} = \frac{1}{n} \sum_{i=0}^2 n_i c_i.$$

In order to obtain the expression (3.1.5), we first show that

$$T_{CA}(c) = \frac{\left(\sum_{i=0}^2 r_i (c_i - \bar{c}) \right)^2}{pq \sum_{i=0}^2 n_i (c_i - \bar{c})^2}.$$

To do so, we show that their respective numerators and denominators are equal.

For the numerator we obtain:

$$\begin{aligned}
 \sum_{i=0}^2 r_i(c_i - \bar{c}) &= \sum_{i=0}^2 r_i \left(c_i - \sum_{i=0}^2 c_i \frac{n_i}{n} \right) \\
 &= \sum_{i=0}^2 r_i c_i - \sum_{i=0}^2 r_i \sum_{i=0}^2 c_i \frac{n_i}{n} \\
 &= \sum_{i=0}^2 r_i c_i - \frac{1}{n} \sum_{i=0}^2 r_i \sum_{i=0}^2 c_i n_i \\
 &= \sum_{i=0}^2 r_i c_i - p \sum_{i=0}^2 c_i (r_i + s_i) \\
 &= (1-p) \sum_{i=0}^2 r_i c_i - p \sum_{i=0}^2 c_i s_i \\
 &= \sum_{i=0}^2 c_i (q r_i - p s_i).
 \end{aligned}$$

For the denominator we have:

$$\begin{aligned}
 \sum_{i=0}^2 n_i (c_i - \bar{c})^2 &= \sum_{i=0}^2 n_i (c_i^2 - 2c_i \bar{c} + \bar{c}^2) \\
 &= \sum_{i=0}^2 n_i c_i^2 - 2\bar{c} \sum_{i=0}^2 c_i n_i + \bar{c}^2 \sum_{i=0}^2 n_i \\
 &= \sum_{i=0}^2 n_i c_i^2 - 2n\bar{c}^2 + \bar{c}^2 n \quad (\text{recall that } n\bar{c} = \sum_{i=0}^2 c_i n_i) \\
 &= n \left(\sum_{i=0}^2 n_i c_i^2 - \bar{c}^2 \right) \\
 &= n \left(\sum_{i=0}^2 c_i^2 \frac{n_i}{n} - \left(\sum_{i=0}^2 c_i \frac{n_i}{n} \right)^2 \right).
 \end{aligned}$$

Now, it is left to show that

$$\frac{\left(\sum_{i=0}^2 r_i(c_i - \bar{c})\right)^2}{pq \sum_{i=0}^2 n_i(c_i - \bar{c})^2} = \frac{b^2}{pq} \sum_{i=0}^2 n_i(c_i - \bar{c})^2.$$

Since $b = \frac{\sum n_i(p_i - p)(c_i - \bar{c})}{\sum n_i(c_i - \bar{c})^2}$, it is easy to see that

$$\frac{b^2}{pq} \sum_{i=0}^2 n_i(c_i - \bar{c})^2 = \frac{\left(\sum_{i=0}^2 n_i(p_i - p)(c_i - \bar{c})\right)^2}{pq \sum_{i=0}^2 n_i(c_i - \bar{c})^2}.$$

Thus to finish the proof we show that $\sum_{i=0}^2 n_i(p_i - p)(c_i - \bar{c}) = \sum_{i=0}^2 r_i(c_i - \bar{c})$:

$$\begin{aligned} \sum_{i=0}^2 n_i(p_i - p)(c_i - \bar{c}) &= \sum_{i=0}^2 (r_i - n_i p)(c_i - \bar{c}) \quad (\text{recall that } r_i = n_i p_i) \\ &= \sum_{i=0}^2 r_i c_i - \bar{c} \sum_{i=0}^2 r_i - p \sum_{i=0}^2 n_i c_i + p \bar{c} \sum_{i=0}^2 n_i \\ &= \sum_{i=0}^2 r_i c_i - \bar{c} r - p n \bar{c} + p \bar{c} n \quad (\text{recall that } n \bar{c} = \sum_{i=0}^2 c_i n_i) \\ &= \sum_{i=0}^2 r_i(c_i - \bar{c}). \end{aligned}$$

Remark 3.1.1 (On the Choice of Scores for the CATT). To test H_0 using a CATT, the score $c = (c_1, c_2, c_3)$ is assigned to the genotypes (aa, aA, AA) such that $0 \leq c_0 \leq c_1 \leq c_2$ or $c_2 \leq c_1 \leq c_0 \leq 0$. In a 1997 paper [46], Sasieni assigned $c = (0, 0, 1)$ to the recessive model, $c = (0, 1/2, 1)$ to the additive model, and $c = (0, 1, 1)$ to the dominant model. The intuition underlying the scores is the following: for the

recessive model, the relative risks of the genotypes aa and aA are the same, so the same score is assigned to aa and aA . Likewise, for the dominant model, the same score is assigned to aA and AA . For the additive model, the effect of aA should be the average of the effect of aa and AA .

3.1.3 Relationship between the Pearson and Cochran-Armitage statistics

Recall that, in the case of CATT applied to genetics, one assumes that a 2×3 contingency table represents the sampling of 3 independent binomial random variables of parameters (ρ_i, n_i) where ρ_i is the penetrance (the probability of being a case given the genotype).

CATT works by looking for the presence of a linear trend among penetrance, across the 3 genotypes. The trend test may give strong evidence of positive or increasing linear trends, of constant or stable trends over time, or of negative or decreasing trends. Cochran and Armitage used a linear model of the form $\rho_i = \alpha + \beta c_i$, fitted by the ordinary least squares method. Recall that $p_i = r_i/n_i$, $p = r/n$ and $\bar{c} = \sum_i n_i c_i / n$. The prediction equation for ρ_i , given by Agresti (section 5.3.5 [1]) is:

$$\hat{\rho}_i = p + b(c_i - \bar{c}),$$

where

$$b = \frac{\sum_i n_i (p_i - p)(c_i - \bar{c})}{\sum_i n_i (c_i - \bar{c})^2}.$$

To build the Cochran-Armitage statistic, Cochran partitioned the Brandt-Snedecor formula (3.1.3) of the Pearson Chi-square statistic. Indeed, in 1954, Cochran noted that the *Brandt-Snedecor formula* (3.1.3) decomposes into

$$T_{\chi^2} = T_{CA}(x) + T_{fit}, \tag{3.1.6}$$

where $T_{fit} = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i (p_i - \hat{\rho}_i)^2$.

To verify decomposition (3.1.6), we first note that

$$\begin{aligned} \sum_{i=0}^2 n_i (p_i - \hat{\rho}_i)^2 &= \sum_{i=0}^2 n_i ((p_i - p) - b(c_i - \bar{c}))^2 \\ &= \sum_{i=0}^2 n_i ((p_i - p)^2 - 2(p_i - p)b(c_i - \bar{c}) - b^2(c_i - \bar{c})^2) \\ &= \sum_{i=0}^2 n_i (p_i - p)^2 - 2b \sum_{i=0}^2 n_i (p_i - p)(c_i - \bar{c}) + b^2 \sum_{i=0}^2 n_i (c_i - \bar{c})^2. \end{aligned}$$

Hence, we have the following equality:

$$T_{CA}(x) + T_{fit} = T_{\chi^2} - \frac{2b^2}{p(1-p)} \sum_{i=0}^2 n_i (p_i - p)(c_i - \bar{c}) + \frac{2b^2}{p(1-p)} \sum_{i=0}^2 n_i (c_i - \bar{c})^2.$$

Thus, it is left to show that $\sum n_i (p_i - p)(c_i - \bar{c}) = b \sum n_i (c_i - \bar{c})^2$. But this is true by construction since the denominator of b is $\sum n_i (c_i - \bar{c})^2$ and the numerator of b is $\sum n_i (p_i - p)(c_i - \bar{c})$.

When the linear probability model holds, T_{fit} is asymptotically chi-squared with $df = 1$ and the statistic z^2 , based on $df = 1$, tests for a linear trend in the proportions. Note that an equivalent way to express the null hypothesis is to say that the slope β of the linear model is zero. That is, $H_0 : \beta = 0$.

3.2 The function T

In this section, we compare the Kantorovich-Rubinstein distance with both the Cochran-Armitage statistic and the Pearson statistic. To do so, we write these two statistics as follows.

Definition 3.2.1. Consider (\mathcal{X}, μ) , a probability space and define $L_2(\mu)_*$ and $L_1^+(\mu)$

as follow:

$$L_2(\mu)_* = \{f \in L_2(\mu) : f \neq \text{constant}\},$$

$$L_1^+(\mu) = \left\{f \in L_1(\mu) : f(x) \geq 0 \mu\text{-a.e.}, \text{ and } 0 \leq \|f\|_1 \leq 1\right\}.$$

We now define the application $T : L_2(\mu)_* \times L_1^+(\mu) \longrightarrow \mathbb{R}$ by:

$$T(c, \alpha) = \frac{\left(\int c(\alpha - m) d\mu\right)^2}{V_\mu(c)}, \quad (3.2.1)$$

where $V_\mu(c) = \int c^2 d\mu - \left(\int c d\mu\right)^2$ and $m = \int \alpha d\mu$.

Remark 3.2.2. If the function $\alpha \in L_1^+(\mu)$ is such that $0 \leq \alpha \leq 1$, then $T(c, \alpha) = T(c, 1 - \alpha)$.

Proof: (Remark 3.2.2)

It is enough to show that $\left(\int c(\alpha - m_\alpha) d\mu\right)^2 = \left(\int c((1 - \alpha) - m_{1-\alpha}) d\mu\right)^2$.

Since $m_{1-\alpha} = \int (1 - \alpha) d\mu = \int 1 d\mu - \int \alpha d\mu = 1 - m_\alpha$, we have:

$$\begin{aligned} \left(\int c((1 - \alpha) - m_{1-\alpha}) d\mu\right)^2 &= \left(\int c(1 - \alpha - (1 - m_\alpha)) d\mu\right)^2 \\ &= \left(\int c(-\alpha + m_\alpha) d\mu\right)^2 = \left(\int c(\alpha - m_\alpha) d\mu\right)^2. \end{aligned}$$

■

Remark 3.2.3. Consider $V_\mu(c) = \int c^2 d\mu - \left(\int c d\mu\right)^2$. Then,

- (i) $V_\mu(c) \geq 0$ for all $c \in L_2(\mu)$,
- (ii) For $c \in L_2(\mu)$, $V_\mu(c) = 0$ if and only if the $c = \text{const}$, μ -a.e.
- (iii) $\int c(\alpha - m) d\mu = 0$ if $c = \text{const}$, μ -a.e.

Proof: (Remark 3.2.3)

Recall the Cauchy-Schwarz inequality:

$$\left| \int fg \, d\mu \right|^2 \leq \int f^2 \, d\mu \int g^2 \, d\mu.$$

By the Cauchy-Schwarz inequality applied to $c \in L_2(\mu)$, we obtain:

$$(i) \int c^2 \, d\mu \int 1 \, d\mu - \left(\int c \, d\mu \right)^2 \geq 0$$

(ii) Moreover, since $\int 1 \, d\mu = \mu(\mathcal{X}) = 1$, then $\int c^2 \, d\mu \int 1 \, d\mu - \left(\int c \, d\mu \right)^2 = 0$ if and only if $\int c^2 \, d\mu = \left(\int c \, d\mu \right)^2$. That is, if and only if $V_\mu(c) = 0$.

(iii) Since $\mu(\mathcal{X}) = 1$ and $\int \alpha \, d\mu = m$, then $\int (\alpha - m) \, d\mu = \int \alpha \, d\mu - \int m \, d\mu = 0$. Thus, $\int c(\alpha - m) \, d\mu = c \int (\alpha - m) \, d\mu = 0$.

■

To better understand the behaviour of the function T , we now prove important properties:

Let $G \subset GL_2(\mathbb{R})$ be the group defined by

$$G = \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}; a \neq 0, b \in \mathbb{R} \right\}.$$

and its action on $L_2(\mu)_*$ given, for $g \in G$ and $c \in L_2(\mu)_*$, by

$$(gc)(x) = ac(x) + b, \text{ for all } x \in \mathcal{X}.$$

Note that if, for every $g \in G$, $c \in L_2(\mu)_*$, then $gc \in L_2(\mu)_*$.

Let \sim_G denote the equivalence relation on $L_2(\mu_*)$ given by

$$c_1 \sim_G c_2 \text{ if and only if } \exists g \in G \text{ such that } gc_1 = c_2.$$

Lemma 3.2.4. *Let $c_1, c_2 \in L_2(\mu)_*$ be such that $c_1 \sim_G c_2$. Then the two functions $T(c_1, \cdot)$ and $T(c_2, \cdot)$, from $L_1^+(\mu)$ to \mathbb{R}_+ are equal.*

Proof: (Lemma 3.2.4)

Let $c_1, c_2 \in L_2(\mu)_*$ and $g = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \in G$ be such that $gc_1 = c_2$. We show separately that $V_\mu(c_2) = a^2V_\mu(c_1)$ and that $\int c_2(\alpha - m) d\mu = a \int c_1(\alpha - m) d\mu$:

$$\begin{aligned} V_\mu(c_2) &= \int (ac_1 + b)^2 d\mu - \left(\int (ac_1 + b) d\mu \right)^2 \\ &= \int (ac_1)^2 d\mu + b^2 + 2b \int c_1 d\mu - \left(\int c_1 d\mu + b \right)^2 \\ &= a^2V_\mu(ac_1) + b^2 + 2b \int c_1 d\mu - b^2 - 2b \int c_1 d\mu = a^2V_\mu(c_1). \end{aligned}$$

$\int c_2(\alpha - m) d\mu = \int (ac_1 + b)(\alpha - m) d\mu = a \int c_1(\alpha - m) d\mu + b \int (\alpha - m) d\mu$. Since we know by Remark 3.2.3 that $\int (\alpha - m) d\mu = 0$, we obtain the desired result. ■

To prove the converse (Property 3.2.8), we will need the following two lemmas:

Lemma 3.2.5. *Let $c \in L_2(\mu)_*$. Then, there exists $\tilde{c} \in L_2(\mu)_*$ such that*

$$c \sim_G \tilde{c} \quad \text{and} \quad T(\tilde{c}, \alpha) = \left(\int \tilde{c}\alpha d\mu \right)^2,$$

for all $\alpha \in L_1^+(\mu)$.

Proof: (Lemma 3.2.5)

Define $g_1, g_2 \in G$ by

$$g_1 = \begin{pmatrix} 1 & -\int c d\mu \\ 0 & 1 \end{pmatrix}, \quad g_2 = \begin{pmatrix} a_2 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{where } a_2 = \left(\int (g_1 c)^2 d\mu \right)^{-1/2},$$

and set $\tilde{c} = g_2 g_1 c$. Then $\int \tilde{c} d\mu = 0$ and $\int (\tilde{c})^2 d\mu = 1$. Indeed,

$$\int g_1 c d\mu = \int c d\mu - \mu(\mathcal{X}) \int c d\mu = 0.$$

Therefore,

$$\int \tilde{c} d\mu = \int g_2 g_1 c d\mu = \frac{1}{a_2} \int g_1 c d\mu = 0, \text{ and } \int \tilde{c}^2 d\mu = \frac{1}{a_2^2} \int (g_1 c)^2 d\mu = 1. \text{ Thus } V_\mu(\tilde{c}) = 1.$$

Since $V_\mu(\tilde{c}) = 1$ and $\int \tilde{c} d\mu = 0$, we have $T(\tilde{c}, \alpha) = \left(\int \tilde{c} \alpha d\mu \right)^2$. ■

Lemma 3.2.6. *Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a finite probability space and let c_0 and c_1 be two measurable functions from \mathcal{X} to \mathbb{R} . Given $p \in (0, 1)$, define the set $L_\infty^+(\mu)_p$ by*

$$L_\infty^+(\mu)_p = \left\{ f \in L_\infty(\mu) : \int f d\mu = p, f(x) \geq 0, \mu\text{-a.e.} \right\}.$$

If, for all $f \in L_\infty^+(\mu)_p$,

$$\left(\int c_1 f d\mu \right)^2 = \left(\int c_0 f d\mu \right)^2, \text{ then either } c_1(x) = c_0(x) \text{ or } c_1(x) = -c_0(x), \mu\text{-a.e..}$$

As for any $B \in \mathcal{F}$, $\mu(B) > 0$, the function $f = \frac{p}{\mu(B)} \mathbb{1}_B \in L_\infty^+(\mu)_p$.

Lemma 3.2.6 is a direct consequence of the following lemma:

Lemma 3.2.7. *Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a probability space and let c_0 and c_1 be two measurable functions from \mathcal{X} to \mathbb{R} . If, for all $A \in \mathcal{F}$,*

$$\left(\int c_1 \mathbb{1}_A d\mu \right)^2 = \left(\int c_0 \mathbb{1}_A d\mu \right)^2, \text{ then either } c_1(x) = c_0(x) \text{ or } c_1(x) = -c_0(x), \mu\text{-a.e..}$$

Proof: (Lemma 3.2.7)

Let us define $Y = \{x \in \mathcal{X}; |c_1(x)| > |d(x)|\}$, $Z = \{x \in \mathcal{X}; |c_1(x)| < |d(x)|\}$ and

$\mathcal{X}_e = \mathcal{X} \setminus (Y \cup Z)$. We will show that $\mu(Y) = 0$. Note that, by interchanging the functions c_0 and c_1 in Y and Z , we can show that $\mu(Z) = 0$.

Let us partition Y as $\coprod_{i,j \in \{0,1\}} Y_{i,j}$ where

$$Y_{i,j} = \{x \in Y; \operatorname{sgn}(c(x)) = (-1)^i \text{ and } \operatorname{sgn}(d(x)) = (-1)^j\}$$

and show that $\mu(Y_{i,j}) = 0$, for all i, j .

Suppose that there exist i_0, j_0 such that $\mu(Y_{i_0, j_0}) > 0$ and denote, to simplify notation, Y_{i_0, j_0} by Y_0 . For $x \in Y_0$, $0 \leq (-1)^{j_0} d(x) < (-1)^{i_0} c(x)$. Therefore,

$$\int_{Y_0} (-1)^{j_0} d \, d\mu < \int_{Y_0} (-1)^{i_0} c \, d\mu \quad \text{and} \quad \left| \int_{Y_0} d \, d\mu \right| < \left| \int_{Y_0} c \, d\mu \right|$$

which is impossible since, by assumption, $\left(\int_{Y_0} d \, d\mu \right)^2 = \left(\int_{Y_0} c \, d\mu \right)^2$.

For $i, j \in \{0, 1\}$, set $\mathcal{X}_e^{ij} = \{x \in \mathcal{X}; (-1)^i c(x) = (-1)^j d(x) > 0\}$ and $\tilde{\mathcal{X}}_e = \{x \in \mathcal{X}; c(x) = d(x) = 0\}$. For $i \in \{0, 1\}$, let $\mathcal{X}_e^i = \{x \in \mathcal{X}; d(x) = (-1)^i c(x) \neq 0\}$. Then $\mathcal{X}_e^i = \mathcal{X}_e^{i, i+1} \coprod \mathcal{X}_e^{i+1, i}$, where the indices are computed modulo 2.

Claim 1 below completes the proof of the Lemma while Claim 2 is used in the proof of Claim 1.

Claim 1: If $\mu(\mathcal{X}_e^i) > 0$, then $\mu(\mathcal{X}_e^{i+1}) = 0$.

Claim 2: Let $k \in \{i, j\}$. If $\mu(\mathcal{X}_e^{i,j}) > 0$ (respectively, $\mu(\mathcal{X}_e^{k,k}) > 0$), then

$$\int_{E_{i,j}} c_1 \, d\mu = \epsilon \int_{E_{i,j}} c_0 \, d\mu,$$

where $E_{i,j} = \mathcal{X}_e^{i,j} \coprod \mathcal{X}_e^{k,k}$ and $\epsilon = (-1)^{i-j} = -1$ (respectively $\epsilon = (-1)^{k-k} = 1$).

Proof of Claim 2: By the assumption of the lemma,

$$\int_{E_{i,j}} c_1 \, d\mu = \epsilon \int_{E_{i,j}} c_0 \, d\mu, \quad \text{with } \epsilon \in \{\pm 1\}.$$

Then, as $\int_{E_{i,j}} c_1 d\mu = -\int_{\mathcal{X}_e^{i,j}} c_0 d\mu + \int_{\mathcal{X}_e^{k,k}} c_0 d\mu$, we obtain

$$(\epsilon + 1) \int_{\mathcal{X}_e^{i,j}} c_0 d\mu = (1 - \epsilon) \int_{\mathcal{X}_e^{k,k}} c_0 d\mu.$$

Then, if $\mu(\mathcal{X}_e^{i,j}) > 0$ (respectively $\mu(\mathcal{X}_e^{k,k}) > 0$), then $\epsilon = -1$ (respectively $\epsilon = 1$). Indeed, if not, we would have $\int_{\mathcal{X}_e^{i,j}} c_0 d\mu = 0$ (respectively $\int_{\mathcal{X}_e^{k,k}} c_0 d\mu = 0$), which is impossible.

Proof of Claim 1: As $\mathcal{X}_e^i = \mathcal{X}_e^{i,0} \amalg \mathcal{X}_e^{i+1,1}$, if $\mu(\mathcal{X}_e^i) > 0$, then either $\mu(\mathcal{X}_e^{i,0}) > 0$ or $\mu(\mathcal{X}_e^{i+1,1}) > 0$. To prove the claim, we then need to show that $\mu(\mathcal{X}_e^{i+1,0}) = \mu(\mathcal{X}_e^{i,1}) = 0$, as $\mathcal{X}_e^{i+1} = \mathcal{X}_e^{i+1,0} \amalg \mathcal{X}_e^{i,1}$.

We start with the case $i = 0$. If $\mu(\mathcal{X}_e^{k,k}) > 0$, for $k = 0$ or $k = 1$, then by Claim 2 applied to $\mathcal{X}_e^{j,j+1} \amalg \mathcal{X}_e^{k,k}$, for $j = 0$ or $j = 1$, we obtain

$$\int_{E_{j,j+1}} c_1 d\mu = \int_{E_{j,j+1}} c_0 d\mu \quad \text{with } E_{j,j+1} = \mathcal{X}_e^{j,j+1} \amalg \mathcal{X}_e^{k,k}.$$

$$\text{As } \int_{\mathcal{X}_e^{j,j+1}} c_1 d\mu = -\int_{\mathcal{X}_e^{j,j+1}} c_0 d\mu, \text{ then } \int_{E_{j,j+1}} c_0 d\mu = 0,$$

and therefore $\mu(\mathcal{X}_e^{j,j+1}) = 0$, for $j = 0$ or 1 ($\text{sgn}(c)$ is constant on $\mathcal{X}_e^{j,j+1}$).

For the case $i = 1$, ie. $\mu(\mathcal{X}_e^1) > 0$, then $\mu(\mathcal{X}_e^{1,0}) > 0$ or $\mu(\mathcal{X}_e^{0,1})$. Then, to show that $\mu(\mathcal{X}_e^{0,0}) = \mu(\mathcal{X}_e^{1,1}) = 0$, we apply, as above, Claim 2 to $\mathcal{X}_e^{j,j+1} \amalg \mathcal{X}_e^{k,k}$, for $j = 0$ or 1 and $k = 0$ or 1 . ■

To prove Property 3.2.8, we define, for $p \in (0, 1)$ fixed, the set $L_1^+(\mu)_p$ defined by

$$L_1^+(\mu)_p = \{f \in L_1(\mu); \int f d\mu = p, f(x) \leq 0, \mu\text{-a.e.}\}.$$

Consider, for $c \in L_2(\mu)_*$ fixed, the function $T_c : L_1^+(\mu) \rightarrow \mathbb{R}$ defined by $T_c(\alpha) =$

$T(c, \alpha)$. Let us denote by $T_{p,c}$ the restriction of T_c to $L_1^+(\mu)_p$.

Property 3.2.8. Let $c_0, c_1 \in L_2(\mu)_*$ and $p \in (0, 1)$ be fixed. If $T_{p,c_0} = T_{p,c_1}$, then, $c_0 \sim_G c_1$.

Proof: (Property 3.2.8)

Consider $c_0, c_1 \in L_2(\mu)_*$. By Lemma 3.2.5, we can suppose, without loss of generality, that

$$T(c_0, \alpha) = \left(\int c_0 \alpha \, d\mu \right)^2 \quad \text{and} \quad T(c_1, \alpha) = \left(\int c_1 \alpha \, d\mu \right)^2, \quad \text{for all } \alpha \in L_1^+(\mu)_p. \quad (3.2.2)$$

Since both equalities in (3.2.2) are true for all $\alpha \in L_1^+(\mu)_p$, they remain true for all $\alpha \in L_\infty^+(\mu)_\circ$.

As well, as seen in the proof of lemma 3.2.5, $\int c_i \, d\mu = 0$ and $\int c_i^2 \, d\mu = 1$ for $i = 0, 1$.

Now, by assumption, $T_{p,c_0}(\alpha) = T_{p,c_1}(\alpha)$, for all $\alpha \in L_1^+(\mu)_p$. Hence, we obtain that

$$\left(\int c_0 \alpha \, d\mu \right)^2 = \left(\int c_1 \alpha \, d\mu \right)^2, \quad \text{for all } \alpha \in L_\infty^+(\mu)_\circ.$$

Therefore, by Lemma 3.2.6, we know that either $c_1 = c_2$ or $c_1 = -c_2$. Thus $c_0 \sim_G c_1$. ■

From now on, for any $\alpha \in L_1^+(\mu)_p$ fixed, we study the function $T_\alpha : L^2(\mu)_* \rightarrow \mathbb{R}$ defined by $T_\alpha(c) = T(c, \alpha)$.

In Corollary (3.2.11), we show that, $\sup T_\alpha = T(\alpha, \alpha)$, where the supremum is taken over all $c \in L_2(\mu)_*$. Corollary (3.2.11) will be a consequence of propositions (3.2.9) and (3.2.10).

Proposition 3.2.9. Let $p \in (0, 1)$ be fixed and $\alpha \in L_1^+(\mu)_p$. Consider the function $b : L_2(\mu)_* \rightarrow \mathbb{R}$ defined by

$$b(c) = \frac{\int (\alpha - p)(c - \int c \, d\mu) \, d\mu}{\int (c - \int c \, d\mu)^2 \, d\mu}. \quad (3.2.3)$$

where $p = \int \alpha \, d\mu$. Then we have:

(i) If $d \in \mathbb{R}$ and $e \in \mathbb{R}^*$, then $b(c + d) = b(c)$ and $b(ec) = \frac{1}{e}b(c)$;

(ii) The denominator of $b(c)$ is equal to $V_\mu(c)$. That is,

$$\int (c - \int c \, d\mu)^2 \, d\mu = \int c^2 \, d\mu - \left(\int c \, d\mu \right)^2;$$

(iii) The numerator of $b(c)$ can be simplified to $\int c(\alpha - p) \, d\mu$;

(iv) $T(c, \alpha) = b^2(c)V_\mu(c)$.

Proof: (Proposition 3.2.9)

(i) The linearity of integrals makes this result clear.

(ii)

$$\begin{aligned} \int (c - \int c \, d\mu)^2 \, d\mu &= \int \left[c^2 + \left(\int c \, d\mu \right)^2 - 2c \left(\int c \, d\mu \right) \right] \, d\mu \\ &= \int c^2 \, d\mu + \left(\int c \, d\mu \right)^2 - 2 \left(\int c \, d\mu \right)^2 \\ &= \int c^2 \, d\mu - \left(\int c \, d\mu \right)^2 = V_\mu(c). \end{aligned}$$

(iii) $\int (\alpha - p)(c - \int c \, d\mu) \, d\mu = \int c(\alpha - p) \, d\mu - \int c \, d\mu \int (\alpha - p) \, d\mu$.
 Since $\int \alpha \, d\mu = p$, we have $\int (\alpha - p) \, d\mu = 0$.

(iv) A direct application of (ii) and (iii) yields the first equation of (iv). ■

Proposition 3.2.10. *Let $p \in (0, 1)$ be fixed and $\alpha \in L_1^+(\mu)_p \cap L_2^+(\mu)$. Consider the function $S_\alpha : L_2(\mu)_* \rightarrow \mathbb{R}$ defined by*

$$S_\alpha(c) = \int \left(\alpha - p - b(c) \left(c - \int c \, d\mu \right) \right)^2 \, d\mu. \quad (3.2.4)$$

Then we have:

- (i) $S_\alpha(c) \geq 0$;
- (ii) If $c_0 \sim_G c_1$, then $S_\alpha(c_0) = S_\alpha(c_1)$;
- (iii) $T_\alpha(\alpha) = T_\alpha(c) + S_\alpha(c)$.

Proof: (Proposition 3.2.10)

(i) $S_\alpha \geq 0$ since it is defined as an integral of a positive function with respect to a probability measure.

(ii) If $c_0 \sim_G c_1$, then $b(c_1) = \frac{1}{p}b(c_0)$ and $c_1 - \int c_1 \, d\mu = p \left(c - \int c \, d\mu \right)$. Hence $S_\alpha(c_0) = S_\alpha(c_1)$.

(iii)

$$\begin{aligned} S_\alpha(c) + T_\alpha(c) &= \int \left(\alpha - p - b(c) \left(c - \int c \, d\mu \right) \right)^2 \, d\mu + b^2(c) \int \left(c - \int c \, d\mu \right)^2 \, d\mu \\ &= \int \left((\alpha - p)^2 + b^2(c) \left(c - \int c \, d\mu \right)^2 - 2b(c)(\alpha - p) \left(c - \int c \, d\mu \right) \right) \, d\mu \\ &\quad + b^2(c) \int \left(c - \int c \, d\mu \right)^2 \, d\mu \\ &= \int (\alpha - p)^2 \, d\mu + 2b^2(c) \int \left(c - \int c \, d\mu \right)^2 \, d\mu \\ &\quad - 2b(c) \int (\alpha - p) \left(c - \int c \, d\mu \right) \, d\mu \end{aligned}$$

$$= \int (\alpha - p)^2 d\mu + 2b(c) \left(b(c) \int (c - \int c d\mu)^2 d\mu - \int c(\alpha - p) d\mu \right). \quad (3.2.5)$$

Now, recall from Proposition 3.2.9 that the function b can be written as

$$b(c) = \frac{1}{V_\mu(c)} \int c(\alpha - p) d\mu.$$

Hence Equation (3.2.5) becomes:

$$\int (\alpha - p)^2 d\mu + 2b(c) \left(b(c)V_\mu(c) - V_\mu(c)b(c) \right) = \int (\alpha - p)^2 d\mu.$$

Using formula (3.2.1) that defines T , we obtain the following:

$$T(\alpha, \alpha) = \frac{\left(\int \alpha(\alpha - p) d\mu \right)^2}{V_\mu(\alpha)} = \frac{\left(\int \alpha^2 d\mu - p \int \alpha d\mu \right)^2}{\int \alpha^2 d\mu - \left(\int \alpha d\mu \right)^2} = \int \alpha^2 d\mu - p^2,$$

To finish the proof, one needs to show that $T(\alpha, \alpha) = \int (\alpha - p)^2 d\mu$. By expanding $(\alpha - p)^2$, one obtains

$$\int (\alpha - p)^2 d\mu = \int \alpha^2 d\mu - 2p \int \alpha d\mu + p^2 \int d\mu = \int \alpha^2 d\mu - p^2.$$

■

Corollary 3.2.11. *Let $p \in (0, 1)$ be fixed and $\alpha \in L_1^+(\mu)_p \cap L_2^+(\mu)$. Then,*

$$\sup\{T_\alpha(c); c \in L_2(\mu)_*\} = T(\alpha, \alpha), \quad \text{where } T(\alpha, \alpha) = \|\alpha - p\|_2^2.$$

Proof: (Corollary 3.2.11)

Recall the function $S_\alpha : L_2(\mu)_* \rightarrow \mathbb{R}$ defined in Equation (3.2.4) in Proposition 3.2.10

and the function $b : L_2(\mu)_* \rightarrow \mathbb{R}$ defined in Equation (3.2.3) in Proposition 3.2.9. Evaluated for $c = \alpha$, we obtain $b(\alpha) = 1$ and thus $S_\alpha(\alpha) = 0$. Since, by Proposition 3.2.10, we know that $T_\alpha(c) = T_\alpha(\alpha) - S_\alpha(c)$ with $S(c, \alpha) \geq 0$, T_α attains its supremum at α .

Now we show that $T_\alpha(\alpha) = \|\alpha - p\|_2^2$. As shown in the proof of Proposition 3.2.10 part (iii), $T_\alpha(\alpha) = \int (\alpha - p)^2 d\mu$. ■

In the last part of this section, we study the function T_α in the following special case:

- (i) $(\mathcal{X}, \mu) = (\{x_0, x_1, x_2\}, \{\mu_0, \mu_1, \mu_2\})$ where $0 \leq \mu_i \leq 1$ and $\sum_{i=0}^2 \mu_i = 1$.
- (ii) $p \in (0, 1)$ and $\alpha = (\alpha_0, \alpha_1, \alpha_2)$ with $\sum_{i=0}^2 \alpha_i \mu_i = p$, $\alpha_i \geq 0$ and $\alpha \neq \text{constant}$.
If not, then $\alpha_0 = \alpha_1 = \alpha_2 = p$ and thus $T_\alpha(c) = 0, \forall c$.

The remark below shows that, without loss of generality, we can assume that $0 \leq \alpha_0 \leq \alpha_1 \leq \alpha_2$:

Remark 3.2.12. Let $\sigma \in \mathcal{S}_3$ be a permutation and ν be the probability measure $\nu = \mu \circ \sigma$ i.e. $\nu(j) = \mu(\sigma(j))$. Then,

- (i) $V_\nu(c \circ \sigma) = V_{\mu \circ \sigma}(c \circ \sigma) = V_\mu(c)$ where V_μ is as defined in Equation (3.2.1).

Here is the proof of (i):

$$\begin{aligned} V_{\mu \circ \sigma}(c \circ \sigma) &= \int (c \circ \sigma)^2(x) d\mu \circ \sigma(x) - \left(\int c \circ \sigma(x) d\mu \circ \sigma(x) \right)^2 \\ &= \int c^2(\sigma(x)) d\mu(\sigma(x)) - \left(\int c(\sigma(x)) d\mu(\sigma(x)) \right)^2 \\ &= \int c^2(y) d\mu(y) - \left(\int c(y) d\mu(y) \right)^2 = V_\mu(c) \end{aligned}$$

- (ii) $T(c \circ \sigma, \alpha \circ \sigma) = T(c, \alpha)$.

Indeed,

$$\begin{aligned} \int c \circ \sigma(\alpha \circ \sigma - p) d\mu \circ \sigma &= \int c(\sigma(x))(\alpha(\sigma(x)) - p) d\mu(\sigma(x)) \\ &= \int c(y)(\alpha(y) - p) d\mu(y). \end{aligned}$$

Case 1: $|\text{supp } \mu| = 3$.

Subcase (a): $c_o \neq c_2$.

In this case, there is 3 facts of interest:

- (i) $T(c, \alpha) = T((0, x, 1), \alpha)$
- (ii) $\max_c T(c, \alpha) = T((0, x^*, 1), \alpha)$ where $x^* = \frac{\alpha_1 - \alpha_o}{\alpha_2 - \alpha_o} \in [0, 1]$.
- (iii) If $\alpha_o = \alpha_1$, then $x^* = 0$, and $\max_c T(c, \alpha) = T((0, 0, 1), \alpha)$.
If $\alpha_o = \alpha_1$, then $x^* = 1$, and $\max_c T(c, \alpha) = T((0, 1, 1), \alpha)$.

Subcase (b): $c_o = c_2$.

In this case, there is 3 facts of interest:

- (i) $T(c, \alpha) = T((0, x, 0), \alpha)$ where $x \neq 0$ since c cannot be constant.

After computation, one obtains

$$T((0, x, 0), \alpha) = \frac{(\alpha_1 - p)^2 \mu_1^2}{\mu_1 - \mu_1^2} = \frac{(\alpha_1 - p)^2 \mu_1}{1 - \mu_1}.$$

- (ii) $\lim_{|x| \rightarrow \infty} T((0, x, 1), \alpha) = T((0, x, 0), \alpha)$.

Indeed, when $x \rightarrow \infty$,

$$T((0, x, 1), \alpha) = \frac{\left(x(\alpha_1 - p)\mu_1 + (\alpha_2 - p)\mu_2\right)^2}{x^2(\mu_1 + \mu_2) - (x\mu_1 + \mu_2)^2} \longrightarrow \frac{(\alpha_1 - p)^2 \mu_1^2}{\mu_1 - \mu_1^2}.$$

Case 2: $|\text{supp } \mu| = 2$

Without loss of generality, we can assume that $\text{supp } \mu = \{0, 1\}$ and that $c_o < c_1$. Moreover, by Lemma 3.2.4, we can assume that $c_o = 0$ and $c_1 = 1$. Under these

assumptions, we have

$$T(c, \alpha) = \frac{(\alpha_1 - p)^2 \mu_1}{1 - \mu_1}.$$

Since $\text{supp } \mu = \{0, 1\}$ we can write $\mu_o = 1 - \mu_1$ and therefore $p = \alpha_o \mu_o + \alpha_1(1 - \mu_o) = \alpha_1 + (\alpha_o - \alpha_1)\mu_o$. Replacing μ_o and p in T , yields

$$T(c, \alpha) = (\alpha_1 - \alpha_o)^2 \mu_o \mu_1. \quad (3.2.6)$$

3.3 Generalization of the Pearson and Cochran-Armitage Statistics

The purpose of this section is to construct the Generalized Cochran-Armitage statistic and the Generalized Pearson statistic using the function T constructed in Definition 3.2.1.

Definition 3.3.1 (Generalized Cochran-Armitage Statistic). *Let \mathcal{X} be a Borel space and $\tilde{\mu}_r, \tilde{\mu}_s$ be two positive measures on \mathcal{X} .*

Consider the two positive measures μ_r, μ_s defined by $\mu_r = \frac{\tilde{\mu}_r}{n}$ and $\mu_s = \frac{\tilde{\mu}_s}{n}$, where $n = \tilde{\mu}_r(\mathcal{X}) + \tilde{\mu}_s(\mathcal{X})$.

Note that the measure $\mu = \mu_r + \mu_s$ is a probability measure.

Let α_r, α_s be the respective Radon-Nikodym derivative of μ_r and μ_s with respect to μ . Then, we can define the generalized Cochran-Armitage statistic $\tilde{T}_{CA} : L_2(\mu)_ \rightarrow \mathbb{R}$ by*

$$\tilde{T}_{CA}(c) = \frac{n}{p_r p_s} T(c, \alpha_r),$$

where $p_r = \int \alpha_r d\mu$ and $p_s = \int \alpha_s d\mu$.

Note that, by definition of the Radon-Nikodym derivative, and as μ is a probability measure, $p_r = \mu_r(\mathcal{X})$ and $p_s = \mu_s(\mathcal{X})$. Hence $p_r + p_s = 1$. As well, $\alpha_r + \alpha_s = 1$ since $\alpha_r + \alpha_s = \frac{d(\mu_r + \mu_s)}{d\mu}$.

Remark 3.3.2. Using the same assumptions as in Definition 3.3.1, we notice that

$$\int c(\alpha_r - p_r) d\mu = \int c(p_s \alpha_r - p_r \alpha_s) d\mu.$$

Indeed,

$$\begin{aligned} \int c(p_s \alpha_r - p_r \alpha_s) d\mu &= \int c((1 - p_r)\alpha_r - p_r(1 - \alpha_r)) d\mu \\ &= \int c(\alpha_r - p_r \alpha_r - p_r + p_r \alpha_r) d\mu \\ &= \int c(\alpha_r - p_r) d\mu. \end{aligned}$$

Definition 3.3.3 (Generalized Pearson Statistic). *Using the same assumptions as in Definition 3.3.1, we can define the Generalized Pearson Statistic \tilde{T}_{χ^2} by*

$$\tilde{T}_{\chi^2} = \frac{n}{p_r p_s} T(\alpha_r, \alpha_r).$$

3.3.1 Standard Statistics: a Special Case of the Generalized Statistics

The purpose of this subsection is to study the Generalized Cochran-Armitage statistic and the Generalized Pearson statistic in the case of a three point space equipped with the discrete distance $1_{x \neq y}$. The generalized Cochran-Armitage statistic and the generalized Pearson statistic simplify to the well known Cochran-Armitage statistic and Pearson statistic, respectively.

Particular Case 3.3.4 (Cochran-Armitage). Let \mathcal{X} be a Borel space. Let us define two atomic probability measures

$$\mu_1 = \sum_{i=1}^n \mu_i^{(1)} \delta_{x_i} \quad \text{and} \quad \mu_2 = \sum_{i=1}^n \mu_i^{(2)} \delta_{x_i},$$

supported on the same countable subset S of \mathcal{X} , and such that $\mu = \mu_1 + \mu_2$ is a probability measure. The respective Radon-Nikodym derivative of μ_1 and μ_2 are defined by

$$\alpha_r(x_i) = \frac{\mu_r(x_i)}{\mu(x_i)} \quad \text{and} \quad \alpha_s(x_i) = \frac{\mu_s(x_i)}{\mu(x_i)}, \quad x_i \in S.$$

Then, the Generalized Cochran-Armitage statistic, defined in 3.3.1, can be written as

$$\tilde{T}_{CA}(c) = \frac{n}{p_r p_s} \frac{\left(\sum_{x_i \in S} c_i (p_s \mu_r(x_i) - p_r \mu_s(x_i)) \right)^2}{\sum_{x_i \in S} c_i^2 \mu(x_i) - \left(\sum_{x_i \in S} c(x_i) \mu(x_i) \right)^2},$$

where $p_r = \sum_{x_i \in S} \alpha_r(x_i) \mu(x_i) = \sum_{x_i \in S} \mu_r(x_i) = \mu_r(S)$ and, likewise, $p_s = \mu_s(S)$.

If μ_r and μ_s are supported on a three-point set $S = \{x_0, x_1, x_2\}$, we set $\mu_r(x_i) = r_i/n$, $\mu_r(S) = r/n$, $\mu_s(x_i) = s_i/n$, $\mu_s(S) = s/n$, $\mu(x_i) = n_i/n$ and $\mu(S) = 1$, and we observe that

$$\frac{n}{p_r p_s} = \frac{1}{n \frac{p_r}{n} \frac{p_s}{n}} = \frac{1}{npq}.$$

By doing so, we obtain the "classic" Cochran-Armitage statistic given in equation (3.1.4).

Particular Case 3.3.5 (Pearson). Using the same assumptions as in the particular case 3.3.4, and recalling the expression of $T(\alpha, \alpha)$ obtained in the proof of Corollary 3.2.11, the Generalized Pearson Statistic \tilde{T}_{χ^2} , defined in 3.3.3, can be written as

$$\tilde{T}_{\chi^2} = \frac{n}{p_r p_s} \sum_{x_i \in S} \left(\frac{\mu_r(x_i)}{\mu(x_i)} - p_r \right)^2 \mu(x_i).$$

Moreover, if μ_r and μ_s are supported on a three-point set $S = \{x_0, x_1, x_2\}$, and we set $\mu_r(x_i)$, $\mu_r(S)$, $\mu_s(x_i)$, $\mu_s(S)$, $\mu(x_i)$ and $\mu(S)$ as in Particular Case 3.3.4, we obtain the Brandt-Snedecor formula (3.1.3) of the "classic" Pearson statistic.

If μ_r and μ_s are supported on a two point set $S = \{x_0, x_1\}$, we recall the equation

(3.2.6) of $T(c, \alpha)$ and set $\mu_r(x_i)$, $\mu_r(S)$, $\mu_s(x_i)$, $\mu_s(S)$, $\mu(x_i)$ and $\mu(S)$ as above to obtain

$$T(c, \alpha) = \left(\frac{r_1/n}{n_1/n} - \frac{r_0/n}{n_0/n} \right)^2 \frac{n_0}{n} \frac{n_1}{n} = \frac{(s_0 r_1 - r_0 s_1)^2}{n_1 n_0} \frac{1}{n^2}$$

This is interesting as it means that T is equivalent to the odds ratio (OR) measure of association commonly used in case-control studies with two-by-two frequency tables. the OR represents between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. The OR measure is given by $r_0 s_1 / r_1 s_0$. If $OR = 1$, the exposure does not affect odds of outcome and if $OR \neq 1$, the exposure is associated with the odds of the outcome.

3.4 Relationship between the Kantorovich-Rubinstein Distance and Trend Tests

In the case of a metric space equipped with the k -discrete distance, we have already proved the following (see Theorem 2.4.5):

Theorem 3.4.1. *Let \mathcal{X} be a Polish space equipped with the k -discrete distance $k_{x \neq y}$. Let μ_1 and μ_2 be two Borel probability measures on \mathcal{X} . Then,*

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \frac{k}{2} \|\mu_1 - \mu_2\|_{TV}.$$

Recall that the Total Variation norm is related to the L_1 -norm:

Lemma 3.4.2. *Let μ_1 and μ_2 be two Borel probability measures on a Borel space \mathcal{X} absolutely continuous with respect to a Borel probability measure μ . Let $\alpha_1 = \frac{d\mu_1}{d\mu}$ and $\alpha_2 = \frac{d\mu_2}{d\mu}$ be the respective Radon-Nikodym derivatives of μ_1 and μ_2 . Then,*

$$\|\mu_1 - \mu_2\|_{TV} = \frac{1}{2} \|\alpha_1 - \alpha_2\|_1.$$

Note that $\nu = \mu_1 + \mu_2$ satisfies the condition of the lemma.

As a direct consequence of Theorem 2.4.5 and Lemma 3.4.2, we have the following lemma:

Lemma 3.4.3. *Let \mathcal{X} be a Polish space equipped with the k -discrete distance $k_{x \neq y}$. Let μ_1 and μ_2 be two probability measures on \mathcal{X} and α_1 and α_2 be their respective Radon-Nikodym derivatives, with respect to some Borel measure ν . Then,*

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \frac{k}{4} \|\alpha_1 - \alpha_2\|_1.$$

We now have all the necessary results to compare the Kantorovich-Rubinstein distance with both the Cochran-Armitage statistic and the Pearson statistic:

Proposition 3.4.4. *Let \mathcal{X} be a Polish space equipped with the k -discrete distance $k_{x \neq y}$ and $\tilde{\mu}_r, \tilde{\mu}_s$ two finite positive measures on \mathcal{X} .*

Consider the two positive measures μ_r, μ_s defined by $\mu_r = \frac{\tilde{\mu}_r}{n}$ and $\mu_s = \frac{\tilde{\mu}_s}{n}$, where $n = \tilde{\mu}_r(\mathcal{X}) + \tilde{\mu}_s(\mathcal{X})$.

Let α_r, α_s be the respective Radon-Nikodym derivative of μ_r, μ_s with respect to the measure $\mu = \mu_r + \mu_s$ and $p_r = \int \alpha_r d\mu, p_s = \int \alpha_s d\mu$. Then, the following inequalities hold:

$$np_r p_s \frac{4^2}{k^2} W_{\mathcal{X}}^2\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right) \leq \tilde{T}_{\chi^2} \leq \frac{4n}{k} \|\alpha_r - p_r\|_{\infty} W_{\mathcal{X}}\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right),$$

where \tilde{T}_{χ^2} denotes the Generalized Pearson Statistic as defined in Definition 3.3.3.

Proof: (Proposition 3.4.4)

By Lemma 3.4.3, we know that

$$W_{\mathcal{X}}\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right) = \frac{k}{4} \left\| \frac{1}{p_r} \alpha_r - \frac{1}{p_s} \alpha_s \right\|_1.$$

Recall that $p_r + p_s = 1$ and $\alpha_r + \alpha_s = 1, \mu$ -a.e. Then,

$$\left\| \frac{1}{p_r} \alpha_r - \frac{1}{p_s} \alpha_s \right\|_1 = \frac{1}{p_r p_s} \left\| (1 - p_r) \alpha_r - p_r (1 - \alpha_r) \right\|_1 = \frac{1}{p_r p_s} \|\alpha_r - p_r\|_1,$$

and therefore

$$\frac{4n}{k} W_{\mathcal{X}}\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right) = \frac{n}{p_r p_s} \|\alpha_r - p_r\|_1. \quad (3.4.1)$$

Now, we first show that

$$\tilde{T}_{\chi^2} \leq \frac{4n}{k} \|\alpha_r - p_r\|_{\infty} W_{\mathcal{X}}\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right). \quad (3.4.2)$$

Basic properties of L_p distances inequality, yield:

$$\|\alpha_r - p_r\|_2^2 = \int |\alpha_r - p_r|^2 d\mu \leq \|\alpha_r - p_r\|_{\infty} \int |\alpha_r - p_r| d\mu = \|\alpha_r - p_r\|_{\infty} \|\alpha_r - p_r\|_1.$$

By multiplying on both sides of the inequality by $\frac{n}{p_r p_s}$ and using equation (3.4.1), one obtains:

$$\frac{n}{p_r p_s} \|\alpha_r - p_r\|_2^2 \leq \frac{4n}{k} \|\alpha_r - p_r\|_{\infty} W_{\mathcal{X}}\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right).$$

By Corollary 3.2.11, we know that $\tilde{T}_{\chi^2} = \frac{n}{p_r p_s} \|\alpha_r - p_r\|_2^2$. Hence we obtain inequality (3.4.2). It is now left to show that

$$n p_r p_s \frac{4^2}{k^2} W_{\mathcal{X}}^2\left(\frac{1}{p} \mu_r, \frac{1}{q} \mu_s\right) \leq \tilde{T}_{\chi^2}.$$

Using Hölder inequality, we have: $\frac{n}{p_r p_s} \|\alpha_r - p_r\|_1^2 \leq \frac{n}{p_r p_s} \|\alpha_r - p_r\|_2^2 = T_{\chi^2}$. Using equation (3.4.1) on the lower bound gives:

$$\begin{aligned} \frac{n}{p_r p_s} \|\alpha_r - p_r\|_1^2 &= \frac{p_r p_s}{n} \frac{n^2}{p_r^2 p_s^2} \|\alpha_r - p_r\|_1^2 \\ &= \frac{p_r p_s}{n} \frac{(4n)^2}{k^2} W_{\mathcal{X}}^2\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right) \\ &= n p_r p_s \frac{4^2}{k^2} W_{\mathcal{X}}^2\left(\frac{1}{p_r} \mu_r, \frac{1}{p_s} \mu_s\right), \end{aligned}$$

which completes the proof. ■

Chapter 4

K-Lipschitz Maps and the Kantorovich-Rubinstein distance

This chapter is short and technical but very important as it is a building bloc for Chapter 7. All the results presented are new. We study the functorial properties of the Kantorovich-Rubinstein distance. The goal is to compare the Kantorovich-Rubinstein distances between two measures $\mu_1, \mu_2 \in P_{\mathcal{X}}$ and their respective push-forward measures $f(\mu_1), f(\mu_2) \in P_{\mathcal{Y}}$, for a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ with particular properties.

In the first section, we obtain interesting results when f is a k -Lipschitz function and when f is an isometry. These results are given in Theorem 4.1.1 and Proposition 4.1.9.

The second section considers the cases of canonical projection $\pi_a : \mathcal{X} \rightarrow \mathcal{X}_A$, where A is a Cartesian product of Polish spaces and \mathcal{X}_A is a canonical subset of \mathcal{X} . We compare the Kantorovich-Rubinstein distances between two measures $\mu_1, \mu_2 \in P_{\mathcal{X}}$ and their respective push-forward measures $\pi_A(\mu_1), \pi_A(\mu_2) \in P_{\mathcal{X}_A}$. It is a particular case of the first section as canonical projections are Lipschitz functions.

4.1 Functorial properties of the Kantorovich-Rubinstein distance

Following the notation introduced in Section 2.2, if \mathcal{X} is a Polish space and d is a lower semi-continuous distance on \mathcal{X} , we denote by $P_{\mathcal{X}}$ the (convex) set of all Borel probability measures ν on \mathcal{X} such that

$$\int_{\mathcal{X}} d(x, x_0) d\nu(x) < \infty,$$

for some (and therefore for any) $x_0 \in \mathcal{X}$.

Recall also that $W_{\mathcal{X}}$ denotes the Kantorovich-Rubinstein distance on $P_{\mathcal{X}}$ and that $\mathcal{D}_{\mathcal{X}}$ denotes the set of all lower semi-continuous metrics d on \mathcal{X} . A metric $d \in \mathcal{D}_{\mathcal{X}}$ does not necessarily define the topology on \mathcal{X} .

Theorem 4.1.1. *Let \mathcal{X} and \mathcal{Y} be two Polish spaces, equipped with $d_{\mathcal{X}} \in \mathcal{D}_{\mathcal{X}}$ and $d_{\mathcal{Y}} \in \mathcal{D}_{\mathcal{Y}}$, respectively. Let $k \in \mathbb{R}_+^*$ and $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$ be a k -Lipschitz, measurable mapping.*

Let μ_1, μ_2 be two probability measures in $P_{\mathcal{X}}$, and $\varphi(\mu_1), \varphi(\mu_2)$ their respective push-forward measures on \mathcal{Y} . Then, for $i = 1, 2$, $\varphi(\mu_i) \in P_{\mathcal{Y}}$ and

$$W_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2)) \leq k W_{\mathcal{X}}(\mu_1, \mu_2).$$

To prove Theorem 4.1.1, we will need Lemma 4.1.3. Let us first recall a notation:

Notation 4.1.2. For any two sets Z_1 and Z_2 , let π_j , $j = 1, 2$, denote the canonical projection from the Cartesian product $Z_1 \times Z_2$ onto Z_j .

Lemma 4.1.3. *Let \mathcal{X}_i and \mathcal{Y}_i , for $i = 1, 2$, be Borel spaces and $\varphi_i : \mathcal{X}_i \rightarrow \mathcal{Y}_i$ be Borel maps. Let $\varphi_1 \times \varphi_2 : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \mathcal{Y}_1 \times \mathcal{Y}_2$ denote the Borel map defined by $(\varphi_1 \times \varphi_2)(x_1, x_2) = (\varphi_1(x_1), \varphi_2(x_2))$.*

For $i = 1, 2$, let μ_i be a probability measure on \mathcal{X}_i and $\varphi_i(\mu_i)$ be its push-forward measure on \mathcal{Y}_i .

If $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$ is a coupling of μ_1 and μ_2 , then its push-forward measure $(\varphi_1 \times \varphi_2)(\vartheta)$ belongs to $\mathcal{N}(\varphi_1(\mu_1), \varphi_2(\mu_2))$.

Proof: (Lemma 4.1.3)

Since, $\pi_i \circ (\varphi_1 \times \varphi_2) = \varphi_i \circ \pi_i$, for $i = 1, 2$, we have

$$\pi_i\left((\varphi_1 \times \varphi_2)(\vartheta)\right) = \pi_i \circ (\varphi_1 \times \varphi_2)(\vartheta) = \varphi_i \circ \pi_i(\vartheta),$$

for any probability measure ϑ on $\mathcal{X}_1 \times \mathcal{X}_2$.

Then, if $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$, we have, for $i = 1, 2$,

$$\pi_i\left((\varphi_1 \times \varphi_2)(\vartheta)\right) = \varphi_i \circ \pi_i(\vartheta) = \varphi(\mu_i).$$

Hence $(\varphi_1 \times \varphi_2)(\vartheta) \in \mathcal{N}(\varphi_1(\mu_1), \varphi_2(\mu_2))$. ■

We can now prove Theorem 4.1.1:

Proof: (Theorem 4.1.1)

Let μ_1 and μ_2 be two probability measures in $P_{\mathcal{X}}$ and let $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$. As $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are lower semi-continuous,

$$\begin{aligned} \int_{\mathcal{Y}} d_{\mathcal{Y}}(\varphi(x_0), y) d\varphi(\mu_i)(y) &= \int_{\mathcal{X}} d_{\mathcal{Y}}(\varphi(x_0), \varphi(x)) d\mu_i(x) \\ &\leq k \int_{\mathcal{X}} d_{\mathcal{X}}(x_0, x) d\mu_i(x) < \infty, \end{aligned}$$

for any $x_0 \in \mathcal{X}$ fixed. Therefore, $\varphi(\mu_i) \in P_{\mathcal{Y}}$.

Since $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$,

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x, y) d\vartheta(x, y) &\leq \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x, x_0) d\vartheta(x, y) + \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x_0, y) d\vartheta(x, y) \\ &\leq \int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0) d\mu_1(x) + \int_{\mathcal{X}} d_{\mathcal{X}}(x_0, y) d\mu_2(y). \end{aligned}$$

Thus $d \in L_1(\vartheta)$. By definition of push-forward measures, we get:

$$\int_{\mathcal{Y} \times \mathcal{Y}} d_{\mathcal{Y}}(y_1, y_2) d(\varphi \times \varphi)(\vartheta)(y_1, y_2) = \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{Y}}(\varphi(x_1), \varphi(x_2)) d\vartheta(x_1, x_2).$$

Since φ is k -Lipschitz, we obtain:

$$\int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{Y}}(\varphi(x_1), \varphi(x_2)) d\vartheta(x_1, x_2) \leq k \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x_1, x_2) d\vartheta(x_1, x_2).$$

By Lemma 4.1.3, $(\varphi \times \varphi)(\vartheta)$ belongs to $\mathcal{N}(\varphi(\mu_1), \varphi(\mu_2))$ and therefore

$$\begin{aligned} W_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2)) &\leq \int_{\mathcal{Y} \times \mathcal{Y}} d_{\mathcal{Y}}(y_1, y_2) d(\varphi \times \varphi)(\vartheta)(y_1, y_2) \\ &\leq k \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x_1, x_2) d\vartheta(x_1, x_2). \end{aligned}$$

By definition of the Kantorovich-Rubinstein distance, $W_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2)) \leq kW_{\mathcal{X}}(\mu_1, \mu_2)$.

■

Let \mathcal{X} be a measurable space. Recall from Remark 2.4.8 that for any probability measures μ_1 and μ_2 on \mathcal{X} , the total variation distance $\|\mu_1 - \mu_2\|_{TV}$ has a coupling characterization given by

$$\|\mu_1 - \mu_2\|_{TV} = 2 \inf \mathbb{P}[X \neq Y],$$

where the infimum is over all couplings (X, Y) of (μ_1, μ_2) . We get the following useful result:

Theorem 4.1.4. *Let \mathcal{X} be a Polish space and $d \in D_{\mathcal{X}}$ be a bounded distance. Let μ_1 and μ_2 be two Borel probability measures on \mathcal{X} . Then,*

$$W_{\mathcal{X}}(\mu_1, \mu_2) \leq \Delta \|\mu_1 - \mu_2\|_{TV}, \text{ where } \Delta = \text{diam}(\mathcal{X}).$$

Proof: (Theorem 4.1.4)

Let $1_{x \neq y}$ be the discrete distance on \mathcal{X} . Then, for all $x, y \in \mathcal{X}$, $d_{\mathcal{X}}(x, y) \leq \Delta 1_{x \neq y}$ and therefore

$$\begin{aligned} W_{\mathcal{X}}(\mu_1, \mu_2) &= \inf_{\vartheta \in \mathcal{N}(\mu_1, \mu_2)} \int d(x, y) \, d\vartheta(x, y) \\ &\leq \Delta \inf_{\vartheta \in \mathcal{N}(\mu_1, \mu_2)} \int 1_{x \neq y}(x, y) \, d\vartheta(x, y) \\ &\leq \Delta \|\mu_1 - \mu_2\|_{TV}. \end{aligned}$$

■

In Theorem 4.1.1, we have compared the Kantorovich-Rubinstein distance between two measures and between their respective push-forward measures in the case of a k -Lipschitz map φ from \mathcal{X} to \mathcal{Y} . Now, we study two particular cases: when φ is an optimal function and when φ is an isometric map.

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and μ_1 and μ_2 be probability measures on \mathcal{X} . We recall that an optimal function $\varphi : \mathcal{X} \rightarrow \mathbb{R}$, is a 1-Lipschitz function that realizes the Kantorovich-Rubinstein distance. Hence, $\int_{\mathcal{X}} \varphi \, d(\mu_1 - \mu_2)$.

Proposition 4.1.5. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} and $f : \mathcal{X} \rightarrow \mathbb{R}$ be an optimal 1-Lipschitz function. Then,*

$$W_{\mathcal{X}}(\mu_1, \mu_2) = W_{\mathbb{R}}(f(\mu_1), f(\mu_2)),$$

where $f(\mu_1)$ and $f(\mu_2)$ are the respective push-forward measures of μ_1 and μ_2 on \mathbb{R} .

For the proof of Proposition 4.1.5, we first need the following lemma:

Lemma 4.1.6. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} and $f :$*

$\mathcal{X} \rightarrow \mathbb{R}$ be a measurable, 1-Lipschitz function. Then, the identity function $i : \mathbb{R} \rightarrow \mathbb{R}$, defined by $i(t) = t$, is in $L_1(|f(\mu_1) - f(\mu_2)|)$.

Proof: (Lemma 4.1.6)

Let Δ be the d -diameter of \mathcal{X} . By Lemma 2.3.3, $\text{diam}(f(\mathcal{X})) < \infty$. Therefore there exist $-\infty < a < b < \infty$ such that $f(\mathcal{X}) \subset [a, b]$. Hence, for any Borel set $A \subset [a, b]^c$, $f(\mu_1)(A) = f(\mu_2)(A) = 0$ and thus $|f(\mu_1) - f(\mu_2)|([a, b]^c) = 0$. Then, Therefore,

$$\int_{\mathbb{R}} |t| d|f(\mu_1) - f(\mu_2)|(t) = \int_a^b |t| d|f(\mu_1) - f(\mu_2)|(t) < \infty,$$

which completes the proof. ■

We can now give the proof of Proposition 4.1.5:

Proof: (Proposition 4.1.5)

As f is 1-Lipschitz and measurable, then, by Theorem 4.1.1, we have

$$W_{\mathbb{R}}(f(\mu_1), f(\mu_2)) \leq W_{\mathcal{X}}(\mu_1, \mu_2).$$

To prove the converse inequality, first recall that, by the Kantorovich-Rubinstein Duality Theorem 1.3.2,

$$W_{\mathbb{R}}(f(\mu_1), f(\mu_2)) = \sup \left\{ \int_{\mathbb{R}} \psi(t) d(f(\mu_1) - f(\mu_2))(t), \psi \in L_1(|f(\mu_1) - f(\mu_2)|), \|\psi\|_{Lip} \leq 1 \right\}.$$

Now, by the optimality of f and with a change of variable, we obtain:

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} f(x) d(\mu_1 - \mu_2)(x) = \int_{\mathbb{R}} t d(f(\mu_1) - f(\mu_2))(t).$$

As the identity function $i : \mathbb{R} \rightarrow \mathbb{R}$ is 1-Lipschitz and, by Lemma 4.1.6, is in $L_1(|f(\mu_1) - f(\mu_2)|)$, the proof is complete. ■

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and μ be a positive measure on \mathcal{X} . Recall (see for example Definition 6.7 in [44]) that μ is concentrated on $C \in \mathcal{B}$ if $\mu(B) = \mu(B \cap C)$, for all $B \in \mathcal{B}$, or equivalently, $\mu(B) = 0$, if $B \cap C = \emptyset$.

Proposition 4.1.7. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. For $i = 1, 2$, let μ_i be a probability measure on $(\mathcal{X}, \mathcal{B})$ concentrated on $C_i \in \mathcal{B}$. If $\vartheta \in \mathcal{N}(\mu_1, \mu_2)$ is a coupling of μ_1 and μ_2 , then ϑ is concentrated on $C_1 \times C_2$.*

Proof: (Proposition 4.1.7)

Let $A \in \mathcal{B} \times \mathcal{B}$ be contained in $(C_1 \times C_2)^c$. Let us show that $\vartheta(A) = 0$. Set $A_1 = A \cap (C_1^c \times \mathcal{X})$ and $A_2 = A \cap (C_2^c \times \mathcal{X})$. Therefore, $A_i \in \mathcal{B} \times \mathcal{B}$, $\pi_i(A_i) = \pi_i(A) \cap C_i^c$ and $\vartheta(A_i) \leq \vartheta(\pi_i^{-1}(\pi_i(A_i))) = \mu_i(\pi_i(A_i)) = \mu_i(\pi_i(A_i) \cap C_i^c) = 0$. Hence $\vartheta(A) \leq \vartheta(A_1) + \vartheta(A_2) = 0$. ■

The following proposition is the key technical tool of Theorem 4.1.9. To facilitate the understanding of the proof, we chose to denote by $\mathcal{N}_{\mathcal{X}}$ the set of all couplings of two measures both supported on \mathcal{X} . Likewise, $\mathcal{N}_{\mathcal{Y}}$ is the set of all couplings of two measures both supported on \mathcal{Y} .

Proposition 4.1.8. *Let \mathcal{X} and \mathcal{Y} be two Polish spaces and let $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$ be an injective measurable map.*

Let μ_1, μ_2 be two probability measures in $P_{\mathcal{X}}$. Then,

$$(\varphi \times \varphi)(\mathcal{N}_{\mathcal{X}}(\mu_1, \mu_2)) = \mathcal{N}_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2)).$$

Proof: (Proposition 4.1.8)

By Lemma 4.1.3, we have $(\varphi \times \varphi)(\mathcal{N}_{\mathcal{X}}(\mu_1, \mu_2)) \subset \mathcal{N}_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2))$.

To prove the other inclusion, let us first note that by Theorem 8.3.6 in [15], $\varphi(\mathcal{X}) \subset \mathcal{Y}$ is Borel and $\varphi : \mathcal{X} \rightarrow \varphi(\mathcal{X}) \subset \mathcal{Y}$ is a Borel isomorphism. If $\vartheta \in \mathcal{N}_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2))$, then by Proposition 4.1.7, ϑ is concentrated on $\varphi(\mathcal{X}) \times \varphi(\mathcal{X})$. Then, $(\varphi^{-1} \times \varphi^{-1})(\vartheta)$

is a probability measure on $\mathcal{X} \times \mathcal{X}$, belonging to $\mathcal{N}_{\mathcal{X}}(\mu_1, \mu_2)$. Indeed, denoting by $\pi_i^{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X}$ (respectively, $\pi_i^{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathcal{Y}$) the canonical projections, we have by Lemma 4.1.3, $\pi_i^{\mathcal{X}}(\varphi^{-1} \times \varphi^{-1})(\vartheta) = \varphi^{-1}(\pi_i^{\mathcal{Y}}(\vartheta)) = \varphi^{-1}(\varphi(\mu_i)) = \mu_i$. Then $(\varphi \times \varphi)((\varphi^{-1} \times \varphi^{-1})(\vartheta)) = \vartheta$. ■

We can now give the proof of Theorem 4.1.9:

Theorem 4.1.9 (1.176). *Let \mathcal{X} and \mathcal{Y} be two Polish spaces equipped with corresponding lower semi-continuous distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$. Let $k \in \mathbb{R}_+^*$ and $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$ be a measurable isometry.*

Let μ_1, μ_2 be two probability measures in $P_{\mathcal{X}}$, and $\varphi(\mu_1), \varphi(\mu_2)$ their respective push-forward measures on \mathcal{Y} . Then, $\varphi(\mu_i) \in P_{\mathcal{Y}}$, for $i = 1, 2$, and $W_{\mathcal{X}}(\mu_1, \mu_2) = W_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2))$.

Proof: (Theorem 4.1.9)

As in Theorem 4.1.1, $\varphi(\mu_i) \in P_{\mathcal{Y}}$, for $i = 1, 2$. By Proposition 4.1.7 and Proposition 4.1.8, $\vartheta \in \mathcal{N}_{\mathcal{X}}(\mu_1, \mu_2)$ if and only if $(\varphi \times \varphi)\vartheta \in \mathcal{N}_{\mathcal{Y}}(\varphi(\mu_1), \varphi(\mu_2))$ and

$$\begin{aligned} \int_{\mathcal{Y} \times \mathcal{Y}} d_{\mathcal{Y}}(y_1, y_2) d(\varphi \times \varphi)\vartheta(y_1, y_2) &= \int_{\varphi(\mathcal{X}) \times \varphi(\mathcal{X})} d_{\mathcal{Y}}(y_1, y_2) d(\varphi \times \varphi)\vartheta(y_1, y_2) \\ &= \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(\varphi(x_1), \varphi(x_2)) d\vartheta(x_1, x_2) \\ &= \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x_1, x_2) d\vartheta(x_1, x_2). \end{aligned}$$

Thus, by definition of the Kantorovich-Rubinstein distance, the proof is complete. ■

4.2 Projections and the Kantorovich-Rubinstein distance

After recalling the definition of the ℓ_p -product metric (for $1 \leq p \leq \infty$) given in Section 2.4.2, we introduce notations of canonical projections and state results needed in the following chapters.

Definition 4.2.1. *Let (\mathcal{X}_k, d_k) be r metric spaces and let $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ be the Cartesian product of these r metric spaces. For $p \in [1, +\infty)$, the p -product metric \tilde{d}_p is defined as the ℓ_p -norm of the r -vector of the distances d_i . That is :*

$$\tilde{d}_p(x, y) = \left(\sum_{i=1}^r d_i(x_i, y_i)^p \right)^{1/p} \quad \text{with } x, y \in \mathcal{X}.$$

For $p = \infty$, the ℓ_∞ -product metric is also called the sup metric and is defined as

$$\tilde{d}_\infty(x, y) = \max_{1 \leq i \leq r} d_i(x_i, y_i).$$

Lemma 4.2.2. *Let (\mathcal{X}_k, d_k) be r metric spaces and let $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ be the Cartesian product of these r metric spaces. For $1 \leq q < q' \leq \infty$, let \tilde{d}_q and $\tilde{d}_{q'}$ be two product metrics on the Cartesian product $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$.*

Then, the metrics \tilde{d}_q , $\tilde{d}_{q'}$ and \tilde{d}_∞ are equivalent.

Proof: (Lemma 4.2.2)

The fact that a product metric \tilde{d}_p ($1 \leq p \leq \infty$) is a metric is straight forward.

Now, recall that two metrics d and d' are said to be equivalent if there exist two positive constants c_1, c_2 such that, for all $(x, y) \in \mathcal{X} \times \mathcal{X}$, we have the inequalities $c_1 d(x, y) \leq d'(x, y) \leq c_2 d(x, y)$.

Clearly, $\tilde{d}_\infty \leq \tilde{d}_q$ and $\tilde{d}_\infty \leq \tilde{d}_{q'}$. Now, since $d_k(x_k, y_k) \leq \tilde{d}_\infty(x, y)$, for all $k = 1, \dots, r$, we have

$$\tilde{d}_q(x, y) \leq n^{1/q} \tilde{d}_\infty(x, y) \quad \text{and} \quad \tilde{d}_{q'}(x, y) \leq n^{1/q'} \tilde{d}_\infty(x, y).$$

Thus \tilde{d}_∞ is equivalent to \tilde{d}_q and $\tilde{d}_{q'}$, respectively.

We can then write $(1/r^{1/q})\tilde{d}_q(x, y) \leq \tilde{d}_{q'}(x, y) \leq r^{1/q'}\tilde{d}_q(x, y)$, showing that \tilde{d}_q and $\tilde{d}_{q'}$ are equivalent. ■

We now give the definition of a canonical projection as it is used in the rest of the chapter.

Definition 4.2.3. *Let (\mathcal{X}_k, d_k) be r metric spaces and let $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ be the Cartesian product of these r metric spaces. For any non-empty subset $A = \{i_1, i_2, \dots, i_a\} \subset \{1, \dots, r\}$, consider the Cartesian product $\mathcal{X}_A = \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_a}$.*

(i) *The canonical projection $\pi_A : \mathcal{X} \rightarrow \mathcal{X}_A$ is defined by $\pi_A(x) = (x_{i_1}, x_{i_2}, \dots, x_{i_a})$.*

(ii) *The ℓ_p -distance $\tilde{d}_{p,a}$ on \mathcal{X}_A is defined by*

$$\tilde{d}_{p,a}(x, y) = \left(\sum_{k=1}^a d_{i_k}(x_{i_k}, y_{i_k})^p \right)^{\frac{1}{p}}, \text{ for } 1 \leq p < \infty, \text{ and}$$

$$\tilde{d}_{\infty,a}(x, y) = \max_{1 \leq k \leq a} d_{i_k}(x_{i_k}, y_{i_k}).$$

The following result will be necessary in Theorem 4.2.5:

Lemma 4.2.4. *Let r be a positive integer. For $1 \leq k \leq r$, let \mathcal{X}_k be a Polish space and $d_k \in \mathcal{D}_{\mathcal{X}_k}$. Consider the product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ endowed with the product topology, and \tilde{d}_p , ($1 \leq p \leq \infty$), the ℓ_p -distance on \mathcal{X} . Then, \tilde{d}_p is lower semi-continuous distance on \mathcal{X} .*

Proof: (Lemma 4.2.4)

For $1 \leq k \leq r$, let $\pi_k : \mathcal{X} \rightarrow \mathcal{X}_k$ be the canonical projection and $d^k = d_k \circ \pi_k$. For any $\alpha \geq 0$,

$$F_\alpha = (d^k)^{-1}([0, \alpha]) = \mathcal{X}_1 \times \dots \times \mathcal{X}_{k-1} \times F_\alpha^k \times \mathcal{X}_{k+1} \times \dots \times \mathcal{X}_r,$$

where $F_\alpha^k = d_k^{-1}([0, \alpha])$.

As d_k is lower semi-continuous, F_α^k is a closed subset of \mathcal{X}_k and therefore F_α is closed in \mathcal{X} . Hence, d^k is lower semi-continuous. As the sum of lower semi-continuous functions and the product of positive lower semi-continuous functions are lower semi-continuous (see [21]), \tilde{d}_p is lower semi-continuous for $1 \leq p \leq \infty$. As the supremum of lower semi-continuous functions is lower semi-continuous (see [21]), \tilde{d}_∞ is also lower semi-continuous. ■

Theorem 4.2.5. *Let r be a positive integer. For $1 \leq k \leq r$, let \mathcal{X}_k be a Polish space and $d_k \in \mathcal{D}_{\mathcal{X}_k}$. Consider the Polish Cartesian product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$, and its subset $\mathcal{X}_A = \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_a}$, where $S = \{i_1, i_2, \dots, i_a\} \subset \{1, \dots, r\}$ is non-empty. Let μ_1 and μ_2 be two probability measures in $P_{\mathcal{X}}$, then the push-forward measures $\pi_A(\mu_1)$ and $\pi_A(\mu_2)$ are in $P_{\mathcal{X}_A}$ and*

$$W_{\mathcal{X}_A}(\pi_A(\mu_1), \pi_A(\mu_2)) \leq W_{\mathcal{X}}(\mu_1, \mu_2), \quad (4.2.1)$$

where $W_{\mathcal{X}}$ and $W_{\mathcal{X}_A}$ are the respective Wasserstein distances associated to $(\mathcal{X}, \tilde{d}_p)$ and $(\mathcal{X}_A, \tilde{d}_{p,a})$.

Proof: (Theorem 4.2.5)

As each \mathcal{X}_k is a Polish space and $d_k \in \mathcal{D}_{\mathcal{X}_k}$, it follows from Lemma 4.2.4 that $\tilde{d}_p \in \mathcal{D}_{\mathcal{X}}$ and $\tilde{d}_{p,a} \in \mathcal{D}_{\mathcal{X}_A}$, for $1 \leq p \leq \infty$.

Now, the projection $\pi_A : \mathcal{X} \rightarrow \mathcal{X}_A$ is open and therefore Borel measurable. Moreover, by definition of π_A and of the distances \tilde{d}_p and $\tilde{d}_{p,a}$, it is clear that $\|\pi_A\|_{\text{Lip}} \leq 1$. Indeed, we have $\tilde{d}_{p,a}(x, y) \leq \tilde{d}_p(x, y)$ since

$$\left(\sum_{k=1}^a d_{i_k}(x_{i_k}, y_{i_k})^p \right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^r d_i(x_i, y_i)^p \right)^{\frac{1}{p}}.$$

Thus, by Theorem 4.1.1, $\pi_A(\mu_i) \in P_{\mathcal{X}_A}$, for $i = 1, 2$, and Equation (4.2.1) is satisfied. ■

Corollary 4.2.6. *Let r be a positive integer. For $1 \leq k \leq r$, let \mathcal{X}_k be a Polish space and $d_k \in \mathcal{D}_{\mathcal{X}_k}$. Let $S = \{i_1, \dots, i_a\} \subset B = \{j_1, \dots, j_b\}$ be non-empty subsets of $\{1, \dots, r\}$. Keeping the notation of Theorem 4.2.5, we denote by $\mathcal{X}_A = \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_a}$, $\mathcal{X}_B = \mathcal{X}_{j_1} \times \dots \times \mathcal{X}_{j_b}$ and $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$, the respective Polish Cartesian product spaces.*

If μ_1 and μ_2 are two probability measures on $P_{\mathcal{X}}$, then the push-forward measures $\pi_A(\mu_i)$ and $\pi_B(\mu_i)$ are respectively in $P_{\mathcal{X}_A}$ and $P_{\mathcal{X}_B}$, and

$$W_{\mathcal{X}_A}(\pi_A(\mu_1), \pi_A(\mu_2)) \leq W_{\mathcal{X}_B}(\pi_B(\mu_1), \pi_B(\mu_2)),$$

where $W_{\mathcal{X}_A}$ and $W_{\mathcal{X}_B}$ are the respective Kantorovich-Rubinstein distances associated to $(\mathcal{X}_A, \tilde{d}_{p,a})$ and $(\mathcal{X}_B, \tilde{d}_{p,b})$.

Proof: (Corollary 4.2.6)

Let $\pi_{AB} : \mathcal{X}_B \rightarrow \mathcal{X}_A$ denote the canonical projection from \mathcal{X}_B to \mathcal{X}_A . Then, by Lemma 4.2.4, $\tilde{d}_{p,a} \in \mathcal{D}_{\mathcal{X}_A}$ and $\tilde{d}_{p,b} \in \mathcal{D}_{\mathcal{X}_B}$. As in Theorem 4.2.5, π_{AB} is measurable and $(\tilde{d}_{p,b}, \tilde{d}_{p,a})$ -Lipschitz. Since $\pi_A = \pi_{AB} \circ \pi_B$, Theorem 4.2.5 yields

$$\begin{aligned} W_{\mathcal{X}_A}(\pi_A(\mu_1), \pi_A(\mu_2)) &= W_{\mathcal{X}_A}(\pi_{AB}(\pi_B(\mu_1)), \pi_{AB}(\pi_B(\mu_2))) \\ &\leq W_{\mathcal{X}_B}(\pi_B(\mu_1), \pi_B(\mu_2)), \end{aligned}$$

which completes the proof. ■

Chapter 5

An Introduction to Classification problems

5.1 Statistical Machine Learning

Machine learning is the field encompassing the study, the conception and the implementation of computer algorithms that can learn. In his book [38], T. M. Mitchell provides the following definition of *learning* :

“An entity is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .”

In order to develop, compare and ultimately improve learning algorithms, one needs a framework to characterise the mathematical representation of the experiences E , and to mathematically define the class of tasks T and the notion of performance measures P . *Statistical learning theory* provides such a framework. Note that statistical learning theory is a framework built for the scenario of *supervised learning*. According to Vapnik (one of the key contributors of the domain), the idea is to consider learning problems through the statistical framework of minimising the expected value of a carefully chosen loss function using labeled data. Hence, statistical learning theory considers machine learning tasks from the perspective of both statistical

inference and optimization theory. The following description of statistical learning is found in Vapnik's *The Nature of Statistical Learning Theory* [50]. It requires three assumptions:

1. a measure space \mathcal{X} of observations (also called an input space) and the existence of a generator of observations, drawn independently from the same unknown probability measure P .
2. a (measurable) output space \mathcal{Y} and the existence of a supervisor that returns an output y for every input x , according to a fixed but unknown conditional distribution $P(\cdot | x)$.
3. the existence of a learning machine capable of implementing any function $g \in \mathcal{G}$ where \mathcal{G} is a fixed subset of all measurable functions mapping \mathcal{X} to \mathcal{Y} . The function g is called a *learning function* and the set \mathcal{G} is called a *model*.

Learning is to choose, from a given model \mathcal{G} , the function g which best predicts the supervisor's response. In order to choose the best available approximation to the supervisor's response, one measures the loss (or discrepancy) $L(y, g(x))$ between the response y of the supervisor to a given input x and the response $g(x)$ provided by the learning machine. Consider the expected value of the loss, given by the *risk functional*:

$$R(g) = \int L(y, g(x)) dP(x, y). \quad (5.1.1)$$

The goal is to find the function g_o which minimises the risk functional R (over the model \mathcal{G}) in the situation where the joint probability distribution $P(x, y)$ is unknown and the only available information is contained in a sample of n independent and identically distributed observations on $\mathcal{X} \times \mathcal{Y}$.

5.2 Classification

Classification is the assignment of any given observation $x \in \mathcal{X}$ to one of m classes ($m \in \mathbb{N}^*$). An observation that has been assigned to a particular class c is said to

be labeled. Hence, to classify is to create a measurable function $g : \mathcal{X} \rightarrow \{1, \dots, m\}$, where \mathcal{X} is the space of observations and $\mathcal{Y} = \{1, \dots, m\}$ is the output space (or label space). Such a measurable function g is called a *classifier*. Given a classifier g , an observation $x \in \mathcal{X}$ is well-classified if $g(x)$ matches the label $y \in \mathcal{Y}$ associated to x . A *classification error* (also known as a *misclassification*) occurs if $g(x) \neq y$. The performance of a classifier g is measured by its accuracy or, conversely, its probability of misclassification. The accuracy of a classifier is the probability that it assigns a given observation $x \in \mathcal{X}$ to the correct class c . Conversely, the classification error is the probability that the classifier labels an observation incorrectly.

To formalize the definition of a classification problem, we consider a probability measure P on the product $\mathcal{X} \times \mathcal{Y}$ of the input space \mathcal{X} and the output space \mathcal{Y} and a family \mathcal{G} of classifying functions mapping \mathcal{X} to \mathcal{Y} .

The problem of classification is to find a classifier $g_o \in \mathcal{G}$ which minimises the probability of misclassification $P(g(x) \neq y)$ in the situation where P is unknown but an independent and identically distributed sample $S_n = ((x_1, y_1), \dots, (x_n, y_n))$ of n labeled observations is given.

For the purpose of this thesis, we limit ourselves to the case of binary classification. That is, we consider a label space \mathcal{Y} of cardinality 2. By convention, we identify \mathcal{Y} with the set $\{-1, 1\}$. Hence, for A good reference covering the theoretical material of multi-class classification is the chapter 8 of *Foundations of Machine Learning* [39].

5.3 Bayes Error and the Estimation Error

Over all measurable classifiers, the infimum of the classification errors is defined as the Bayes error and is denoted by R^* . One can construct the unique measurable classifier that attains the Bayes error. It is called the Bayes classifier and is denoted

by g^* . One can show that the Bayes classifier g^* is defined as follow:

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

where $\eta(x)$ is the conditional probability that the label y is equal to 1, given an observation x .

Since a given model \mathcal{G} is by construction a subset of all measurable functions, *the excess error* of a classifier $g \in \mathcal{G}$ is defined as the discrepancy between the misclassification error of $g \in \mathcal{G}$ and the Bayes error. Thus, the problem of classification defined in section 5.2 is equivalent to the minimization of the excess error over all $g \in \mathcal{G}$. For $g \in \mathcal{G}$, the difference $R(g) - R^*$ can be decomposed as follow:

$$R(g) - R^* = R(g) - \inf_{g \in \mathcal{G}} R(g) + \inf_{g \in \mathcal{G}} R(g) - R^*.$$

The first difference on the right hand side of the equal sign is referred to as the estimation error while the second difference is known as the approximation error. The estimation error measures the quality of the classifier g with respect to the optimal misclassification error of \mathcal{G} while the approximation error measures how well the optimal classification error of \mathcal{G} can approximate the Bayes error. The decomposition of the excess error in terms of the estimation and approximation errors shows that the choice of the model \mathcal{G} is subject to a trade-off: a rich model is more likely to have a small approximation error but at the price of a larger estimation error, and vice-versa. Note that since both the joint probability distribution $P(x, y)$ and the conditional distribution $P(\cdot | x)$ are unknown, both g^* and $R(g)$ are unknown. Even with various noise assumptions, estimating the approximation error is difficult [39]. On the other hand, the estimation error can be bounded.

5.4 Representational Capacity of a Model and Generalization Ability of a Classifier

As written in the subsection 5.3, the focus of statistical learning theory is to bound the estimation error $R(g) - \inf_{g \in \mathcal{G}} R(g)$. The difficulty of this task stems from the fact that the only information we have is contained in the training set of n unlabeled observations drawn independently according to the joint probability measure P on $\mathcal{X} \times \mathcal{Y}$. Thus, as said above, for a classifier $g \in \mathcal{G}$, one cannot obtain $R(g)$ and thus cannot obtain $\inf_{g \in \mathcal{G}} R(g)$ either. One can only measure the agreement of g with the n points in the training set S_n . The standard measurement of agreement is the empirical probability of misclassification:

$$\hat{R}_S(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{g(X_i) \neq Y_i}.$$

Once we compute the empirical probability of misclassification $\hat{R}_S(g)$ we need a guarantee that the “general” probability of misclassification $R(g)$ is close to $\hat{R}_S(g)$. A classifier with a small discrepancy between its empirical error and its general misclassification error $R(g)$ is said to have a high generalization ability. The problem is that the generalization ability of a particular classifier $g \in \mathcal{G}$ is unknown since the general misclassification error is unknown. Therefore, a lot of effort has been spent to construct quantitative predictions on the discrepancy between $\hat{R}_S(g)$ and $R(g)$ over all $g \in \mathcal{G}$. These predictions are based on the concept of *representational capacity* (often just called *capacity*). Informally, the representational capacity of a model \mathcal{G} tries to capture the idea that \mathcal{G} contains classifiers nimble enough to mimic well the unknown classification. Said differently, the capacity of a model \mathcal{G} tries to capture the model’s ability to fit a wide variety of classifications. It is by properly quantifying the capacity of the model \mathcal{G} that statistical learning theory was able to construct, over all $g \in \mathcal{G}$, uniform upper bounds for the difference $R(g) - \hat{R}_S(g)$. These upper bounds are functions of both the cardinality of the training set S_n of the training

set and values that measure the capacity of the model \mathcal{G} . The general form of most generalisation bounds is composed of three different terms and has the following form:

With probability at least $1 - \delta$,

$$R(g) \leq \hat{R}_S(g) + \text{capacity}(\mathcal{G}) + \text{confidence}(n, \delta).$$

The central results in statistical learning theory show that for a given model, the upper bound decreases as the cardinality n of the training set increases while for a given training set, the upper bound is larger for a model with greater capacity. Hence, a given model \mathcal{G} of high capacity increases the risk that a classifier $g \in \mathcal{G}$ has a low generalization ability. On the other hand, a model \mathcal{G} of low capacity increases the risk that a classifier $g \in \mathcal{G}$ has a high generalization ability.

For binary classification problems, the most well known means of quantifying representational capacity is the Vapnik-Chervonenskis dimension (VC dimension). The VC dimension is defined as being the largest possible integer l for which there exists a training set of l different points in \mathcal{X} that the model \mathcal{G} can label arbitrarily. An important aspect of the VC dimension is that it is independent of the distribution P and thus the same upper bound holds for any distribution. The drawback is that the bound may be loose for most distributions. In the early 2000's, several authors considered alternative notions of capacity such as maximum discrepancy, Rademacher averages and Gaussian averages. These new(er) notions of capacity are dependent on the distribution P but give sharper upper bounds and have properties that make their computation possible from the training set only. We study Rademacher averages in more details later in the thesis.

5.5 Underfitting, Overfitting and Regularisation

As explained in the previous sections, the goal of classification is to minimise the probability of misclassification $R(g)$. For a given model \mathcal{G} , we focus on minimising the estimation error even though we cannot compute $R(g)$ as we do not know the

underlying measure of probability. We thus try to infer a classifier $g_s \in \mathcal{G}$, based on the training sample S_n , whose probability of error $R(g_s)$ is close to $\inf_{g \in \mathcal{G}} R(g)$. The most intuitive way to find such a classifier g_s is by:

1. replacing the “general” probability of misclassification R by the empirical probability of misclassification \hat{R}_S , constructed on the basis of the training set S_n .
2. approximating the function g_o that minimises R over \mathcal{G} by g_s that minimises $\hat{R}_S(g)$ over G .

Using the empirically optimal classifier g_s to approximate g_o is known as the *empirical risk minimization* inductive principle (ERM).

Once we compute g_s , there are two factors determining how close the misclassification error $R(g_s)$ is from $\inf_{g \in \mathcal{G}} R(g)$:

1. The empirical misclassification error \hat{R}_S of g_s .
2. The generalization ability of g_s .

These two factors are linked to the two central concerns in the field of Machine Learning: underfitting and overfitting. Underfitting occurs when the classifier $g_s \in \mathcal{G}$ is such that $\hat{R}_S(g_s)$ is not sufficiently small. Overfitting occurs when the difference between the empirical error $\hat{R}_S(g)$ and the misclassification error $R(g)$ is big. The fundamental results of Statistical Learning Theory show that one can control how likely it will be that g_s underfits or overfits by altering the capacity of the model \mathcal{G} . As seen in the last section, if the model is of large capacity, it is more likely that the empirically optimal classifier g_s overfits while for a small capacity model it is less likely that g_s overfits. On the other hand, a small capacity model increases the risk of underfitting while a model with a large capacity C decreases the risk of underfitting. Thus, there is a tradeoff to make: a larger capacity decreases the risk of underfitting but increases the risk of overfitting while a smaller capacity decreases the risk of overfitting but increases the risk of underfitting.

In order to strike the right balance for the model’s capacity, Vapnik and Chervonenkis developed the concept of *Structural Risk Minimization* (SRM) [51]. SRM is intended to minimise the classification error with respect to both the empirical error \hat{R}_s and

the capacity of the model used. The algorithm is defined as follow: Consider a nested sequence $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \dots\}$ of models with respective capacities C_1, C_2, C_3, \dots , such that the countable union of all \mathcal{G}_i 's is dense in \mathcal{G} . For each i , consider $g_s^{(i)}$, the empirically optimal classifier over the classe \mathcal{G}_i . We select the classifier g_s° minimising the capacity-penalized empirical error:

$$g_s^\circ = \underset{i \in \mathbb{N}}{\text{Argmin}} \left(g_s^{(i)} + r(n, C_i) \right),$$

where the penalty $r(n, C_i)$ can be understood as an estimate of the overfitting magnitude of $g_s^{(i)}$. As such, r depends on the capacity C_i of \mathcal{G}_i . Thus, SRM identifies an optimal model \mathcal{G}_{i_\circ} and returns the classifier $g_s^{(i_\circ)}$ that minimises its empirical error. While SRM has a strong theoretical footing [52, 20], it is often computationally very expensive as it requires determining the solution of multiple empirical risk minimization problems. It is therefore rarely used.

Today, rather than SRM, the approach used to minimise the classification error is *regularisation*. Regularisation based algorithms are inspired by SRM based algorithms but are more general. A regularisation algorithm is defined as follow: Consider an uncountable union of nested models \mathcal{G}_j , $j \in \mathcal{J}$ with respective capacities C_j , such that the union \mathcal{G} of all \mathcal{G}_j 's is dense in the space of continuous functions over \mathcal{X} . As for SRM, we select the classifier g_s° minimising the constrained optimization problem of the form:

$$\min_{j \in \mathcal{J}, g \in \mathcal{G}_j} \left(g_s^{(j)} + r(n, C_j) \right), \quad (5.5.1)$$

where the penalty $r(n, C_j)$ has the same signification than in the case of SRM.

Under quite general assumptions, there exist a function $\Gamma : \mathcal{G} \rightarrow \mathbb{R}$ such that the constrained optimization problem (5.5.1) can be equivalently written as the unconstrained optimization problem:

$$g_s^\circ = \underset{g \in \mathcal{G}}{\text{Argmin}} \left(\hat{R}(g) + \lambda \Gamma(g) \right),$$

where $\lambda \in [0, +\infty)$ is the regularisation parameter. The regularisation parameter λ

is treated as an hyperparameter since its optimal value is not known. Note that the value of λ is set prior to solving the constrained optimization problem. The function Γ is called a *regularizer*. A regularizer (also known as a regularisation term) is a quantity that penalizes one classifier over another in the model \mathcal{G} . Hence, for two classifiers in \mathcal{G} , both are eligible as solutions of the optimization problem but one classifier is preferred. The penalized function is chosen only if it classifies the training data significantly better than the preferred function. Often, the regularizer Γ is designed to express a generic preference for a simpler model class in order to promote generalization. As noted by Bengio *et al.* in *Deep Learning* [7], regularisation is a more general way of controlling a model's capacity than SRM. Penalizing one classifier over another is more general than excluding subsets of classifiers from the model \mathcal{G} . One can think of removing a classifier from the model \mathcal{G} as expressing an infinitely strong penalty against that classifier. The hyperparameter λ weights the relative contribution of the regularisation term $\Gamma(g)$ with respect to the empirical misclassification error. Setting $\lambda = 0$ means no regularisation and increasing λ corresponds to more regularisation.

5.6 Loss Functions and Confidence Margins

A quick introduction to *the problem of learning* was given in section 5.1. As stated in that section, we use the definition of the learning problem given by Vapnik in *The Nature of Statistical Learning Theory* [50]. In his book, Vapnik presents learning problems as particular cases of minimising the risk functional on the basis of empirical data:

Let $(\mathcal{Z}, \mathcal{B})$ be a measurable space and μ be a probability measure on \mathcal{Z} . Consider a set Λ of μ -integrable random variables $Q : \mathcal{Z} \rightarrow \mathbb{R}_+$. For a given random variable $Q \in \Lambda$, the risk functional R is the expected value of Q . The goal is to minimise the risk functional R over all $Q \in \Lambda$, where the probability measure μ is unknown but a independent and identically distributed sample z_1, \dots, z_n is given.

To obtain the particular case of learning problems from the minimising of risk func-

tional on the basis of empirical data, one considers a family \mathcal{H} of measurable functions $h : \mathcal{X} \rightarrow \mathcal{Y}$ and one loss function $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$. Then, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and the μ -integrable random variables $Q : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ are defined by $Q(x, y) = L(h(x), y)$. The goal of a learning problem is to find the learning function h_0 that minimises the risk functional

$$R(h) = \int_{\mathcal{X} \times \mathcal{Y}} L(h(x), y) d\mu(x, y)$$

where the probability measure μ on $\mathcal{X} \times \mathcal{Y}$ is unknown but a independent and identically distributed sample $(x_1, y_1), \dots, (x_n, y_n)$ is given.

Let us now formalise the definition of a loss function L .

Definition 5.6.1 (Loss function). *Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{B})$ be a measurable space, μ be a probability measure on $\mathcal{X} \times \mathcal{Y}$ and let \mathcal{H} be a family of measurable functions $h : \mathcal{X} \rightarrow \mathcal{Y}$. A map $L : \mathcal{Y} \times \mathcal{Y}' \rightarrow \mathbb{R}_+$, with $\mathcal{Y}' \subset \mathcal{Y}$, is a loss function for \mathcal{H} if:*

(i) *the composed function*

$$\begin{aligned} L_h : \mathcal{X} \times \mathcal{Y}' &\rightarrow \mathbb{R}_+ \\ (x, y') &\mapsto L(h(x), y') \end{aligned}$$

is μ -integrable for all $h \in \mathcal{H}$;

(ii) *For all $h \in \mathcal{H}$, for all $(x, y) \in D_h \times I_h$, $h(x) = y$ implies $L(h(x), y) = 0$.*

Remark 5.6.2. Definition 5.6.1 of a loss function is found in *Learning with Kernels* by Schölkopf and Smola [47]. The authors specify that it is possible to relax the non-negativity of L . Indeed, it is enough for the image of L to be bounded from below. In that case, an appropriate translation would recover the non-negativity. Likewise, it is possible to relax condition (ii) because an appropriate translation would recover the condition that exact predictions have a loss of zero.

The definition of *classification* given in section 5.2 can be expressed as a particular case of a *learning problem*:

Consider a model \mathcal{G} of measurable binary classifiers $g : \mathcal{X} \rightarrow \{-1, 1\}$ and the zero-one loss function $L_{z-o} : \{-1, 1\} \times \{-1, 1\} \rightarrow \mathbb{R}_+$ defined by:

$$L_{z-o}(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 = y_2 \\ 1 & \text{if } y_1 \neq y_2 \end{cases}$$

Note that the zero-one loss function can also be written as $L_{z-o}(y_1, y_2) = \mathbb{1}_{y_1 \neq y_2}$.

The goal is thus to find a classifier g_o that minimises the risk functional $R(g)$. But $R(g)$ can be simplified:

$$\begin{aligned} R(g) &= \int_{\mathcal{X} \times \{-1, 1\}} L_{z-o}(g(x), y) \, d\mu(x, y) \\ &= \int_{\mathcal{X} \times \{-1, 1\}} \mathbb{1}_{g(x) \neq y} \, d\mu(x, y) \\ &= \mu(\{(x, y) \in \mathcal{X} \times \{-1, 1\} : g(x) \neq y\}). \end{aligned} \tag{5.6.1}$$

We thus see from the equalities above that if we incur a loss of 1 to any couple $(x, y) \in \mathcal{X} \times \{-1, 1\}$ such that $g(x) \neq y$, and 0 otherwise, the risk functional R is the probability μ of misclassification.

Remark 5.6.3. There are many equivalent formulas to express the risk functional obtained with the zero-one loss:

$$\begin{aligned} \text{(i)} \quad R(g) &= \int_{\mathcal{X} \times \{-1, 1\}} \frac{1}{2} |g(x) - y| \, d\mu(x, y); \\ \text{(ii)} \quad R(g) &= \int_{\mathcal{X} \times \{-1, 1\}} \frac{1}{2} (1 - g(x)y) \, d\mu(x, y). \end{aligned}$$

For a binary classifier $g : \mathcal{X} \rightarrow \{-1, 1\}$, it is often useful to consider its associated *classification function* $f : \mathcal{X} \rightarrow \mathbb{R}$:

Definition 5.6.4. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and $f : \mathcal{X} \rightarrow \mathbb{R}$ be a real-valued measurable function. The sign $\text{sgn}(f)$ of f defines a binary classifier $g : \mathcal{X} \rightarrow \{-1, 1\}$

defined by:

$$g(x) = \begin{cases} 1 & \text{if } f(x) \geq 0 \\ -1 & \text{if } f(x) < 0. \end{cases}$$

Terminology 5.6.5. One says that the function f is the *classification function* of g and one writes g_f .

For a family of classifiers $g_f : \mathcal{X} \rightarrow \{-1, 1\}$, it is often useful to consider both its associated family \mathcal{F} of classification functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and loss functions of the form:

$$\begin{aligned} \mathbb{L} : \mathcal{X} \times \{-1, 1\} &\rightarrow \mathbb{R}_+ \\ (y_1, y_2) &\mapsto \Phi(y_1 y_2), \end{aligned}$$

where $\Phi : \mathbb{R} \rightarrow \mathbb{R}_+$ is a bounded real function.

We can now express, in the spirit of Equation 5.6.1, the risk functional $R(g_f)$ for the classifier g_f linked to a classification function f . To do so, we apply the Disintegration Theorem A.1.1:

For $(\mathcal{X}, \mathcal{B})$ and $(\{-1, 1\}, \mathcal{P}\{-1, 1\})$ two Borel spaces, we consider $(\mathcal{X} \times \{-1, 1\}, \mathcal{B} \otimes \mathcal{P}\{-1, 1\})$ the product space with the natural product measure, a probability measure μ on $\mathcal{X} \times \{-1, 1\}$ and the canonical function $\pi : \mathcal{X} \times \{-1, 1\} \rightarrow \{-1, 1\}$. Then, there exists a $\pi(\mu)$ -almost everywhere determined pair of measures μ_1 and μ_{-1} on $\mathcal{X} \times \{-1, 1\}$ such that

- (i) $\mu_i(\pi^{-1}(\{i\})) = 1$, for $i = -1, 1$. Hence, for $A \subset \mathcal{X} \times \{-1, 1\}$, $\mu_i(A) = \mu(A \cap \pi^{-1}(\{i\})) = \mu(A \cap \mathcal{X} \times \{i\})$,
- (ii) for every Borel measure function $h : \mathcal{X} \times \{-1, 1\} \rightarrow \mathbb{R}_+$,

$$\begin{aligned} \int_{\mathcal{X} \times \{-1, 1\}} h \, d\mu(x, y) &= \int_{\{-1, 1\}} \int_{\pi^{-1}(\{i\})} h \, d\mu_i(x, y) \, d\pi(\mu)(\{i\}) \\ &= \sum_{i=1}^2 \pi(\mu)(\{i\}) \int_{\mathcal{X} \times \{i\}} h \, d\mu_i(x, y). \end{aligned}$$

Thus, in the case where $h(x, y) = \Phi(yf(x))$, we have

$$R(g_f) = \sum_{i=1}^2 \pi(\mu)(\{i\}) \int_{\mathcal{X} \times \{i\}} \Phi(if(x)) \, d\mu_i(x, y).$$

We define two new measures μ_+ and μ_- on \mathcal{X} such that, for any $B \in \mathcal{B}$:

$$\mu_+(B) = \frac{1}{\mu_1(\mathcal{X})} \mu_1(\pi_x^{-1}(B)) \quad \text{and} \quad \mu_-(B) = \frac{1}{\mu_{-1}(\mathcal{X})} \mu_{-1}(\pi_x^{-1}(B)), \quad (5.6.2)$$

With the measures μ_+ and μ_- , we can then write the risk functional as:

$$R(g_f) = \int_{\mathcal{X}} \Phi(f(x)) \, d\mu_+ + \int_{\mathcal{X}} \Phi(-f(x)) \, d\mu_-. \quad (5.6.3)$$

We now give an interpretation of the quantities $\Phi(yf(x))$ and $yf(x)$ of a given point $(x, y) \in \mathcal{X} \times \{-1, 1\}$. The interpretation is based on the notion of *confidence of prediction*. Consider a real-valued function $f \in \mathcal{F}$. By construction, f defines the classifier $g_f = \text{sgn}(f)$. One interprets the absolute value $|f(x)|$ as the confidence of the prediction $g_f(x) = \text{sgn}(f(x))$ made by the classifier g_f . Given a point $(x, y) \in \mathcal{X} \times \{-1, 1\}$, the confidence margin of the prediction $\text{sgn}(f(x))$ is the quantity $yf(x)$. Hence, when the product $yf(x)$ is positive, one can conclude that the classifier $\text{sgn}(f(x))$ classifies the point $x \in \mathcal{X}$ correctly with a confidence $f(x)$. The function $\Phi : \mathbb{R} \rightarrow [a, b]$ represents the loss that one wants to incur on a classification $\text{sgn}(f(x))$ with confidence $|f(x)|$.

It is interesting to note that the zero-one loss defined earlier in this subsection can be written in the $\Phi(yf(x))$ form:

Proposition 5.6.6. *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space, μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a real-valued function and g_f be the classifier linked to f . Then, for any $(x, y) \in \mathcal{X} \times \{-1, 1\}$,*

$$L_{z-o}(g_f(x), y) = 1 - H(yf(x)) = \begin{cases} 1 & \text{if } yf(x) \leq 0 \\ 0 & \text{if } yf(x) > 0 \end{cases}$$

where H is the Heaviside function.

We observe that for $\Phi = 1 - H$, the value of the composed function $(x, y) \mapsto 1 - H(yf(x))$ is only dependent on the sign of f and not on its value. Said differently, a misclassified point (x, y) incurs a loss of 1 no matter the value of $|f(x)|$. Likewise, a well classified point incurs no loss no matter the value $|f(x)|$. If one wants to incur a loss that takes into consideration the value $|f(x)|$ of a point $(x, y) \in \mathcal{X} \times \{-1, 1\}$, one needs to find an appropriate function Φ .

We give two examples of loss functions of the form $\Phi(y_1 y_2)$. First we consider the α -translated zero-one loss. It penalizes both misclassified points and correctly classified points with confidence smaller than α . The penalty is 1. The α -translated zero-one loss function can be written in this form:

Definition 5.6.7 (α -translated zero-one loss). *For any $\alpha > 0$, the α -translated zero-one function $L_\alpha : \mathbb{R} \times \{-1, 1\} \rightarrow \mathbb{R}_+$ is defined by*

$$L_\alpha(y_1, y_2) = \mathbb{1}_{y_1 y_2 \leq \alpha} = \begin{cases} 1 & \text{if } y_1 y_2 \leq \alpha \\ 0 & \text{if } y_1 y_2 > \alpha \end{cases}$$

Next, we consider the α -margin loss function defined in *Foundations of Machine Learning, 2nd edition* [39]. It penalizes misclassified points with the cost of 1 but also penalizes points correctly classified with a confidence smaller or equal to α with a linear penalty of slope $-1/\alpha$. The α -margin loss function can be written in the following format:

Definition 5.6.8 (α -margin loss function). *For any $\alpha > 0$, the α -margin loss function $L_{\Phi_\alpha} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ is defined by $L_{\Phi_\alpha}(y_1, y_2) = \Phi_\alpha(y_1 y_2)$ where*

$$\Phi_\alpha(u) = \min \left(1, \max \left(0, 1 - \frac{u}{\alpha} \right) \right) = \begin{cases} 1 & \text{if } u \leq 0 \\ 1 - \frac{u}{\alpha} & \text{if } 0 \leq u \leq \alpha \\ 0 & \text{if } u \geq \alpha \end{cases}$$

Note that, for any $u \in \mathbb{R}$, $\Phi_\alpha(u) \leq \mathbb{1}_{u \leq \alpha}$. Thus, for all $(x, y) \in \mathcal{X} \times \{-1, 1\}$, the α -margin loss function is smaller or equal to the α -translated zero-one loss function. We therefore have the following inequality between their respective risk functionals:

$$\int_{\mathcal{X} \times \{-1, 1\}} \Phi_\alpha(yf(x)) \, d\mu(x, y) \leq \int_{\mathcal{X} \times \{-1, 1\}} \mathbb{1}_{yf(x) \leq \alpha} \, d\mu(x, y). \quad (5.6.4)$$

The α -translated zero-one loss admits an interpretation: it is a measure of the set of points that has been misclassified or classified with a confidence smaller than α . The limitation of the α -translated zero-one loss is, like for the zero-one loss, the lack of proportionality between the loss incurred by a couple (x, y) and the confidence of the prediction $f(x)$. The margin loss function addresses this short-coming: the penalty for a well classified point with a confidence smaller or equal to α decreases by a slope of $1/\alpha$ with respect to its confidence. If the confidence is greater than α , the penalty is zero, signalling that the impact of points with very large confidence should be limited.

5.7 Rademacher Averages and Generalization Bounds

We now present strong theoretical justifications for the use of loss functions that take into consideration the confidence of predictions. These justifications are based on generalization bounds and the concept of representational capacity, briefly presented in subsection 5.4. As mentioned in that subsection, a properly defined representational capacity allows to construct generalization bounds. In this subsection, we present the definition of, and the important results linked to, *Rademacher averages*. The notion of Rademacher average quantifies the representational capacity of a model \mathcal{G} . We rely heavily on chapter 3, 4 and 5 of *Foundations of Machine Learning, 2nd edition* [39] and on the paper *Rademacher and Gaussian Complexities: Risk Bounds and Structural Results* by Bartlett and Mendelson [5].

Note that many papers use the term *Rademacher complexity* instead of *Rademacher averages*. We do not use the term *Rademacher complexity* as we think that it can

be confused for a descriptor of classification problem complexity that we introduce in Chapter 7. Hence, in this thesis, we use the term *Rademacher averages*, as Bousquet *et al.* [12].

Definition 5.7.1 (Empirical Rademacher Average). *Let $(\mathcal{Z}, \mathcal{B})$ be a measurable space and let μ be a probability measure on \mathcal{Z} . Let $S_n = \{z_1, \dots, z_n\}$ be an independent and identically distributed sample with respect to μ . For a given family \mathcal{H} of functions $h : \mathcal{Z} \rightarrow \mathbb{R}$, we define its empirical Rademacher average $\hat{\mathcal{R}}_S(\mathcal{H})$ by*

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma \left(\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i h(z_i) \right) \right),$$

where $\sigma = (\sigma_1, \dots, \sigma_n)$, with σ_i 's independent uniform random variables taking values in $\{-1, 1\}$. The random variables σ_i are called *Rademacher variables*.

Definition 5.7.2 (Rademacher average). *Let $(\mathcal{Z}, \mathcal{B})$ be a measurable space and let μ be a probability measure on \mathcal{Z} . Let $S_n = \{z_1, \dots, z_n\}$ be an independent and identically distributed sample with respect to μ . Let \mathcal{H} be a family of functions $h : \mathcal{Z} \rightarrow \mathbb{R}$ and $\hat{\mathcal{R}}_S(\mathcal{H})$ be its the empirical Rademacher average. The Rademacher average of \mathcal{H} is the expected value of the empirical Rademacher averages $\hat{\mathcal{R}}_S(\mathcal{H})$ over all samples S_n of size n :*

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E} \left(\hat{\mathcal{R}}_S(\mathcal{H}) \right).$$

According to Bartlett and Mendelson [5], Rademacher empirical averages are reasonable as measures of representational capacity as they quantify the extend to which the functions of \mathcal{H} can be correlated with a noise sequence of length n . The noise sequence is represented by the sequence of Rademacher variables. If a family of functions \mathcal{H} has a large Rademacher empirical average, there is, on average, more chance that one can find a function $h \in \mathcal{H}$, such that $\sum \sigma_i h(z_i)$ is large.

We can now present the first generalization bounds theorem that uses the Rademacher averages.

Theorem 5.7.3. *Let $(\mathcal{Z}, \mathcal{B})$ be a measurable space and let μ be a probability measure on \mathcal{Z} . Let $S_n = \{z_1, \dots, z_n\}$ be an independent and identically distributed sample with respect to μ . Let \mathcal{H} be a family of functions $h : \mathcal{Z} \rightarrow [0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the two following inequalities holds for all $h \in \mathcal{H}$:*

$$\mathbb{E}[h(z)] \leq \frac{1}{n} \sum_{i=1}^n h(z_i) + 2\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}},$$

$$\mathbb{E}[h(z)] \leq \frac{1}{n} \sum_{i=1}^n h(z_i) + 2\hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

Proof: (Theorem 5.7.3)

The proof can be found in page 31 of *Foundations of Machine Learning*, 2nd edition [39]. ■

The Rademacher averages are well defined for any family of functions \mathcal{H} mapping from an arbitrary measurable space \mathcal{Z} to \mathbb{R} . In statistical learning theory, given a model of binary classifier \mathcal{G} and a loss function $L : \mathcal{Y} \times \mathcal{Y}' \rightarrow [0, 1]$, we construct $\mathcal{L}_{\mathcal{G}}$, the family of loss functions associated to \mathcal{G} , mapping from $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to $[0, 1]$ and denoted by $\mathcal{L}_{\mathcal{G}}$:

$$\mathcal{L}_{\mathcal{G}} = \{L_g : (x, y) \mapsto L(g(x), y) : g \in \mathcal{G}\}.$$

Now, in order to make use of Theorem 5.7.3 to construct a generalization bound for a model \mathcal{G} of binary classifiers, one needs to relate the Rademacher averages of the family $\mathcal{L}_{\mathcal{G}}$ of loss functions to the Rademacher averages of the model \mathcal{G} . Recall that in the case of binary classification, $\mathcal{Y}' = \{-1, 1\} \subset \mathbb{R}$.

We first consider the case of the zero-one loss: $L(y_1, y_2) = \mathbb{1}_{y_1 \neq y_2}$:

Lemma 5.7.4. *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Let $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Let \mathcal{G} be a model of binary classifiers $g : \mathcal{X} \rightarrow \{-1, 1\}$ and $\mathcal{L}_{\mathcal{G}}$ be the family of loss*

functions associated to \mathcal{G} for the zero-one loss. Then, $\hat{\mathcal{R}}_S(\mathcal{L}_{\mathcal{G}}) = \frac{1}{2}\hat{\mathcal{R}}_{S_x}(\mathcal{G})$.

The proof of Lemma 5.7.4 is very short and uses a clever trick that will be used again in the proof of Proposition 5.7.6. It can be found in *Foundations of Machine Learning*, 2nd edition [39].

Using both Theorem 5.7.3 and Lemma 5.7.4, we obtain the following result:

Theorem 5.7.5 (Rademacher averages bounds - binary classification). *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Let $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an independent and identically distributed sample with respect to μ . Let \mathcal{G} be a model of binary classifiers $g : \mathcal{X} \rightarrow \{-1, 1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the two following inequalities holds for all $g \in \mathcal{G}$:*

$$R(g) \leq \hat{R}_S(g) + \mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}},$$

$$R(g) \leq \hat{R}_S(g) + \hat{\mathcal{R}}_{S_x}(\mathcal{G}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

Proof: We apply Theorem 5.7.3 to the family $\mathcal{L}_{\mathcal{G}}$ defined by $\mathcal{L}_{\mathcal{G}} = \{L_g : (x, y) \mapsto \mathbb{1}_{g(x) \neq y} : g \in \mathcal{G}\}$. Hence we obtain the inequation:

$$\mathbb{E}[\mathbb{1}_{g(x) \neq y}] \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{g(x_i) \neq y_i} + 2\mathcal{R}_n(\mathcal{L}_{\mathcal{G}}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}},$$

$$\mathbb{E}[\mathbb{1}_{g(x) \neq y}] \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{g(x_i) \neq y_i} + 2\hat{\mathcal{R}}_{S_x}(\mathcal{L}_{\mathcal{G}}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

By definition of the risk R , $\mathbb{E}[\mathbb{1}_{g(x) \neq y}] = R(g)$ and $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{g(x_i) \neq y_i} = \hat{R}_S(g)$.

By Lemma 5.7.4, $\hat{\mathcal{R}}_S(\mathcal{L}_{\mathcal{G}}) = \frac{1}{2}\hat{\mathcal{R}}_{S_x}(\mathcal{G})$ and by taking the expectation we get $\mathcal{R}_n(\mathcal{L}_{\mathcal{G}}) = \frac{1}{2}\mathcal{R}_n(\mathcal{G})$. ■

Now, we consider the case of loss functions of the form $L(y_1, y_2) = \Phi(y_1 y_2)$, where $\Phi : \mathbb{R} \rightarrow [0, 1]$ is a k -Lipschitz function. We start with the following simple but useful proposition that is unrelated to the Lipschitz property:

Proposition 5.7.6. *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Let $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Let \mathcal{F} be a family of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and define $\tilde{\mathcal{F}} = \{\tilde{f} : (x, y) \mapsto yf(x) : f \in \mathcal{F}\}$. Then, $\hat{\mathcal{R}}_S(\tilde{\mathcal{F}}) = \hat{\mathcal{R}}_{S_x}(\mathcal{F})$.*

Proof: (Proposition 5.7.6)

The proof uses the fact that, for a fixed $y \in \{-1, 1\}$, a Rademacher variable σ and the variable $y\sigma$ are both distributed in the same way and take the same values. That is, they are both uniform random variables taking their values in $\{-1, 1\}$. Thus:

$$\begin{aligned} \hat{\mathcal{R}}_S(\tilde{\mathcal{F}}) &= \mathbb{E}_\sigma \left(\sup_{\tilde{f} \in \tilde{\mathcal{F}}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i \tilde{f}(x_i, y_i) \right) \right) \\ &= \mathbb{E}_\sigma \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i y_i f(x_i) \right) \right) \\ &= \mathbb{E}_\sigma \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right) \right) \\ &= \hat{\mathcal{R}}_{S_x}(\mathcal{F}). \end{aligned}$$

■

The Lipschitz property of Φ is necessary to apply *Talagrand's Lemma*. This lemma bounds the empirical Rademacher averages of $\mathcal{L}_{\mathcal{F}}$ (for the loss functions of the form $L(y_1, y_2) = \Phi(y_1 y_2)$) in terms of the Rademacher averages of the family \mathcal{F} of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$:

Lemma 5.7.7 (Talagrand's lemma). *Let Φ_1, \dots, Φ_m be k -Lipschitz functions from \mathbb{R} to \mathbb{R} and $\sigma_1, \dots, \sigma_m$ be Rademacher random variables. Then, for any family \mathcal{F} of*

real-valued functions, the following inequality holds:

$$\frac{1}{m} \mathbb{E}_\sigma \left(\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^m \sigma_i (\Phi_i \circ f)(x_i) \right) \right) \leq \frac{k}{m} \mathbb{E}_\sigma \left(\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^m \sigma_i f(x_i) \right) \right).$$

In particular, if $\Phi_i = \Phi$ for all i , then the following holds:

$$\hat{\mathcal{R}}_S(\Phi \circ \mathcal{F}) \leq k \hat{\mathcal{R}}_S(\mathcal{F}).$$

Proof: (Talagrand's Lemma 5.7.7)

The version of Talagrand's Lemma above is given in *Foundations of Machine Learning, 2nd edition* [39]. It is a simpler and more concise version of a more general version given by Ledoux and Talagrand in [33]. ■

Using both Proposition 5.7.6, and Talagrand's Lemma 5.7.7, we obtain a particular case of Theorem 5.7.3 for the case of loss functions of the form $L(y_1, y_2) = \Phi(y_1 y_2)$:

Theorem 5.7.8. *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Let $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Let \mathcal{F} be a family of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and $\Phi : \mathbb{R} \rightarrow [0, 1]$ be a k -Lipschitz function. Denote by $\mathcal{L}_{\mathcal{F}}$ the family of loss functions associated with \mathcal{F} for the loss function $L(y_1, y_2) = \Phi(y_1 y_2)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the two following inequalities holds for all $g \in \mathcal{F}$:*

$$\begin{aligned} \mathbb{E}[\Phi(yf(x))] &\leq \frac{1}{n} \sum_{i=1}^n \Phi(y_i f(x_i)) + 2k \mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}, \\ \mathbb{E}[\Phi(yf(x))] &\leq \frac{1}{n} \sum_{i=1}^n \Phi(y_i f(x_i)) + 2k \hat{\mathcal{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}. \end{aligned}$$

Proof: (Theorem 5.7.8)

We apply Theorem 5.7.3 to the family $\mathcal{L}_{\mathcal{F}}$ of loss functions where $L(y_1, y_2) = \Phi(y_1 y_2)$.

We obtain:

$$\mathbb{E}[\Phi(\tilde{f})] \leq \frac{1}{n} \sum_{i=1}^n \Phi(\tilde{f}(x_i)) + 2\mathcal{R}_n(\Phi \circ \tilde{\mathcal{F}}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

where $\tilde{\mathcal{F}} = \{\tilde{f} : (x, y) \mapsto yf(x) : g \in \mathcal{F}\}$. By Talagrand's lemma 5.7.7, we have $\mathcal{R}_S(\Phi \circ \tilde{\mathcal{F}}) = k \mathcal{R}_S(\tilde{\mathcal{F}})$ and by Proposition 5.7.6 we obtain $\mathcal{R}_S(\tilde{\mathcal{F}}) = \mathcal{R}_{S_x}(\mathcal{F})$. ■

Let us now consider the α -margin loss function $L_{\Phi_\alpha}(y_1, y_2) = \Phi_\alpha(y_1 y_2)$ with $\Phi_\alpha(u) = \min(1, \max(0, 1 - \frac{u}{\alpha}))$. We obtain the following corollary of Theorem 5.7.8:

Corollary 5.7.9. *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Let $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Let \mathcal{F} be a family of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and denote by $\mathcal{L}_{\mathcal{F}}$ the family of loss functions associated with \mathcal{F} for the loss function $L(y_1, y_2) = \Phi_\alpha(y_1 y_2)$ with $\Phi_\alpha(u) = \min(1, \max(0, 1 - \frac{u}{\alpha}))$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the two following inequalities holds for all $f \in \mathcal{F}$:*

$$R(f) \leq \hat{R}_{S, \Phi_\alpha}(f) + \frac{2}{\alpha} \mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}},$$

$$R(f) \leq \hat{R}_{S, \Phi_\alpha}(f) + \frac{2}{\alpha} \hat{\mathcal{R}}_{S_x}(\mathcal{F}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}},$$

where $\hat{R}_{S, \Phi_\alpha}(f) = \frac{1}{n} \sum_{i=1}^n \Phi_\alpha(y_i f(x_i))$.

While the Rademacher average is a good theoretical quantifier for the representational capacity, it suffers from practical limitations: for most models \mathcal{G} , computing the empirical Rademacher average is an NP-hard problem [39]. For some particular family of classification functions, the Rademacher empirical average can be upper bounded. We consider two family of classification functions: the set of linear functions with bounded weight vectors: $\mathcal{F}_w = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$ and the set of 1-Lipschitz real-functions $\mathcal{F}_{Lip} = \{f : \mathcal{X} \rightarrow \mathbb{R} : \|f\|_{Lip} \leq 1\}$. In both cases, the met-

ric space (\mathcal{X}, d) is totally bounded. Recall that a metric space \mathcal{X} is totally bounded if and only if, for every $\varepsilon > 0$, there exists a finite collection of open balls of radius ε whose centres lie in \mathcal{X} and whose union contains \mathcal{X} . That is, For \mathcal{F}_w , we have the following theorem:

Theorem 5.7.10. *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. For $r > 0$, let $S_n = \{(x_1, y_1), \dots, (x_n, y_n) : \|x_i\| \leq r\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Let $\mathcal{F}_w = \{\mathcal{X} \ni x \mapsto w \cdot x : \|w\| \leq \Lambda\}$, with $\Lambda > 0$. Then, the empirical Rademacher complexity of \mathcal{F}_w can be bounded as follows:*

$$\hat{\mathcal{R}}_S(\mathcal{F}_w) \leq \sqrt{\frac{r^2 \Lambda^2}{n}}.$$

Proof: The proof can be found on page 97 of *Foundations of Machine Learning*, 2nd edition [39]. ■

Combining Corollary 5.7.9 and Theorem 5.7.10, we obtain the following general margin bound for \mathcal{F}_w . Recall that $\Phi_\alpha(u) = \min(1, \max(0, 1 - \frac{u}{\alpha}))$.

Corollary 5.7.11. *Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. For $r > 0$, let $S_n = \{(x_1, y_1), \dots, (x_n, y_n) : \|x_i\| \leq r\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Let $\mathcal{F}_w = \{\mathcal{X} \ni x \mapsto w \cdot x : \|w\| \leq \Lambda\}$, with $\Lambda > 0$ and denote by $\mathcal{L}_{\mathcal{F}_w}$ the family of loss functions associated with \mathcal{F}_w for the loss function $L(y_1, y_2) = \Phi_\alpha(y_1 y_2)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $f \in \mathcal{F}_w$:*

$$R(f) \leq \hat{R}_{S, \Phi_\alpha}(f) + 2\sqrt{\frac{r^2 \Lambda^2 / \alpha^2}{n}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

We can construct a similar bound for the model of 1-Lipschitz functions \mathcal{F}_{Lip} . To do so, we use the notion of covering numbers:

Definition 5.7.12. Let (\mathcal{X}, d) be a totally bounded space. The covering number $N(\mathcal{X}, \epsilon, d)$ of \mathcal{X} is the smallest number of balls of radius ϵ with centers in \mathcal{X} which can cover \mathcal{X} completely.

Using the covering number of a totally bounded space (\mathcal{X}, d) , we obtain an upper bound on the Rademacher average of the unit ball of $Lip(\mathcal{X})$. It is important to note that for the theorem below, the norm associated with the space $Lip(\mathcal{X})$ is

$$\|f\|_{Lip} = \max \left\{ L(f), \frac{\|f\|_{\infty}}{\text{diam}(\mathcal{X})} \right\}$$

This norm was suggested by Bousquet and von Luxbourg in [13].

Theorem 5.7.13. Let (\mathcal{X}, d) be a totally bounded space and Δ denote the d -diameter of \mathcal{X} . If $\mathcal{R}_n(B)$ denotes the unit ball in $Lip(\mathcal{X})$, then, for any $\epsilon > 0$,

$$\mathcal{R}_n(B) \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{2\Delta} \sqrt{N(\mathcal{X}, \frac{u}{4}, d) \ln \left(2 \left\lceil \frac{2\Delta}{u} \right\rceil + 1 \right)} du.$$

Proof: The proof can be found on page 684 of *Distance-Based Classification with Lipschitz Functions* [13]. ■

Using Theorem 5.7.13, we can write a general margin bound for \mathcal{F}_{Lip} :

Corollary 5.7.14. Let (\mathcal{X}, d) be a totally bounded space and Δ denote the d -diameter of \mathcal{X} . Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Let $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Let $\mathcal{F}_{Lip} = \{f : \mathcal{X} \rightarrow \mathbb{R} : \|f\|_{Lip} \leq 1\}$ and denote by $\mathcal{L}_{\mathcal{F}_{Lip}}$ the family of loss functions associated with \mathcal{F}_{Lip} for the loss function $L(y_1, y_2) = \Phi_{\alpha}(y_1 y_2)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $f \in \mathcal{F}_{Lip}$:

$$R(f) \leq \hat{R}_{S, \Phi_{\alpha}}(f) + \frac{2}{\alpha} \inf_{\epsilon > 0} \left(2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{2\Delta} \sqrt{N(\mathcal{X}, \frac{u}{4}, d) \ln \left(2 \left\lceil \frac{2\Delta}{u} \right\rceil + 1 \right)} du \right) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

The two generalisation bounds obtained in Corollary 5.7.11 and Corollary 5.7.14 do not depend directly on the dimension of the space \mathcal{X} . It depends only on the margin α . It suggests that a small generalisation error can be achieved when α is large while the empirical margin loss $\hat{R}_{S, \phi_\alpha}(f)$ remains relatively small. The latter occurs when few points are either misclassified or well classified but with confidence smaller than α . A favourable margin situation depends on the probability measure μ : eventhough the generalisation bound is independent of μ , the existence of a large margin is μ -dependent.

Chapter 6

Classifiers and the Kantorovich-Rubinstein Distance

Chapter 6 is the central theoretical chapter of this thesis. All the results presented are new. We study thoroughly the association between misclassification errors and the Kantorovich-Rubinstein distance.

The first section starts with measurable classifiers and how their error is linked to the total variation distance. We also show that the definition of classifiers and error of classification used in classification theory (see [20] and [27]) are particular cases of our definition of measurable classifiers. In the second part of section 1, we express the risk function of a classifier with two new measures that we construct.

The second section studies Lipschitz classifiers on a metric space. On a space equipped with a distance, we can then define (ϵ, δ) -Lipschitz classifiers and then prove the two most important results of the chapter given in Theorem 6.2.3 and Theorem 6.2.6. In these two theorems, we bound from above the classification error of (ϵ, δ) -Lipschitz classifiers with the Kantorovich-Rubinstein distance.

6.1 Definition of measurable classifiers

Definition 6.1.1. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. To any $B \in \mathcal{B}$, the measurable function $c_B : \mathcal{X} \rightarrow \{-1, 1\}$ given by $c_B = \mathbb{1}_B - \mathbb{1}_{B^c}$ is called a binary classifier.

Notation 6.1.2. For any measurable function $f : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and constant $\gamma \in \mathbb{R}$, we denote by $c_{f,\gamma}$ the family of binary classifiers defined by

$$c_{f,\gamma} = \mathbb{1}_{\{x \in \mathcal{X}; f(x) > \gamma\}} - \mathbb{1}_{\{x \in \mathcal{X}; f(x) \leq \gamma\}}.$$

To make the notations lighter, when $\gamma = 0$, we write c_f instead of $c_{f,0}$.

Remark 6.1.3. We make two remarks:

- (i) On $f^{-1}(\mathbb{R}_*)$, we have $c_{f,0} = \text{sgn}(f)$.
- (ii) For any $\gamma \in \mathbb{R}$, we have $c_{f,\gamma} = c_{f-\gamma,0}$.

Definition 6.1.4. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, μ_1 and μ_2 be two finite measures on \mathcal{X} and $B \in \mathcal{B}$. To the classifier c_B on \mathcal{X} , we associate the quantity ε of c , defined by

$$\varepsilon(c_B; \mu_1, \mu_2) = \mu_1(B^c) + \mu_2(B).$$

Remark 6.1.5. For μ_1, μ_2 and c as in Definition 6.1.4 we have:

- (i) $\varepsilon(c; \mu_1, \mu_2) = \varepsilon(-c; \mu_2, \mu_1)$,
- (ii) $\varepsilon(c; \mu_1, \mu_2) + \varepsilon(c; \mu_2, \mu_1) = \mu_1(\mathcal{X}) + \mu_2(\mathcal{X})$.
- (iii) For any measurable function $f : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and constant $\gamma \in \mathbb{R}$, we have $c_{f,\gamma} = c_B$ when $B = \{x \in \mathcal{X}; f(x) > \gamma\}$. Therefore,

$$\varepsilon(c_{f,\gamma}; \mu_1, \mu_2) = \mu_1(\{x \in \mathcal{X}; f(x) \leq \gamma\}) + \mu_2(\{x \in \mathcal{X}; f(x) > \gamma\}).$$

Recall (see Definition A.2.4) that the total variation norm of a signed measure μ on $(\mathcal{X}, \mathcal{B})$ is given by $\|\mu\| = |\mu|(\mathcal{X})$. The two properties below follow from Proposition A.2.7.

Proposition 6.1.6. *Let $(\mathcal{X}, \mathcal{B})$ be a Borel measurable space and μ_1, μ_2 be two finite measures on \mathcal{X} . Then, there exists a binary classifier $c : \mathcal{X} \rightarrow \{-1, 1\}$ such that*

$$\varepsilon(c; \mu_1, \mu_2) = \frac{1}{2} (\mu_1(\mathcal{X}) + \mu_2(\mathcal{X}) - \|\mu_1 - \mu_2\|).$$

Proof: (Proposition 6.1.6)

Let λ be a common dominating measure for the μ_i 's (ie., $\mu_i \leq \lambda$) on $(\mathcal{X}, \mathcal{B})$ and let m_i be the respective Radon-Nikodym derivatives of μ_i with respect to λ , for $i = 1, 2$. Let M be the measurable set defined by $M = \{x \in \mathcal{X}; m_1(x) \geq m_2(x)\}$ and let $c = \mathbb{1}_M - \mathbb{1}_{M^c}$. Therefore, $\varepsilon(c; \mu_1, \mu_2) = \mu_1(M^c) + \mu_2(M)$. By Proposition A.2.7, we have

$$\begin{aligned} \|\mu_1 - \mu_2\| &= \int c \, d(\mu_1 - \mu_2) = \int_M d(\mu_1 - \mu_2) - \int_{M^c} d(\mu_1 - \mu_2) \\ &= \mu_1(M) - \mu_2(M) - \mu_1(M^c) + \mu_2(M^c) \\ &= \mu_1(\mathcal{X}) + \mu_2(\mathcal{X}) - 2(\mu_1(M^c) + \mu_2(M)). \end{aligned}$$

Hence, $\varepsilon(c; \mu_1, \mu_2) = \frac{1}{2} (\mu_1(\mathcal{X}) + \mu_2(\mathcal{X}) - \|\mu_1 - \mu_2\|)$. ■

Proposition 6.1.7. *Let $(\mathcal{X}, \mathcal{B})$ be a Borel measurable space and μ_1, μ_2 be two finite measures on \mathcal{X} . Then for any measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$ we have*

$$\varepsilon(c_f; \mu_1, \mu_2) \geq \frac{1}{2} (\mu_1(\mathcal{X}) + \mu_2(\mathcal{X}) - \|\mu_1 - \mu_2\|).$$

Proof: (Proposition 6.1.7)

We keep the notations given in the proof of Proposition 6.1.6.

For a measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$, let $B = \{x \in \mathcal{X}; f(x) > 0\}$. By Definition

6.1.4, we have

$$\begin{aligned}\varepsilon(c_f; \mu_1, \mu_2) &= \mu_1(B^c) + \mu_2(B) \\ &= \mu_1(B^c \cap M) + \mu_1(B^c \cap M^c) + \mu_2(B \cap M) + \mu_2(B \cap M^c).\end{aligned}$$

Since $m_2 > m_1$ on M^c , we obtain

$$\mu_2(B \cap M^c) = \int_{B \cap M^c} m_2 \, d\lambda \geq \int_{B \cap M^c} m_1 \, d\lambda = \mu_1(B \cap M^c).$$

Similarly, $\mu_1(B^c \cap M) \geq \mu_2(B^c \cap M)$.

Therefore, we have the following inequality:

$$\begin{aligned}\varepsilon(c_f; \mu_1, \mu_2) &= \mu_1(B^c \cap M^c) + \mu_2(B \cap M^c) + \mu_1(B^c \cap M) + \mu_2(B \cap M) \\ &\geq \mu_1(M^c) + \mu_2(M).\end{aligned}$$

Since $\mu_1(M^c) + \mu_2(M) = \varepsilon(c; \mu_1, \mu_2)$, the proof is complete. ■

Recall that for any pair of probability measures μ_1, μ_2 on $(\mathcal{X}, \mathcal{B})$, the total variation distance is $\|\mu_1 - \mu_2\|_{TV} = \frac{1}{2}\|\mu_1 - \mu_2\|$. Proposition 6.1.6 and Proposition 6.1.7 then become:

Proposition 6.1.8. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and μ_1, μ_2 be two probability measures on \mathcal{X} . Then, there exists a binary classifier $c : \mathcal{X} \rightarrow \{-1, 1\}$ such that*

$$\varepsilon(c; \mu_1, \mu_2) = 1 - \|\mu_1 - \mu_2\|_{TV}.$$

Proposition 6.1.9. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and μ_1, μ_2 be two probability measures on \mathcal{X} . Then, for any measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$, we have*

$$\varepsilon(c_f; \mu_1, \mu_2) \geq 1 - \|\mu_1 - \mu_2\|_{TV}.$$

In the example below, we associate the Bayes decision function introduced in Section 2.1 of *A Probabilistic Theory of Pattern Recognition* [20] to the optimal measurable classifier of Proposition 6.1.6. We keep the notation used in [20] even if it does not correspond to the one used in this thesis.

Example 6.1.10. Let $\nu \in P(\Omega)$, where $P(\Omega)$ denotes the space of Borel probability measures on Ω . Let (h, φ) be a pair of Borel maps from Ω taking their respective values in the Borel measurable space \mathcal{X} and $\{0, 1\}$. Let $\mu = h(\nu)$ denote the push-forward measure on \mathcal{X} and $\eta : \mathcal{X} \rightarrow [0, 1]$ be the regression of φ on h , that is $\eta(x) = \nu(\varphi = 1/h = x)$, for $x \in \mathcal{X}$.

Recalling the Disintegration Theorem A.1.1, we know that there exists a Borel map $x \in \mathcal{X} \mapsto \nu_x \in P(\mathcal{X})$ such that $\nu_x(h^{-1}(x)) = 1$, μ -a.e. and $\nu = \int \nu_x d\mu(x)$. Then, $\eta(x) = \int_{\Omega} \varphi d\nu_x$, for $x \in \mathcal{X}$.

In section 2.1 of [20], the authors define a decision function g as a measurable function $g : \mathcal{X} \rightarrow \{0, 1\}$ given by

$$g(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \in B^c \end{cases} \quad \text{for some } B \in \mathcal{B},$$

where \mathcal{B} is the Borel σ -algebra associated to \mathcal{X} .

We associate to the decision function g , its probability of error $L(g)$ given by $\nu(\{\omega \in \Omega; g \circ h(\omega) \neq \varphi(\omega)\})$, that we denote by $\nu(g \circ h \neq \varphi)$.

With all the relevant notations from section 2.1 of [20], we can now associate the Bayes decision function to the optimal measurable classifier. Let μ_1, μ_2 be two measures on \mathcal{X} given by $\mu_1 = \eta\mu$ and $\mu_2 = (1 - \eta)\mu$ and let $B = \{x \in \mathcal{X}; g(x) = 1\} = g^{-1}(\{1\})$. Then,

Claim: $L(g) = \varepsilon(c_g; \mu_1, \mu_2)$, where c_g is the binary classifier associated to g .

Proof of Claim: By definition, we have

$$L(g) = \nu(\varphi \neq g \circ h) = \nu(\{\omega; \varphi(\omega) = 1, g \circ h(\omega) = 0\}) + \nu(\{\omega; \varphi(\omega) = 0, g \circ h(\omega) = 1\}).$$

The first term of the sum above can be written as follow:

$$\begin{aligned} \nu(\{\omega; \varphi(\omega) = 1, g \circ h(\omega) = 0\}) &= \int_{\mathcal{X}} \nu_x(\{\omega; \varphi(\omega) = 1, g(x) = 0\}) \, d\mu(x) \\ &= \int_{\mathcal{X}} \mathbb{1}_{g^{-1}(0)}(x) \nu_{\mathcal{X}}(\{\omega; \varphi(\omega) = 1\}) \, d\mu(x) \\ &= \int_{\mathcal{X}} \mathbb{1}_{g^{-1}(0)}(x) \eta(x) \, d\mu(x) \\ &= \int_{\mathcal{X}} \mathbb{1}_{g^{-1}(0)} \, d\mu_1(x) \\ &= \mu_1(g^{-1}(0)). \end{aligned}$$

Similarly, we obtain

$$\nu(\{\omega; \varphi(\omega) = 0, g \circ h(\omega) = 1\}) = \int_{\mathcal{X}} \mathbb{1}_{g^{-1}(1)}(1 - \eta(x)) \, d\mu(x) = \int_{\mathcal{X}} \mathbb{1}_{g^{-1}(1)} \, d\mu_2(x) = \mu_2(g^{-1}(1)).$$

As $\varepsilon(c_g; \mu_1, \mu_2) = \mu_1(g^{-1}(0)) + \mu_2(g^{-1}(1))$, the claim is proved.

Following section 2.1, in [20], the Bayes decision function $g^* : \mathcal{X} \rightarrow \{0, 1\}$ is defined by

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ 0 & \text{if } \eta(x) \leq \frac{1}{2} \end{cases}$$

Hence, $g^* = \mathbb{1}_E$ where the set E is defined by $E = \{x \in \mathcal{X}; \eta(x) > \frac{1}{2}\}$.

By construction, η is the Radon-Nikodim derivative of μ_1 and $1 - \eta$ is the Radon-Nikodim derivative of μ_2 . Moreover, the inequality $\eta(x) > \frac{1}{2}$ can easily be rewritten as $1 - \eta(x) < \eta(x)$ and hence E can be written as $E = \{x \in \mathcal{X}; 1 - \eta(x) < \eta(x)\}$. As

in Proposition 6.1.6, we have

$$\|\mu_1 - \mu_2\| = \mu_1(\mathcal{X}) + \mu_2(\mathcal{X}) - \varepsilon(c_{g^*}; \mu_1, \mu_2).$$

As $\mu_1(\mathcal{X}) + \mu_2(\mathcal{X}) = \int \eta \, d\mu + \int (1 - \eta) \, d\mu = 1$ we get $\varepsilon(c_{g^*}; \mu_1; \mu_2) = \frac{1}{2}(1 - \|\mu_1 - \mu_2\|)$ where c_{g^*} is the Bayes binary classifier defined by

$$c_{g^*}(x) = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ -1 & \text{if } \eta(x) \leq \frac{1}{2} \end{cases}$$

Remark 6.1.11. Let us make two remarks:

- (i) Let μ_1, μ_2 be two probability measures on $(\mathcal{X}, \mathcal{B})$. As defined, for example, in Pollard [43], the expression $1 - \|\mu_1 - \mu_2\|_{TV}$ is called the *affinity* between μ_1 and μ_2 and is denoted by $\alpha_1(\mu_1, \mu_2)$. Denote by $\mu_1 \wedge \mu_2$ the largest measure on $(\mathcal{X}, \mathcal{B})$ for which, for all $A \in \mathcal{B}$, the following inequality holds:

$$\mu_1 \wedge \mu_2(A) \leq \min(\mu_1(A), \mu_2(A)).$$

For some dominating measure λ of μ_1 and μ_2 , the measure $\mu_1 \wedge \mu_2$ is such that

$$\frac{d(\mu_1 \wedge \mu_2)}{d\lambda} = \min\left(\frac{d\mu_1}{d\lambda}, \frac{d\mu_2}{d\lambda}\right), \quad \lambda\text{-a.e.}$$

and $\mu_1 \wedge \mu_2(\mathcal{X}) = \|\mu_1 \wedge \mu_2\| = \alpha_1(\mu_1, \mu_2)$.

- (ii) Let ν be a Borel probability measure on Ω . Let (h, φ) be a pair of Borel maps from Ω taking their respective values in \mathcal{X} and $\{0, 1\}$, and assume that $\nu(\{\omega \in \Omega; \varphi(\omega) = 0\}) = \nu(\{\omega \in \Omega; \varphi(\omega) = 1\}) = 1/2$.

As in Exemple 6.1.10, let $\mu = h(\nu)$ and $\eta : \mathcal{X} \rightarrow [0, 1]$ be the regression of φ on h , let μ_1 and μ_2 be two measures on \mathcal{X} given by $\mu_1 = \eta\mu$ and $\mu_2 = (1 - \eta)\mu$. Then, $\mu_1(\mathcal{X}) = \mu_2(\mathcal{X}) = 1/2$.

Let us define two probability measures

$$\tilde{\mu}_1 = \frac{1}{\mu_1(\mathcal{X})} \mu_1 \quad \text{and} \quad \tilde{\mu}_2 = \frac{1}{\mu_2(\mathcal{X})} \mu_2,$$

and let $g^* : \mathcal{X} \rightarrow \{0, 1\}$ be the Bayes decision function given by $g^* = \mathbb{1}_E$, where $E = \{x \in \mathcal{X}; \eta(x) > 1/2\}$. Then,

$$\varepsilon(1 - 2g^*; \tilde{\mu}_1, \tilde{\mu}_2) = 1 - \|\tilde{\mu}_1 - \tilde{\mu}_2\|_{TV}.$$

Consider a classifier c_f linked to a classification function f . In the case of the particular measures μ_+ and μ_- defined in Equation 5.6.2 and specific loss functions, there exist an interesting relation between its risk function $R(c_f)$ and its quantity ε defined in Definition 6.1.4. First, let us recall the definition of μ_+ and μ_- :

Definition 6.1.12. For $(\mathcal{X}, \mathcal{B})$ and $(\{-1, 1\}, \mathcal{P}\{-1, 1\})$ two Borel spaces, we consider $(\mathcal{X} \times \{-1, 1\}, \mathcal{B} \otimes \mathcal{P}\{-1, 1\})$ the product space with the natural product σ -algebra, a probability measure μ on $\mathcal{X} \times \{-1, 1\}$ and the canonical functions $\pi : \mathcal{X} \times \{-1, 1\} \rightarrow \{-1, 1\}$ and $\pi_x : \mathcal{X} \times \{-1, 1\} \rightarrow \mathcal{X}$. We define two new measures μ_+ and μ_- on \mathcal{X} such that:

$$\mu_+ = \frac{1}{\mu_1(\mathcal{X})} \pi_x(\mu_1) \quad \text{and} \quad \mu_- = \frac{1}{\mu_{-1}(\mathcal{X})} \pi_x(\mu_{-1}),$$

where μ_i are measures in $\mathcal{X} \times \{-1, 1\}$ such that $\mu_i(A) = \mu(A \cap \pi^{-1}(\{i\}))$.

We can now write the following proposition:

Proposition 6.1.13. Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{B})$ be a measurable space and let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$. Consider μ_1, μ_{-1} and μ_+, μ_- be as defined in Definition 6.1.12. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a real-valued function and c_f be the classifier linked to the classification function f . Then, for all $(x, y) \in \mathcal{X} \times \{-1, 1\}$,

$$\int_{\mathcal{X} \times \{-1, 1\}} \left(1 - H(yf(x))\right) d\mu(x, y) = \varepsilon(c_f; \mu_+, \mu_-)$$

where H is the Heaviside function and $L_\alpha(u) = \mathbb{1}_{u \leq \alpha}$.

Proof: (Proposition 6.1.13)

Using equality (5.6.3) with $\Phi(yf(x)) = 1 - H(yf(x))$ we can write:

$$\begin{aligned} R(g_f) &= \int_{\mathcal{X}} \left(1 - H(f(x))\right) d\mu_+ + \int_{\mathcal{X}} \left(1 - H(-f(x))\right) d\mu_- \\ &= \mu_+(\{x \in \mathcal{X} : f(x) \leq 0\}) + \mu_-(\{x \in \mathcal{X} : f(x) > 0\}) \\ &= \varepsilon(c_f; \mu_+, \mu_-). \end{aligned}$$

■

Proposition 6.1.13 shows that the quantity ε defined in Definition 6.1.4 is linked to the risk functional of c_f . Using this result as a starting point, we define the notion of ϵ -error for a binary classifier c_f :

Definition 6.1.14. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, μ_1 and μ_2 be two finite measures on \mathcal{X} , and $f : \mathcal{X} \rightarrow \mathbb{R}$ be a measurable function. For any $\epsilon \geq 0$, the ϵ -error of f with respect to (μ_1, μ_2) is given by*

$$\text{err}(f, \epsilon; \mu_1, \mu_2) = \min(\varepsilon(c_f, \epsilon; \mu_1, \mu_2), \varepsilon(c_f, \epsilon; \mu_2, \mu_1)),$$

where $\varepsilon(c_f, \epsilon; \mu_i, \mu_j) = \mu_i(\{x \in \mathcal{X}; f(x) \leq \epsilon/2\}) + \mu_j(\{x \in \mathcal{X}; f(x) > -\epsilon/2\})$

Remark 6.1.15. There are two remarks worth making:

- (i) $\text{err}(c_f, \epsilon; \mu_1, \mu_2) \geq \text{err}(c_f, 0; \mu_1, \mu_2)$
- (ii) $\text{err}(c_{-f}, \epsilon; \mu_1, \mu_2) = \mu_1(\{x \in \mathcal{X}; f(x) \leq -\epsilon/2\}) + \mu_2(\{x \in \mathcal{X}; f(x) > \epsilon/2\})$,

hence we have

$$\text{err}(c_{-f}, \epsilon; \mu_1, \mu_2) + \mu_1(\{x \in \mathcal{X}; f(x) = \epsilon/2\}) = \text{err}(c_f, \epsilon; \mu_1, \mu_2) + \mu_2(\{x \in \mathcal{X}; f(x) = -\epsilon/2\}).$$

Definition 6.1.16 ((ϵ, δ) -classification function). *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and μ_1 and μ_2 be two finite measures on \mathcal{X} . For ϵ and $\delta \geq 0$, a measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$ is a (ϵ, δ) -classification function if*

$$\text{err}(f, \epsilon; \mu_1, \mu_2) \leq \delta.$$

Let \mathcal{X} be a Polish space and μ_1 and μ_2 be two Borel probability measures on \mathcal{X} . The next proposition links (ϵ, δ) -classification function to the two distribution functions of the probability measures $f(\mu_1)$ and $f(\mu_2)$. This result will be useful in the next section 6.2.

Proposition 6.1.17. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let μ_1 and μ_2 be two probability measures on \mathcal{X} . Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a (ϵ, δ) -classification function and F_1, F_2 be the respective distribution functions of the probability measures $f(\mu_1)$ and $f(\mu_2)$. Then,*

$$\int |F_1(t) - F_2(t)| dt \geq (1 - \delta)\epsilon.$$

Proof: (Proposition 6.1.17)

We first suppose that $\varepsilon(c_f, \epsilon; \mu_1, \mu_2) \leq \varepsilon(c_f, \epsilon; \mu_2, \mu_1) \leq \delta$.

Using F_1 and F_2 , we have

$$\varepsilon(c_f, \epsilon; \mu_1, \mu_2) = 1 - (F_2(-\epsilon/2) - F_1(\epsilon/2)) \text{ and thus } F_2(-\epsilon/2) - F_1(\epsilon/2) \geq 1 - \delta.$$

Now, we show that, for all $-\epsilon/2 \leq s \leq \epsilon/2$, $F_2(s) \geq (1 - \delta) + F_1(s)$. Since F_i is monotone and non-decreasing we have: $F_2(s) \geq F_2(-\epsilon/2)$ and $F_1(s) \leq F_1(\epsilon/2)$. Therefore, for all $-\epsilon/2 \leq s \leq \epsilon/2$, we obtain: $F_2(s) \geq F_2(-\epsilon/2) \geq (1 - \delta) + F_1(\epsilon/2) \geq (1 - \delta) + F_1(s)$.

Hence,
$$\int_{-\epsilon/2}^{\epsilon/2} F_2(t) - F_1(t) dt \geq \epsilon(1 - \delta).$$

Likewise, for the case $\varepsilon(c_f, \epsilon; \mu_2, \mu_1) \leq \varepsilon(c_f, \epsilon; \mu_1, \mu_2) \leq \delta$, we obtain

$$\int_{-\epsilon/2}^{\epsilon/2} F_1(t) - F_2(t) dt \geq \epsilon(1 - \delta)$$

Therefore,
$$\int |F_2(t) - F_1(t)| dt \geq \int_{-\epsilon/2}^{\epsilon/2} |F_2(t) - F_1(t)| dt \geq \epsilon(1 - \delta). \quad \blacksquare$$

Lemma 6.1.18. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let μ_1 and μ_2 be two probability measures on \mathcal{X} . Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a measurable function and F_1, F_2 be the respective distribution functions of the probability measures $f(\mu_1)$ and $f(\mu_2)$. Then, for any $t \in \mathbb{R}$,*

$$\min(\varepsilon(c_{f,t}; \mu_1, \mu_2), \varepsilon(c_{f,t}; \mu_2, \mu_1)) = 1 - |F_1(t) - F_2(t)|.$$

Proof: (Lemma 6.1.18)

By the definition of F_1 and F_2 , we have:

$$\begin{aligned} \mu_1(\{x \in \mathcal{X} \mid f(x) > t\}) &= \mu_1(\{x \in \mathcal{X} \mid f(x) \leq t\}^c) = \mu_1(\mathcal{X}) - F_1(t) = 1 - F_1(t) \\ \mu_2(\{x \in \mathcal{X} \mid f(x) > t\}) &= \mu_2(\{x \in \mathcal{X} \mid f(x) \leq t\}^c) = \mu_2(\mathcal{X}) - F_2(t) = 1 - F_2(t). \end{aligned}$$

Using F_1 and F_2 , we can write $\varepsilon(c_{f,t}; \mu_1, \mu_2)$ and $\varepsilon(c_{f,t}; \mu_2, \mu_1)$ as follow:

$$\begin{aligned} \varepsilon(c_{f,t}; \mu_1, \mu_2) &= \mu_1(\{x \in \mathcal{X} \mid f(x) \leq t\}) + \mu_2(\{x \in \mathcal{X} \mid f(x) > t\}) \\ &= \mu_2(\mathcal{X}) - F_2(t) + F_1(t) \\ &= 1 + (F_1(t) - F_2(t)) \\ \varepsilon(c_{f,t}; \mu_2, \mu_1) &= \mu_1(\{x \in \mathcal{X} \mid f(x) > t\}) + \mu_2(\{x \in \mathcal{X} \mid f(x) \leq t\}) \\ &= \mu_1(\mathcal{X}) - F_1(t) + F_2(t) \\ &= 1 - (F_1(t) - F_2(t)). \end{aligned}$$

Hence, $\min(\varepsilon(c_{f,t}; \mu_1, \mu_2), \varepsilon(c_{f,t}; \mu_2, \mu_1)) = 1 - |F_1(t) - F_2(t)|$. ■

6.2 Lipschitz classifiers and the Kantorovich-Rubinstein distance

Generally, the datasets we want to classify are subsets of Euclidean spaces, and more generally of metric spaces. Therefore, in the rest of this section, \mathcal{X} will be equipped

with a distance $d_{\mathcal{X}}$. More precisely, as we will use notions and results of Chapter 2, \mathcal{X} will be a Polish space, with its Borel structure and the distance d will be lower semi-continuous. With a metric space, we can focus on Lipschitz classification functions.

Definition 6.2.1 ((ϵ, δ) -Lipschitz classifier). *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two finite measures on $(\mathcal{X}, \mathcal{B})$.*

- (i) *A Lipschitz classifier $c_f : \mathcal{X} \rightarrow \mathbb{R}$ is a binary classifier such that its classification function $f \in L_1(\mathcal{X}, \mu_i)$, for $i = 1, 2$, and f is 1-Lipschitz with respect to the distance d .*
- (ii) *For ϵ and $\delta \geq 0$, a classifier $c_f : \mathcal{X} \rightarrow \mathbb{R}$ is a (ϵ, δ) -Lipschitz classifier if it is a Lipschitz classifier such that $\text{err}(c_f, \epsilon; \mu_1, \mu_2) \leq \delta$.*

Remark 6.2.2. Let \mathcal{X} be a Polish space, and $\mathcal{B}(\mathcal{X})$ be its Borel σ -algebra. Let d be a lower semi-continuous distance on \mathcal{X} . Then a 1-Lipschitz function $f : \mathcal{X} \rightarrow \mathbb{R}$ is not necessarily $\mathcal{B}(\mathcal{X})$ -measurable.

For example, let \mathcal{X} be the real line \mathbb{R} , with its natural Polish topology and let d be the discrete distance on \mathbb{R} . By Lemma 2.4.4, we know that d is lower semi-continuous. For any subset $E \subset \mathbb{R}$, the characteristic function $\mathbb{1}_E$ is 1-Lipschitz, but is Borel measurable only if $E \in \mathcal{B}(\mathbb{R})$.

Let \mathcal{X} be a Polish space and μ_1 and μ_2 be two Borel probability measures on \mathcal{X} . In the next theorem, we show that an (ϵ, δ) -Lipschitz classifier determines a lower-bound of the Kantorovich-Rubinstein distance of μ_1 and μ_2 :

Theorem 6.2.3. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} .*

If, for $0 \leq \epsilon \leq 1$, c_f is a (ϵ, δ) -Lipschitz classifier, then $W_{\mathcal{X}}(\mu_1, \mu_2) \geq \epsilon(1 - \delta)$.

Proof: (Theorem 6.2.3)

Let f be a (ϵ, δ) -Lipschitz classification function. Applying Theorem 4.1.1 and The-

orem 2.5.5, we have

$$W_{\mathcal{X}}(\mu_1, \mu_2) \geq W_{\mathbb{R}}(\mu_1, \mu_2) = \int_{-\infty}^{+\infty} |F_{\mu_1}(x) - F_{\mu_2}(x)| dx.$$

where F_i denotes the distribution functions of the measure $f(\mu_i)$.

Applying Proposition 6.1.17 completes the proof. ■

Proposition 6.2.4. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} . If Δ denotes the d -diameter of \mathcal{X} , then, for any $\eta > 0$, there exists a 1-Lipschitz measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$ such that*

$$W_{\mathcal{X}}(\mu_1, \mu_2) \leq \Delta(1 + \eta) \text{err}(c_f; \mu_1, \mu_2).$$

Proof: (Proposition 6.2.4)

By Proposition 4.1.5, there exist an optimal function $g : \mathcal{X} \rightarrow \mathbb{R}$ such that $W_{\mathcal{X}}(\mu_1, \mu_2) = W_{\mathbb{R}}(g(\mu_1), g(\mu_2))$. On \mathbb{R} equipped with the Euclidean distance, we apply Theorem 2.5.5 and obtain

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{-\infty}^{+\infty} |G_2(t) - G_1(t)| dt,$$

where $G_i(t) = \mu_i(\{x \in \mathcal{X}; g(x) \leq t\})$ denotes the distribution functions of the measure $g(\mu_i)$, for $i = 1, 2$.

Since $\text{diam}(\mathcal{X}) < +\infty$ and g is 1-Lipschitz, $\text{diam}(g(\mathcal{X})) \leq \Delta$. Thus, $|G_2(t) - G_1(t)| \leq 1$, for any $t \in \mathbb{R}$ and $|G_2(t) - G_1(t)| = 0$ if $t \notin [a, b]$ with $a < b$, $b - a \leq \Delta$. We can write:

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_0^{\Delta} |G_{\mu_2}(t) - G_{\mu_1}(t)| dt.$$

We denote, for simplicity, $\sup_t |G_2(t) - G_1(t)|$ by S . For all $\eta > 0$, there exist a

$t_o \in [a, b]$ such that $S - \eta < |G_2(t_o) - G_1(t_o)| \leq S$. Therefore,

$$W_{\mathcal{X}}(\mu_1, \mu_2) \leq \Delta S \leq \Delta(1 + \eta)|G_2(t_o) - G_1(t_o)|.$$

Set $f = g - t_o$ and F_i the distribution function of the probability measure $f(\mu_i)$. For all $t \in \mathbb{R}$, $F_i(t) = G_i(t - t_o)$ and thus $G_i(t_o) = F_i(0)$. Then, f is a 1-Lipschitz, measurable function such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{\mathcal{X}} f \, d(\mu_1 - \mu_2) \leq \Delta(1 + \eta)|F_2(0) - F_1(0)|.$$

Applying Lemma 6.1.18 completes the proof. ■

In Proposition 6.2.4, we have shown the existence of (ϵ, δ) -Lipschitz classifiers whose ϵ -error δ depends on $W_{\mathcal{X}}(\mu_1, \mu_2)$, when $\epsilon = 0$. In Theorem 6.2.6, we consider the case of (ϵ, δ) -Lipschitz classifiers for $\epsilon \geq 0$. It is a more general result, but we obtain a less stringent upper bound. We first give a proposition that we use in the proof of Theorem 6.2.6.

Proposition 6.2.5. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} .*

If Δ denotes the d -diameter of \mathcal{X} then there exists a 1-Lipschitz classification function f such that for all $0 \leq s < t \leq 1$,

$$s\mu_2(\{x \in \mathcal{X}; f(x) > s\Delta\}) + (1 - t)\mu_1(\{x \in \mathcal{X}; f(x) \leq t\Delta\}) \leq 1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta}.$$

Therefore, for $0 < s \leq 1/2$,

$$\mu_2(\{x \in \mathcal{X}; f(x) > s\Delta\}) + \mu_1(\{x \in \mathcal{X}; f(x) \leq (1 - s)\Delta\}) \leq \frac{1}{s} \left(1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} \right).$$

Proof: (Proposition 6.2.5)

By Proposition 2.3.6, there exists a 1-Lipschitz measurable function $g : \mathcal{X} \rightarrow [0, \Delta]$

such that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_0^\Delta (G_2(s) - G_1(s)) \, ds,$$

where $G_i(s) = \mu_i(\{x \in \mathcal{X}; g(x) \leq s\})$, for $i = 1, 2$.

For $0 \leq s < t \leq 1$, we have

$$\begin{aligned} W_{\mathcal{X}}(\mu_1, \mu_2) &= \left(\int_0^{s\Delta} + \int_{s\Delta}^{t\Delta} + \int_{t\Delta}^\Delta \right) (G_2(s) - G_1(s)) \, ds \\ &\leq s\Delta G_2(s\Delta) + (t-s)\Delta + \Delta(1-t)(1-G_1(t\Delta)). \end{aligned} \quad (6.2.1)$$

Dividing by Δ on both sides and rearranging the terms yields

$$\frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} \leq s G_2(s\Delta) + (t-s) + 1 - G_1(t\Delta) - t(1 - G_1(t\Delta)).$$

For $0 \leq r \leq \Delta$, let $A_r = \{x \in \mathcal{X}; g(x) \leq r\Delta\}$. We can now write:

$$\begin{aligned} \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} &\leq s\mu_2(A_s) + (t-s) + (1-t)(1-\mu_1(A_t)) \\ 1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} &\geq s(1-\mu_2(A_s^c)) + (1-s) - (t-s) + (1-t)(1-\mu_1(A_t)) \\ 1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} &\geq s\mu_2(A_s^c) + (1-t)\mu_1(A_t). \end{aligned}$$

For $s \in (0, 1/2]$ and $t = 1 - s$, we obtain

$$\mu_2(A_s^c) + \mu_1(A_{1-s}) \leq \frac{1}{s} \left(1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} \right).$$

■

Theorem 6.2.6. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two probability measures on \mathcal{X} .*

If Δ denotes the d -diameter of \mathcal{X} then, for $0 \leq \varrho < 1$, there exists a $(\varrho\Delta, \delta)$ -Lipschitz classifier $c_{f, \varrho\Delta}$ such that

$$\delta \leq \frac{2}{1 - \varrho} \left(1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} \right).$$

Proof: (Theorem 6.2.6)

Let $s = \frac{1-\varrho}{2}$. By Proposition 6.2.5, there exists a Lipschitz classification function $g : \mathcal{X} \rightarrow [0, \Delta]$ such that

$$\begin{aligned} \mu_2 \left(\left\{ x \in \mathcal{X}; g(x) > \frac{\Delta}{2} - \frac{\varrho\Delta}{2} \right\} \right) + \mu_1 \left(\left\{ x \in \mathcal{X}; g(x) \leq \frac{\Delta}{2} + \frac{\varrho\Delta}{2} \right\} \right) \\ \leq \frac{2}{1 - \varrho} \left(1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} \right). \end{aligned}$$

By setting $f = g - \frac{\Delta}{2}$, the proof is complete. ■

In the case where the two measures μ_1 and μ_2 are not probability measures but bounded measures such that $\mu_1(\mathcal{X}) = \mu_2(\mathcal{X}) = \gamma$, Proposition 6.2.5 and Theorem 6.2.6 can be made slightly more general:

Proposition 6.2.7. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two finite measures on \mathcal{X} such that $\mu_1(\mathcal{X}) = \mu_2(\mathcal{X}) = \gamma$.*

If Δ denotes the d -diameter of \mathcal{X} then there exists a 1-Lipschitz classification function f such that for all $0 \leq s < t \leq 1$,

$$s\mu_2(\{x \in \mathcal{X}; f(x) > s\Delta\}) + (1 - t)\mu_1(\{x \in \mathcal{X}; f(x) \leq t\Delta\}) \leq \gamma - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta}.$$

Therefore, for $0 < s \leq 1/2$,

$$\mu_2(\{x \in \mathcal{X}; f(x) > s\Delta\}) + \mu_1(\{x \in \mathcal{X}; f(x) \leq (1 - s)\Delta\}) \leq \frac{1}{s} \left(\gamma - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} \right).$$

Theorem 6.2.8. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ_1 and μ_2 be two finite measures on \mathcal{X} such that $\mu_1(\mathcal{X}) = \mu_2(\mathcal{X}) = \gamma$.*

If Δ denotes the d -diameter of \mathcal{X} then, for $0 \leq \varrho < 1$, there exists a $(\varrho\Delta, \delta)$ -Lipschitz classifier $c_{f, \varrho\Delta}$ such that

$$\delta \leq \frac{2}{1 - \varrho} \left(\gamma - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta} \right).$$

6.3 A more general notion of error

In section 6.1, we give the definition of an ϵ -error for a classification function f . The Definition 6.1.14 can be considered quite “rough” as it does not take into consideration the value of the function f but only its sign. It therefore seems natural to consider a more general notion of error:

Definition 6.3.1. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let μ_1 and μ_2 be two finite measures on \mathcal{X} . Let $f : \mathcal{X} \rightarrow \mathbb{R}_+$ be a measurable function such that $f \in L_1(\mathcal{X}, \mu_i)$, $i = 1, 2$.*

Then, for $t \geq 0$, let $\mathcal{E}(f, t; \mu_1, \mu_2)$ be defined by

$$\mathcal{E}(f, t; \mu_1, \mu_2) = \int_{B_t^c} f \, d\mu_1 + \int_{B_t} f \, d\mu_2,$$

where $B_t = \{x \in \mathcal{X}; f(x) \leq t\}$.

Remark 6.3.2. There are three remarks worth making:

- (i) If $\mu_i(\{x \in \mathcal{X} : f(x) = 0\}) = 0$, then, $\mathcal{E}(f, 0; \mu_1, \mu_2) = \int_{\mathcal{X}} f \, d\mu_2$
- (ii) If $f : \mathcal{X} \rightarrow [0, \Delta]$, then $\mathcal{E}(f, \Delta; \mu_1, \mu_2) = \int_{\mathcal{X}} f \, d\mu_1$
- (iii) If $f : \mathcal{X} \rightarrow [0, \Delta]$, then

$$\mathcal{E}(\Delta - f, t; \mu_1, \mu_2) = \int_{B_t^c} \Delta - f \, d\mu_1 + \int_{B_t} \Delta - f \, d\mu_2$$

$$= \Delta(\mu_1(B_t^c) + \mu_2(B_t)) - \mathcal{E}(f, t; \mu_1, \mu_2).$$

Therefore, $\mathcal{E}(f, t; \mu_1, \mu_2) + \mathcal{E}(\Delta - f, t; \mu_1, \mu_2) = \Delta(\mu_2(\mathcal{X}) - (\mu_2(B_t) - \mu_1(B_t)))$
and $\mathcal{E}(f, t; \mu_2, \mu_1) + \mathcal{E}(\Delta - f, t; \mu_2, \mu_1) = \Delta(\mu_1(\mathcal{X}) - (\mu_1(B_t) - \mu_2(B_t)))$.

Definition 6.3.3. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let μ_1 and μ_2 be two finite measures on \mathcal{X} . Let $f : \mathcal{X} \rightarrow \mathbb{R}_+$ be a measurable function such that $f \in L_1(\mathcal{X}, \mu_i)$, $i = 1, 2$.

Then, for $t \geq 0$, the ϵ -error of f with respect to (μ_1, μ_2) is given by

$$\text{Err}(f, \epsilon : \mu_1, \mu_2) = \min\left(\mathcal{E}(f, t; \mu_1, \mu_2) + \mathcal{E}(\Delta - f, t; \mu_1, \mu_2), \mathcal{E}(f, t; \mu_2, \mu_1) + \mathcal{E}(\Delta - f, t; \mu_2, \mu_1)\right).$$

In the particular case where μ_1 and μ_2 are probability measures, we obtain that

$$\text{Err}(f, \epsilon; \mu_1, \mu_2) = \Delta(1 - |\mu_2(B_t) - \mu_1(B_t)|)$$

since, for any $r \in \mathbb{R}$, $\max(r, -r) = |r|$.

By Lemma 6.1.18 we get, for $t \in [0, \Delta]$, $\text{Err}(f, \epsilon; \mu_1, \mu_2) = \text{err}(c_{f,t;\mu_1,\mu_2})$.

Now, we consider a particular case: Let μ_1 and μ_2 be two probability measures on \mathcal{X} . Let Δ denotes the d -diameter of \mathcal{X} and $f : \mathcal{X} \rightarrow [0, \Delta]$ be a measurable function such that, for $t \in [0, \Delta]$, $F_2(t) - F_1(t) \geq 0$, where each $F_i(t) = \mu_i(\{x \in \mathcal{X}; f(x) \leq t\})$ is continuous. Let $t_o \in [0, \Delta]$ be such that $\sup\{F_2(t) - F_1(t); t \in [0, \Delta]\} = F_2(t_o) - F_1(t_o)$. We obtain the following proposition and its immediate corollary:

Proposition 6.3.4. Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let $f : \mathcal{X} \rightarrow [0, \Delta]$ be as defined above and $g = \Delta - f$. Then,

$$\mathcal{E}(f, t_o; \mu_1, \mu_2) = \int_{B_{t_o}} f \, d\mu_1 + \int_{B_{t_o}^c} f \, d\mu_2 \quad \text{and} \quad \mathcal{E}(g, \Delta - t_o; \mu_2, \mu_1) = \int_{B_{t_o}^c} f \, d\mu_1 + \int_{B_{t_o}} f \, d\mu_2,$$

where $B_{t_o} = \{x \in \mathcal{X}; f(x) \in [0, t_o]\}$. Therefore,

$$\mathcal{E}(f, t_o; \mu_1, \mu_2) + \mathcal{E}(g, \Delta - t_o; \mu_2, \mu_1) = \Delta (1 - (F_2(t_o) - F_1(t_o))).$$

Corollary 6.3.5. *Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Then, there exist 1-Lipschitz classification functions $f : \mathcal{X} \rightarrow [0, \Delta]$ and $g = \Delta - f$ such that:*

$$\mathcal{E}(f, t_o; \mu_1, \mu_2) + \mathcal{E}(g, \Delta - t_o; \mu_2, \mu_1) \leq \Delta \left(1 - \frac{W_{\mathcal{X}}(\mu_1, \mu_2)}{\Delta}\right).$$

Chapter 7

Applications of the Kantorovich-Rubinstein Distance in Machine Learning

Chapter 7, aims to show that the use of the Kantorovich-Rubinstein distance as a feature-selection criterion function is an efficient and natural thing to do.

The first section makes the case that the Kantorovich-Rubinstein distance is a very good feature-selection criterion function because, by construction, it encompasses both topological and metric information of the two sample distributions. Hence, the Kantorovich-Rubinstein distance gives us a good bound for the classification errors without having to go through the whole classification process (as shown in the results of Chapter 6).

The second section of the chapter constructs a Kantorovich-Rubinstein distance based function J . We then prove interesting properties of J (monotonicity and additivity) that enable us to apply search algorithms from the Branch-and-Bound family. The advantage of such search engines is that they are much quicker than an exhaustive search while remaining optimal.

7.1 The Kantorovich-Rubinstein distance as a descriptor of sample complexity - 1.9

The no free lunch theorem for machine learning states that, averaged over all possible data-generating distributions, no classification algorithm can consistently have the lowest classification error. Said differently:

“The most sophisticated algorithm we can conceive of has the same average performance (over all possible tasks) as merely predicting that every point belongs to the same class” [7].

Hence, since there is no classification algorithm that can have the smallest misclassification error over all data-generating distributions, there might be, for particular datasets, specific classification algorithms that are better suited. The difficulty is to find these better-suited algorithms: Suppose we are given an independent and identically distributed sample $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ with respect to an unknown probability measure μ on $\mathcal{X} \times \{-1, 1\}$ and S_x be the projection of S_n on \mathcal{X} . If, for a chosen classifier g , the empirical misclassification $\hat{R}(g)$ over S_n is large, one can only speculate which of these two explanations is the correct one: is it because the Bayes error R^* is also large or because the choice of the classifier g is inadequate? There is therefore a need for a characterisation of S_n . In the case of the measurable space $(\mathbb{R}^d \times \{-1, 1\}, \mathcal{B})$, with $d \in \mathbb{N}$, the paper *Complexity Measures of Supervised Classification Problems* [6] by Ho and Basu was seminal in defining the notion of complexity of S_n and finding a set of descriptors that are not directly dependent on a classifier g . We first define the notion of complexity. The complexity of S_n is defined using the Kolmogorov complexity concept: the Kolmogorov complexity of S_n is characterised by the length of the shortest algorithm necessary to fully describe the relationship between each point in $x_i \in S_x$ and their respective label. The longer is the algorithm, the more complex is S_n . The worst case scenario would require to list all the $x_i \in S_x$ along with their label. However, if there exist some form of regularity in S_n , a more compact algorithm can be obtained. In practice, the Kolmogorov complex-

ity is algorithmically incomputable. In their paper, Ho and Basu approximate the Kolmogorov complexity with statistical indicators and geometrical descriptors drawn from S_n . They refer to these indicators and descriptors as *complexity measures*. A subset of the complexity measures is called the *geometrical complexity measures*. It contains descriptors and indicators that describe the regularities and irregularities of the boundary that separates the classes of S_n . Ho and Basu assume that the family of these descriptors and indicators are sufficient to give a good approximation of S_n 's complexity as most classifiers in \mathbb{R}^d can be characterised by geometrical descriptions of their decision regions [6].

Following Ho and Basu, many papers were published with new descriptors. A thorough survey was done in 2019 by Lorena *et al.* [37]. The survey groups geometrical complexity measures according to 4 categories:

1. **Feature-based measures**, which characterise how informative the available features are to separate the classes;
2. **Linearity measures**, which try to quantify whether the classes can be linearly separated;
3. **Neighbourhood measures**, which characterise the presence and density of same or different classes in local neighbourhoods;
4. **Network measures**, which extract structural information from the dataset by modeling it as a graph.

Lorena *et al.* describes the descriptors of category 2 as follows:

“These descriptors try to quantify to what extent the classes are linearly separable, that is, if it is possible to separate the classes by a hyperplane. They are motivated by the assumption that a linearly separable problem can be considered simpler than a problem requiring a non-linear decision boundary.”

There is a major drawback to the linearity descriptors: they are computed using quantities obtained from a linear classifier (i.e., a separating hyperplane). As is standard nowadays, the linear classifier considered is the one constructed using the Support Vector Machines (SVM) algorithm. SVM is considered one of the most

theoretically well motivated and practically most effective classification algorithms. Notwithstanding the quality of SVM, this is a situation in which one constructs a classifier to verify if the dataset can be classified. That is, as said above, precisely what one tries to avoid. What would be preferable are predictors that can be computed solely from the dataset while quantifying whether the classes can be separated by a specific family of classifiers. With the results obtained in the previous chapters, we can do just that: construct a predictor based solely on the dataset that predicts to what extent we can separate the classes using 1-Lipschitz classifiers:

Claim 7.1.1. Let \mathcal{X} be a Polish space and $d \in \mathcal{D}_{\mathcal{X}}$ be a bounded lower semi-continuous distance on \mathcal{X} . Let μ be a probability measure on $\mathcal{X} \times \{-1, 1\}$ and $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be an independent and identically distributed sample with respect to μ and S_x be the projection of S_n on \mathcal{X} . Keeping the notations of Equation (5.6.2), we define:

$$\mu_+ = \frac{1}{2} \frac{1}{|S_+|} \sum_{x \in S_+} \delta_x \quad \text{and} \quad \mu_- = \frac{1}{2} \frac{1}{|S_-|} \sum_{x \in S_-} \delta_x,$$

where $S_+ = \{x_i \in S_x : (x_i, 1) \in S_n\}$ and $S_- = \{x_i \in S_x : (x_i, -1) \in S_n\}$.

We say that $W_{\mathcal{X}}(\mu_+, \mu_-)$ is a good complexity measure for S_n with respect to 1-Lipschitz classifiers if the ratio $W_{\mathcal{X}}(\mu_+, \mu_-)/\Delta$ being close to 1/2 implies that there exists a 1-Lipschitz classifier for which most $x \in S_x$ have a large enough confidence.

We claim that the Kantorovich-Rubinstein distance $W_{\mathcal{X}}(\mu_+, \mu_-)$ between the measures μ_+ and μ_- is a good complexity measure for S_n with respect to 1-Lipschitz classifiers.

We present below what we believe to be a strong theoretical argument to support Claim 7.1.1:

Keeping the notations of Definition 5.6.7, we recall that the risk functional of the classifier c_f with the α -translated zero-one loss is given by

$$R(c_f) = \int_{\mathcal{X} \times \{-1, 1\}} \mathbb{1}_{yf(x) \leq \alpha} d\mu(x, y).$$

Thus, $R(c_f) = \mu(\{(x, y) \in \mathcal{X} \times \{-1, 1\} : yf(x) \leq \alpha\})$. The α -translated zero-one loss penalizes both misclassified points and correctly classified points with confidence smaller than α . The penalty is 1. Thus, the risk $R(c_f)$ measures the set of points that are both misclassified and correctly classified but with a confidence smaller than α . Hence, if $R(c_f)$ is small, the measure of the 3 sets $\{(x, -1) : f(x) \geq \alpha\}$, $\{(x, 1) : f(x) \leq -\alpha\}$ and $\{x \in \mathcal{X} : -\alpha \leq f(x) \leq \alpha\}$ are also small.

In the case of a finite sample S_n , the empirical risk functional is:

$$\hat{R}(c_f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{y_i f(x_i) \leq \alpha}.$$

Thus, a small $\hat{R}(c_f)$ implies that the cardinality of the 3 sets $\{x_i \in S_+ : f(x_i) \geq \alpha\}$, $\{x_i \in S_- : f(x_i) \leq -\alpha\}$ and $\{x_i \in S_n : -\alpha \leq f(x_i) \leq \alpha\}$ is small. This means that the classification function f is such that there are few misclassified points and, moreover, there are few points of S_x in $f^{-1}([-\alpha, \alpha])$. Said differently, the 1-Lipschitz classifier c_f is such that most $x \in S_x$ have a confidence larger than α .

To support our claim, we thus need to show that a large Kantorovich-Rubinstein distance between μ_+ and μ_- implies that $\hat{R}(c_f)$ will be small:

Let $\varphi = \alpha/\Delta$. Then, by Theorem 6.2.8, there exists a (α, δ) -Lipschitz classifier $c_{f,\alpha}$ such that

$$\text{err}(c_{f,\alpha}; \mu_+, \mu_-) \leq \frac{2}{1 - \varphi} \left(\frac{1}{2} - \frac{W_{\mathcal{X}}(\mu_+, \mu_-)}{\Delta} \right).$$

Note that the right hand side of the inequality is always positive since $W_{\mathcal{X}}(\mu_+, \mu_-) \leq \frac{1}{2}\Delta$.

By Proposition 6.1.13, we know that, for the classification function of c_f ,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{y_i f(x_i) \leq \alpha} = \text{err}(c_{f,\alpha}; \mu_+, \mu_-).$$

And thus we obtain:

$$\hat{R}(c_f) \leq \frac{2}{1 - \varphi} \left(\frac{1}{2} - \frac{W_{\mathcal{X}}(\mu_+, \mu_-)}{\Delta} \right).$$

7.2 The Kantorovich-Rubinstein distance and feature-selection algorithms - 1.64

Many machine learning problems become exceedingly difficult when the sample set lives in a high-dimensional space. These difficulties are referred to as *the curse of dimensionality*. These difficulties stem from the fact that, in high-dimensional space, a sample is most likely sparse. Most classification algorithms from the “pre deep learning” era are designed to encourage the construction of smooth classification functions. The idea being that a smooth classification function would have little overfitting. But when your data is sparse, it becomes extremely complicated to generalise well. To avoid the curse of dimensionality, one can either add assumptions about the unknown data generating distribution (as in the case in deep learning and manifold learning) or can reduce the dimensionality of the learning space. There are two different approaches to dimensionality reduction: the feature selection technique and the feature extraction technique. Suppose $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ is a sample in $\mathcal{X} \times \{-1, 1\}$ with $\dim(\mathcal{X}) = r$, and S_x be the projection of S_n on \mathcal{X} . For feature selection, given $0 < a < r$, a score J is assigned to all subsets $A = \{i_1, \dots, i_a\} \subset \{1, \dots, r\}$. The sample S_x is then canonically projected on $\mathcal{X}_{A^*} = \mathcal{X}_{i_{a^*}} \times \dots \times \mathcal{X}_{i_{a^*}}$, where A^* is the subset with the optimal value of J . A feature selection algorithm is in fact a canonical projection on a chosen dimensions of \mathcal{X} .

On the other hand, feature extraction algorithm includes transformations of the space \mathcal{X} before projecting on a space of lower dimension [27]. Note that both techniques can be used on the same set of observations \mathcal{X} .

Well known examples of feature extraction techniques include Principal Component Analysis (PCA) and Random Projections. PCA is extensively covered in Chapter 12 of *Pattern Recognition and Machine Learning* [11] while Random Projections is well summarized in section 4.2 of H. Duan’s Master’s thesis [22]. Another feature extraction technique worth mentioning is the *Borel isomorphic dimensionality reduction technique*. It was introduced by V. Pestov in [41] and further studied by S.

Hatko in [26].

One can often obtain a set of transformed features generated by feature extraction that provide a better discriminative ability than the best subset chosen by feature selection. The drawback is that feature extracted subsets often lose their original physical/biological meaning which makes it less useful for applications. For that reason, we restrict our interest to feature selection algorithms and more precisely the Branch & Bound family of algorithms. A quick overview of feature selection algorithms is given in Appendix B. We recommend you read it before reading subsection 7.2.1.

7.2.1 Branch & Bound algorithms - 1.73

This section gives a brief overview of the initial Branch & Bound algorithm and touches on its newer and improved versions. The section relies heavily on the paper by Somol *et al.* [48] and Frank *et al.* [24]. As explained in the introduction of section 7.2, the goal of feature selection algorithms is to select, from a set E of r elements, an optimal subset of cardinality k . The notion of optimality is measured by an evaluation criterion function J . From Appendix B, we know that, in the case of a set E with large cardinality, it is totally unrealistic to expect to compute J for each subset of cardinality k . Indeed, the exhaustive search is impractical even for problems of small cardinality r as the optimization space of all subsets of cardinality $k < r$ is of combinatorial complexity. To bypass the issue of complexity, two strategies has been used. The first one is to relax the rigour of optimality. (ie. reducing the size of the optimization space). The second strategy is the introduction of a feature selection criterion function (fscf) with specific properties which help to identify sections of the search optimization space that can be left unexplored. We use the standard notation $J : \mathcal{P}(E) \mapsto \mathbb{R}$ for the fscf. The Branch & Bound family of algorithms employs the second strategy by making use of a fscf J that satisfies the set inclusion monotonicity property.

Definition 7.2.1. *Let E be a set of cardinality r and $\mathcal{P}(E)$ be the power set of E .*

Consider $f : \mathcal{P}(E) \mapsto \mathbb{R}$. The set inclusion monotonicity property assumes that, for two subsets, $E_1, E_2 \in \mathcal{P}(E)$,

$$f(E_1) \leq f(E_2) \text{ if } E_1 \subset E_2.$$

It is the use of a fscf J satisfying the inclusion monotonicity property that allows to remove parts of the search space that cannot possibly contain the optimal solution. Given a fscf satisfying the inclusion monotonicity property, one can use an algorithm from the Branch & Bound family. The Branch & Bound principle can be summarized as follows:

First, the algorithm constructs a solutions tree $T(E, k)$ where the root represents the set E of cardinality r and the $C(r, k)$ leaves represent all the subsets of k elements. In the solution tree, any node at depth d represents a subset of E with a cardinality of $r - d$ elements. Hence, the leaves are at depth $r - k$. The generation of a solution tree can be regarded as a recursive procedure that builds solution trees from depth $d = 0$ to $d = r - k$. After obtaining a solution tree, all string-structure subtrees (subtrees in which non-leaves nodes have only one child node) can be pruned to obtain a minimum solution tree. Note that the generation of a solution tree (and of minimum solution tree) is independent of the choice of fscf. For more details on this recursive procedure, see Yu *et al* [55].

Once the minimum solution tree has been built, one traverses the tree to find the optimal subset. The algorithm keeps the fscf of the currently best leaf-node in memory (denoted as the bound). Anytime the fscf of some internal node is found to be lower than the bound, the whole subtree may be cut off and thus many computations can be omitted. Somol *et al.* [48] have observed that: 1) nearer to the root, the fscf computation is usually slower as evaluated feature subsets are larger and 2) nearer to the root, subtree cut-offs are less frequent. Note that, for a given minimum solution tree, the speed of the Branch & Bound algorithm depends on the choice of the fscf. Indeed, "traversing the solution tree" requires to compute the fscf at many (if not most) of the tree nodes.

Over time, considerable effort has been invested into the acceleration of the Branch & Bound algorithm. The article *Fast Branch & Bound Algorithms for Optimal Feature Selection* by Somol *et al.* [48] is a good review of the many improvements since the original Branch & Bound algorithm developed by Narendra and Fukunaga [40].

7.2.2 The use of the Kantorovich-Rubinstein distance to construct an evaluation criterion function - 1.86

Our goal is to use the Branch & Bound algorithm as a dimensionality reduction prior to using a classification algorithm. Hence, a natural choice for the criterion function J is a function that satisfies the set inclusion monotonicity property defined in 7.2.1 and characterises the complexity of S_n in $\mathcal{X}_A = \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_a}$, with $A = \{i_1, \dots, i_a\} \subset \{1, \dots, r\}$. We could then be cautiously optimistic that our “optimal” subset $A^* = \{i_1^*, \dots, i_a^*\}$ would allow the projection of S_n into $\mathcal{X}_{A^*} \times \{-1, 1\}$ to be easily classifiable. We now construct a Kantorovich-Rubinstein distance-based function J and prove two interesting properties of it: monotonicity and additivity.

Let $r \in \mathbb{N}$ be a positive integer. For $1 \leq k \leq r$, let \mathcal{X}_k be a Polish space, $\mathcal{B}_k = \mathcal{B}(\mathcal{X}_k)$ be its Borel σ -algebra and d_k be a lower semi-continuous distance on \mathcal{X}_k (written $d_k \in \mathcal{D}_{\mathcal{X}_k}$). The product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ endowed with the product topology is a Polish space and its Borel σ -algebra is $\mathcal{B}(\mathcal{X}) = \mathcal{B}_1 \otimes \dots \otimes \mathcal{B}_r$. We denote by \tilde{d}_1 (respectively \tilde{d}_∞) the ℓ_1 (respectively ℓ_∞) distance on \mathcal{X} . As seen in Definition 4.2.1, we have, for $x, y \in \mathcal{X}$,

$$\tilde{d}_1(x, y) = \sum_{i=1}^r d_i(x_i, y_i) \quad \text{and} \quad \tilde{d}_\infty(x, y) = \max_{1 \leq k \leq r} d_k(x_k, y_k).$$

By Lemma 4.2.4, we know that \tilde{d}_1 and \tilde{d}_∞ are lower semi-continuous on \mathcal{X} . Consider $\mathcal{X}_A = \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_a}$ where $\{i_1, \dots, i_a\}$ belongs to the power set $\mathcal{P}\{1, \dots, r\}$. Recall that the canonical projection $\pi_a : \mathcal{X} \rightarrow \mathcal{X}_A$ is defined by $\pi_a(x) = (x_{i_1}, x_{i_2}, \dots, x_{i_a})$

and the ℓ_1 -distance on \mathcal{X}_A is defined by

$$\tilde{d}_{1,a}(x, y) = \sum_{k=1}^a d_{i_k}(x_{i_k}, y_{i_k}).$$

We can now define our evaluation function:

Definition 7.2.2. [l.112] *Let r be a positive integer. For $1 \leq k \leq n$, let \mathcal{X}_k be a Polish space and $d_k \in \mathcal{D}_{\mathcal{X}_k}$. Consider the product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ endowed with the product topology.*

We denote by $J : P_{\mathcal{X}} \times P_{\mathcal{X}} \times \mathcal{P}(\{1, \dots, r\}) \rightarrow \mathbb{R}_+$ the evaluation function defined by:

$$J(\mu_1, \mu_2, A) = W_{\mathcal{X}_A}(\pi_a(\mu_1), \pi_a(\mu_2)),$$

where $W_{\mathcal{X}_A}$ is the Kantorovich-Rubinstein distance associated to $(\mathcal{X}_A, \tilde{d}_{1,a})$.

For $(\mu_1, \mu_2) \in P_{\mathcal{X}} \times P_{\mathcal{X}}$ fixed, we define the function $J_{(\mu_1, \mu_2)} : \mathcal{P}(\{1, \dots, r\}) \rightarrow \mathbb{R}_+$ given by the formula $J_{(\mu_1, \mu_2)}(A) = J(\mu_1, \mu_2, A)$.

The criterion function J defined above has two interesting properties:

Proposition 7.2.3. *Let r be a positive integer. For $1 \leq k \leq r$, let \mathcal{X}_k be a Polish space and $d_k \in \mathcal{D}_{\mathcal{X}_k}$. Let $A = \{i_1, \dots, i_a\} \subset B = \{j_1, \dots, j_b\}$ be non-empty subsets of $\{1, \dots, r\}$. Consider the product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ endowed with the product topology and let μ_1 and μ_2 be two probability measures on $P_{\mathcal{X}}$.*

Then, $J_{(\mu_1, \mu_2)}(A) \leq J_{(\mu_1, \mu_2)}(B)$.

Such a function $J_{(\mu_1, \mu_2)}$ is said to be monotone.

Proof: As $J_{(\mu_1, \mu_2)}(A) = W_{\mathcal{X}_A}(\pi_a(\mu_1), \pi_a(\mu_2))$ and $J_{(\mu_1, \mu_2)}(B) = W_{\mathcal{X}_B}(\pi_b(\mu_1), \pi_b(\mu_2))$, a direct application of Corollary 4.2.6 yields the desired result. ■

Proposition 7.2.4 (l.125). *Let r be a positive integer. For $1 \leq k \leq r$, let \mathcal{X}_k be a Polish space, $d_k \in \mathcal{D}_{\mathcal{X}_k}$ and $\mu_1^{(k)}, \mu_2^{(k)} \in P_{\mathcal{X}_k}$ be two probability measures. Consider the*

product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_r$ endowed with the metric d and $\mu_1 = \mu_1^{(1)} \otimes \dots \otimes \mu_1^{(r)}$ and $\mu_2 = \mu_2^{(1)} \otimes \dots \otimes \mu_2^{(r)}$, the corresponding product measure on \mathcal{X} . Then, for any $A, B \in \mathcal{P}(\{1, \dots, r\})$, $A \cap B = \emptyset$,

$$J_{(\mu_1, \mu_2)}(A \cup B) = J_{(\mu_1, \mu_2)}(A) + J_{(\mu_1, \mu_2)}(B).$$

Such a function $J_{(\mu_1, \mu_2)}$ is said to be additive.

Proof: Without loss of generality, it is enough to prove that additivity holds for singleton. Let $A = \{a\}$ and $B = \{b\}$. As μ is a product measure, we have, by construction, that $\pi_a(\mu) = \mu_a$, $\pi_b(\mu) = \mu_b$ and $\pi_{a \cup b}(\mu) = \mu_a \otimes \mu_b$. Using Theorem 2.2.7, we obtain:

$$\begin{aligned} J_{(\mu_1, \mu_2)}(A \cup B) &= W_{\mathcal{X}_A \times \mathcal{X}_B}(\pi_{a \cup b}(\mu_1), \pi_{a \cup b}(\mu_2)) \\ &= W_{\mathcal{X}_A \times \mathcal{X}_B}(\mu_1^{(a)} \otimes \mu_1^{(b)}, \mu_2^{(a)} \otimes \mu_2^{(b)}) \\ &= W_{\mathcal{X}_A}(\mu_1^{(a)}, \mu_2^{(a)}) + W_{\mathcal{X}_B}(\mu_1^{(b)}, \mu_2^{(b)}) \\ &= J_{(\mu_1, \mu_2)}(A) + J_{(\mu_1, \mu_2)}(B). \end{aligned}$$

■

Chapter 8

Conclusion

The idea to use the Kantorovich-Rubinstein distance for dimensionality reduction was presented to Hubert Duan and I by Professor Vladimir Pestov. In his Master’s thesis [22], H. Duan constructed the “simplest” dimensionality reduction algorithm: for a sample S_n of points in $\mathcal{X} \times \{-1, 1\}$ (where \mathcal{X} is a product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_d$) and S_x the projection of S_n on \mathcal{X} , he considered the d projection maps $\pi_i : \mathcal{X} \rightarrow \mathcal{X}_i$ and associated the zero-one distance to each \mathcal{X}_i . Such a choice of distance made his d evaluation functions J_i equal to the total variation distance on the d samples $\pi_i(S_x)$. He then fixed a threshold α and collected the coordinates $\{i_1, \dots, i_a\} \subset \{1, \dots, d\}$ with an evaluation function greater than α . This allowed him to define a dimension reduction map $T : \mathcal{X} \rightarrow \mathcal{X}_A$ where $\mathcal{X}_A = \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_a}$.

In conjunction with the Random Forest classification algorithm, H. Duan tested his dimensionality reduction algorithm on a portion of the OHGS-2 dataset. He obtained promising results.

In this thesis, we had two main objectives. The first objective was to build a strong theoretical foundation to justify the use of the Kantorovich-Rubinstein distance to construct dimensionality reduction techniques applicable to GWAS datasets such as the OHGS dataset. The second objective was to generalise H. Duan dimensionality reduction algorithm in two ways. First by considering not only the d projection maps

$\pi_i : \mathcal{X} \rightarrow \mathcal{X}_i$ but the projections over all possible combination of the d coordinates. Secondly, by considering any distance d_i associated to each \mathcal{X}_i .

As this thesis had two main objectives, each objective will have its own list of suggestions for future research.

In the case of the theoretical foundations to justify the use of the Kantorovich-Rubinstein distance, here are additional ideas that could be developed:

- (i) Here, we keep the notations of Chapter 5 and Chapter 6. In Chapter 5, we give an interpretation of the quantities $yf(x)$ and $\Phi(yf(x))$: $yf(x)$ is the confidence margin of the prediction $g_f(x)$ and $\Phi(yf(x))$ is the loss incurred by c_f at the point (x, y) . In the two particular cases of $\Phi(yf(x)) = 1 - H(yf(x))$ and $\Phi(yf(x)) = L_\alpha(yf(x))$ (where H is the Heaviside function and L_α is the α -translated zero-one loss), we showed that $R(g_f) = \text{err}(c_{f,\alpha}; \mu_+, \mu_-)$. This equality does not stand for all loss functions. In particular, one cannot write $R(g_f) = \text{Err}(f, \alpha; \mu_+, \mu_-)$ when we consider the α -margin loss function i.e., $\Phi_\alpha(yf(x)) = \min(1, \max(0, 1 - \frac{y}{\alpha}))$. It would therefore be interesting to construct all loss functions that allow the risk functional to be written as an error. It would require to find the structure of all loss functions $\Phi : \mathcal{X} \times \{-1, 1\} \rightarrow \mathbb{R}$ that satisfy the equalities

$$\int_{\mathcal{X}} \Phi_\alpha(f(x)) d\mu_+ = \int_B \Phi_\alpha(f(x)) d\mu_+ \text{ and } \int_{\mathcal{X}} \Phi_\alpha(-f(x)) d\mu_- = \int_{B^c} \Phi_\alpha(-f(x)) d\mu_-$$

where $B \subset \mathcal{X}$.

- (ii) Here, we keep the notations of Chapter 6. In the proof of Proposition 6.2.5, we construct an upper bound for the Kantorovich-Rubinstein distance (see Equation (6.2.1)). This upper bound is constructed using a geometrical approach: since the Kantorovich-Rubinstein distance $W_{\mathcal{X}}(\mu_1, \mu_2)$ can be expressed as the area between the distribution functions G_1 and G_2 (where $G_i(s) = \mu_i(\{x \in g(x) : g(x) \leq s\})$), we built a covering of $W_{\mathcal{X}}(\mu_1, \mu_2)$ with three rectangles. The

sum of each rectangle's area could then be associated to the notion of error $\text{err}(c_{f,\alpha}; \mu_1, \mu_2)$ defined in 6.1.14. We think that we can build a better covering of $W_{\mathcal{X}}(\mu_1, \mu_2)$ using geometrical areas that are given in Equation (D.0.1) and Equation (D.0.2) of Appendix D. The sum of these areas would be associated with the notion of error $\text{Err}(f, \epsilon : \mu_1, \mu_2)$ defined in 6.3.3.

In the case of the second objective (i.e., the generalisation of H. Duan's algorithm), the natural next step is to test our new algorithm on the OHGS-2 dataset. For this thesis, many programs were written to study the Kantorovich-Rubinstein distance as a tool of dimensionality reduction. These programs are written in R. They are accessible upon request. These programs are separated in two categories. The first category regroups programs preparing the OHGS-2 dataset for computations and computing the Kantorovich-Rubinstein distance. The second category regroups the analysis programs. The computation should be improved in two ways. First, the Kantorovich-Rubinstein computation needs to be upgraded with the latest (and thus fastest) algorithms. Today, there exists many efficient algorithms to compute approximations of the Kantorovich-Rubinstein distance. The major breakthrough came in 2013 when Cuturi [17] showed that the empirical Kantorovich-Rubinstein distance can be regularised by an entropic term. The regularisation turns the Kantorovich-Rubinstein distance into a strictly convex problem which can be solved very quickly with the Sinkhorn-Knopp's matrix scaling algorithm. Such a convex problem exhibits linear convergence and can be trivially parallelised. Since Cuturi's paper, many improved algorithms have been published. In particular, the greedy coordinate descent algorithm *Greenhorn* developed by Altschuler *et al.* [2] runs in near-linear time (i.e., runs in time $\mathcal{O}(d^2)$ where d is the space dimension). The expertise in Kantorovich-Rubinstein distance computation exists at the department of Mathematics and Statistics of the University of Ottawa. Secondly, we would need to write the code for an efficient Branch & Bound algorithm. Paired with the Greenhorn algorithm, we would obtain a state-of-the-art dimensionality reduction algorithm for which we could fix the number of dimensions we want. From the biology perspective, this would allow

us to choose the number n of snps to consider and then obtain the most significant combination of n snps in a timely fashion.

Appendix A

Elementary Definitions and Results

This Appendix features some known definitions and results used in the thesis but not directly related to optimal transport or classification problems. The Appendix is separated in two sections that are totally independent of each other. The first section states the disintegration theorem that is used often in the thesis. The second section is a quick overview of the total variation distance.

A.1 Disintegration Theorem

Theorem A.1.1 (Disintegration Theorem - 1.7). *Let \mathcal{X} and \mathcal{Y} be two standard Borel spaces and $P(\mathcal{X})$ and $P(\mathcal{Y})$ be their respective collection of Borel probability measures. Let $\mu \in P(\mathcal{Y})$, let $\pi : \mathcal{Y} \rightarrow \mathcal{X}$ be a Borel-measurable function, and let $\nu \in P(\mathcal{X})$ be the push-forward measure $\nu = \pi * (\mu) = \mu(\pi^{-1})$.*

Then there exists a ν -almost everywhere uniquely determined family of probability measures $\{\mu_x\}_{x \in \mathcal{X}} \subset P(\mathcal{Y})$ such that :

- *the function $x \mapsto \mu_x$ is Borel measurable, in the sense that $x \mapsto \mu_x(B)$ is a Borel measurable function for each Borel measurable set $B \subset \mathcal{Y}$;*
- *μ_x "lives on" the fiber $\pi^{-1}(x)$: for ν -almost all $x \in \mathcal{X}$, $\mu_x(\mathcal{Y} \setminus \pi^{-1}(x)) = 0$. Hence $\mu_x(B) = \mu_x(B \cap \pi^{-1}(x))$, for $B \in \mathcal{Y}$.*

- for every Borel measurable function $f : \mathcal{Y} \rightarrow [0, +\infty]$,

$$\int_{\mathcal{Y}} f(y) \, d\mu(y) = \int_{\mathcal{X}} \int_{\pi^{-1}(x)} f(y) \, d\mu_x(y) \, d\nu(x)$$

In particular, for any event $B \subset \mathcal{Y}$, taking f to be the indicator function of B ,

$$\mu(B) = \int_{\mathcal{X}} \mu_x(B) \, d\nu(x).$$

Proof: (Theorem A.1.1)

For a proof of the Disintegration Theorem, one can consult Kechris ([30], p.115). ■

A.2 Total Variation Norm

Definition A.2.1 (signed measure). Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and let μ be a function on \mathcal{F} with values in $(-\infty, +\infty)$.

If μ is countably additive and satisfies $\mu(\emptyset) = 0$, it is a finite signed measure.

Note that a signed measure is a function that result if the requirement of non-negativity is removed from the definition of a measure.

Theorem A.2.2 (Hahn decomposition theorem - 1.39). Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, and let μ be a signed measure on $(\mathcal{X}, \mathcal{F})$. Then there are disjoint subsets P and N of \mathcal{X} such that P is a positive set for μ , N is a negative set for μ and $\mathcal{X} = P \cup N$.

Proof: For a proof of the Hahn decomposition theorem, one can consult Theorem 4.1.4 in *Measure Theory* by Cohn [15]. ■

Corollary A.2.3. [Jordan decomposition theorem - 1.46] Every signed measure is the difference of two positive measures, at least one of which is finite. Indeed, for any

measure μ , one can write $\mu = \mu_+ - \mu_-$, with $\mu_+(A) = \mu(A \cap P)$, $\mu_-(A) = -\mu(A \cap N)$ where P is a positive set for μ and N is a negative set for μ .

The representation $\mu = \mu_+ - \mu_-$ is called the Jordan decomposition of μ .

Proof: For a proof of the Jordan decomposition theorem, one can consult Corollary 4.1.5 in *Measure Theory* by Cohn[15]. ■

The Definition A.2.4 given below is given in *Measure Theory* by Cohn.

Definition A.2.4 (Total Variation norm - 1.59). Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and let μ be a signed measure on \mathcal{X} .

The variation of the signed measure μ is the positive measure $|\mu|$ defined by

$$|\mu| = \mu_+ + \mu_-.$$

The total variation norm $\|\mu\|$ of the signed measure μ is defined by

$$\|\mu\| = |\mu|(\mathcal{X}).$$

Remark A.2.5. As written in David Pollard's manuscript [43], the total variation norm can equivalently be defined as $\|\mu\| = \sup(|\mu|(A) + |\mu|(A^c))$, where the supremum is taken over all $A \in \mathcal{B}$.

Example A.2.6 (Total variation norm for atomic measures). Let S be a finite subset of \mathcal{X} . Consider the atomic signed measure $\mu = \sum_{x \in S} \mu_x \delta_x$. Let us define, on \mathcal{X} , two positive measures

$$\mu_+ = \sum_{x \in S_+} \mu_x \delta_x \quad \text{and} \quad \mu_- = \sum_{x \in S_-} (-\mu_x) \delta_x,$$

where $S_+ = \{x \in S; \mu_x \geq 0\}$ and $S_- = \{x \in S; \mu_x \leq 0\}$. Then,

$$|\mu| = \sum_{x \in S} |\mu_x| \delta_x \quad \text{and} \quad \|\mu\| = |\mu|(\mathcal{X}) = \sum_{x \in S} |\mu_x|.$$

Proposition A.2.7. [1.81] Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and let $\|\mu\|$ be the total variation norm of a signed measure μ . Then,

(i) $\|\mu\| = \int_{\mathcal{X}} |m| \, d\lambda$, where $m = \frac{d\mu}{d\lambda}$ for some dominating measure λ ,

(ii) $\|\mu\| = \frac{1}{k} \sup \left\{ \left| \int f \, d\mu \right| ; f : \mathcal{X} \rightarrow [-k, k], f \text{ measurable} \right\}$.

The supremum over all measurable f , $|f| \leq k$, is attained with the function g defined by $g = k\mathbb{1}_M - k\mathbb{1}_{M^c}$, where $M = \{x \in \mathcal{X}; m \geq 0\}$.

Proof:

(i) Since μ is absolutely continuous with respect to λ , μ_+ and μ_- are also absolutely continuous with respect to λ . It is well known that the Radon-Nikodym derivative for μ_+ and μ_- are, respectively, the positive and negative parts of m , denoted by m_+ and m_- . Hence,

$$|\mu| = \int m_+ \, d\lambda + \int m_- \, d\lambda = \int |m| \, d\lambda \text{ and thus } \|\mu\| = \int_{\mathcal{X}} |m| \, d\lambda.$$

(ii) Since $|f| \leq k$, we have

$$\left| \int f \, d\mu \right| = \left| \int f m \, d\lambda \right| \leq \int |f| |m| \, d\lambda \leq k \int |m| \, d\lambda.$$

$$\text{Hence } \sup \left\{ \left| \int f \, d\mu \right| ; f : \mathcal{X} \rightarrow [-1, 1], f \text{ measurable} \right\} \leq k \|\mu\|.$$

Let $M = \{x \in \mathcal{X}; m \geq 0\}$ and define the function f by $f = k\mathbb{1}_M - k\mathbb{1}_{M^c}$. Then,

$$\int f m \, d\lambda = \int k m_+ \, d\lambda - \int (-k m_-) \, d\lambda = \int k m_+ + k m_- \, d\lambda = k \int |m| \, d\lambda.$$

Therefore, $\|\mu\| = \frac{1}{k} \sup \left\{ \left| \int f \, d\mu \right| ; f : \mathcal{X} \rightarrow [-k, k], f \text{ measurable} \right\}$. ■

Proposition A.2.8. [l.110] Let $(\mathcal{X}, \mathcal{B})$ be a Borel measurable space, μ be a signed

measure on $(\mathcal{X}, \mathcal{B})$ such that $\mu(\mathcal{X}) = 0$ and μ_+ , respectively $\|\mu\|$, be the positive part, respectively the total variation norm, of μ . Then,

$$(i) \quad \|\mu\| = 2 \int_{\mathcal{X}} m_+ d\lambda,$$

$$(ii) \quad \|\mu\| = 2 \sup_A \mu(A).$$

Proof:

(i) Since $\mu(\mathcal{X}) = 0$, $\int_{\mathcal{X}} m_+ d\lambda = \int_{\mathcal{X}} m_- d\lambda$. Hence,

$$\|\mu\| = \int_{\mathcal{X}} |m| d\lambda = 2 \int_{\mathcal{X}} m_+ d\lambda = 2 \int_{\mathcal{X}} m_- d\lambda.$$

(ii) Since $\mu(\mathcal{X}) = 0$, $\mu_+(\mathcal{X}) = \mu_-(\mathcal{X})$. Hence, $\|\mu\| = 2\mu_+(\mathcal{X})$. We thus need to show that $\mu_+(\mathcal{X}) = \sup \mu(A)$ where the supremum is taken over $A \in \mathcal{B}$.

By the Hahn decomposition, we can write

$$\mu(A) = \mu((A \cap P) \cup (A \cap N)) = \mu(A \cap P) - \mu(A \cap N).$$

Since $\mu(A \cap N)$ is a non-negative measure and P is a subset of \mathcal{X} , we have

$$\mu(A) \leq \mu(A \cap P) \leq \mu(P) = \mu_+(\mathcal{X}).$$

As the supremum is the smallest upper bound, we have $\sup \mu(A) \leq \mu_+(\mathcal{X})$.

Now, suppose that $\sup \mu(A) < \mu_+(\mathcal{X})$ and let $\epsilon = \mu_+(\mathcal{X}) - \sup \mu(A)$. Since the measure μ is regular, for the positive set $P \subset \mathcal{X}$, there exist $A_{\epsilon/2}$ such that $\mu(P) - \mu(A_{\epsilon/2}) \leq \epsilon/2$. But $\mu(P) = \mu_+(\mathcal{X})$ hence we have $\mu_+(\mathcal{X}) - \mu(A_{\epsilon/2}) \leq \epsilon/2$.

Thus $\sup \mu(A) < \mu(A_{\epsilon/2})$, which contradicts our assumption of strict inferiority. ■

Proposition A.2.9. *[probably not useful - l.141] Consider $(\mathcal{X}, \mathcal{B})$, a Borel probability*

space. Let μ_1, μ_2 be two atomic probability measures on \mathcal{X} . Then,

$$\sup_A |\mu_1(A) - \mu_2(A)| = \sum_{x \in \mathcal{X}} |\mu_1(x) - \mu_2(x)|. \quad (\text{A.2.1})$$

where the supremum is taken over all $A \in \mathcal{B}$.

Note that the support of an atomic probability measure on \mathcal{X} contains at most a countable number of elements. Hence we can rewrite equation A.2.1 as

$$\sup_A |\mu_1(A) - \mu_2(A)| = \frac{1}{2} \sum_{i=1}^n |\mu_1(x_i) - \mu_2(x_i)|.$$

Proof: (Proposition A.2.9 - 1.155)

Let $A \in \mathcal{B}$ and define $B = \{x \in \mathcal{X} : \mu_1(x) \geq \mu_2(x)\}$. Then

$$\begin{aligned} \mu_1(A) - \mu_2(A) &= (\mu_1(A \cap B) + \mu_1(A \cap B^c)) - (\mu_2(A \cap B) + \mu_2(A \cap B^c)) \\ &= (\mu_1(A \cap B) - \mu_2(A \cap B)) + (\mu_1(A \cap B^c) - \mu_2(A \cap B^c)) \end{aligned}$$

But for any $x \in B^c$, we have $\mu_1(x) - \mu_2(x) < 0$. Hence,

$$\begin{aligned} \mu_1(A) - \mu_2(A) &\leq \mu_1(A \cap B) - \mu_2(A \cap B) \\ &\leq (\mu_1(A \cap B) - \mu_2(A \cap B)) + (\mu_1(A^c \cap B) - \mu_2(A^c \cap B)) \\ &= \mu_1(B) - \mu_2(B). \end{aligned}$$

Likewise, using the fact that, for any $x \in B$ we have $\mu_2(x) - \mu_1(x) \leq 0$, we obtain

$$\mu_2(A) - \mu_1(A) \leq \mu_2(A \cap B^c) - \mu_1(A \cap B^c) \leq \mu_2(B^c) - \mu_1(B^c).$$

Now, if $\mu_1(A) - \mu_2(A) \geq 0$, we have $|\mu_1(A) - \mu_2(A)| \leq \mu_1(B) - \mu_2(B)$. If $\mu_1(A) - \mu_2(A) < 0$, we have $|\mu_1(A) - \mu_2(A)| \leq \mu_2(B^c) - \mu_1(B^c)$. Hence we obtain

$$|\mu_1(A) - \mu_2(A)| \leq \frac{1}{2} (\mu_1(B) - \mu_2(B) + \mu_2(B^c) - \mu_1(B^c))$$

$$|\mu_1(A) - \mu_2(A)| \leq \frac{1}{2} \left(\sum_{x \in B} \mu_1(x) - \mu_2(x) + \sum_{x \in B^c} \mu_2(x) - \mu_1(x) \right)$$

$$|\mu_1(A) - \mu_2(A)| \leq \frac{1}{2} \left(\sum_{x \in B} |\mu_1(x) - \mu_2(x)| + \sum_{x \in B^c} |\mu_1(x) - \mu_2(x)| \right)$$

$$|\mu_1(A) - \mu_2(A)| \leq \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu_1(x) - \mu_2(x)|.$$

As the supremum over all $A \in \mathcal{B}$ is the smallest upper bound, we have

$$\|\mu_1 - \mu_2\|_{TV} \leq \sum_{x \in \mathcal{X}} |\mu_1(x) - \mu_2(x)|.$$

Now, if we choose A to be equal to B , we have the equality

$$|\mu_1(A) - \mu_2(A)| = \frac{1}{2} (\mu_1(A) - \mu_2(A) + \mu_2(A^c) - \mu_1(A^c)).$$

Thus, for $A = B$, the upper bound is attained, i.e.

$$|\mu_1(A) - \mu_2(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu_1(x) - \mu_2(x)|.$$

Therefore,

$$\sup_A |\mu_1(A) - \mu_2(A)| = \sum_{x \in \mathcal{X}} |\mu_1(x) - \mu_2(x)|.$$

■

Appendix B

A Quick Overview of Feature Selection Algorithms

This Appendix introduces concepts and algorithms of feature selection, and groups different algorithms with a categorisation based on search strategies and evaluation criteria. The information and the text presented relies heavily on the work of Liu and Yu [36], Jovic *et al.* [28] and Liu *et al.* [35].

Feature selection algorithm is a process that selects a subset of an original set of features. Typically, a feature selection process consists of four basic steps: subset generation, subset evaluation, stopping criterion and result validation. In this Appendix, we only consider the first three steps as the fourth one is

B.0.1 Subset Generation

To generate a subset, one has to build a search algorithm which specifies a candidate subset for evaluation at each branching step. The nature of a search engine is determined by two choices: the choice of search strategy and the choice of search starting point (or points). Note that the search starting point influences the search direction.

There are four search starting points that imply four search directions: one may start with an empty set (and then add features); one may start with the complete

set (and then remove features); one may start from both sides (and simultaneously consider smaller and larger feature subsets; or one may start with a subset generated by a heuristic function.

Regarding the search strategies, one can categorize them in three groups: complete search, sequential search and randomized search.

A complete search guarantees to find the optimal result with respect to the chosen evaluation criterion. The Exhaustive search is complete but is computationally prohibitive as a set of r elements has $C(r, k)$ subsets of r elements. Using appropriate feature selection evaluation functions (as defined in Chapter 7) one can reduce the search space without jeopardizing the completeness of the search engine.

A Sequential search is essentially trading its optimality for computational efficiency. Indeed, sequential search algorithm are usually $O(n^2)$ or less. A sequential search adds (respectively, removes) p features in one step and removes (respectively, adds) q features in the next step (with $p > q$). These search processes are suboptimal due to the fact that the best pair of features need not contain the best single feature [27]. In general, good, larger feature sets do not necessarily include good small sets.

Randomized search methods incorporate randomness into their search procedure in order to avoid being trapped in local optima of the search space. The randomness is introduced in the search engine by choosing the first, many, or all subset(s) randomly.

B.0.2 Subset Evaluation

Each subset generated by the search algorithm needs to be evaluated by a feature selection evaluation function, also known as a feature selection criterion function. Clearly, the notion of “goodness” of a subset is dependent on the choice of the evaluation function: one optimal subset selected using one evaluation function may not be optimal with another evaluation function. We note that the mathematical “requirements” for evaluation functions are quite limited. Thus, in practice, the real difficulty is to construct a feature selection evaluation function that represents the intrinsic characteristics of interest in our subset.

B.0.3 Stopping Criteria

A stopping criteria determines when the feature selection process should be terminated. According to Liu and Yu [36], some frequently used stopping criteria are:

1. The search completes
2. Some given bound is reached, where a bound can be a specified number (minimum number of features or maximum number of iterations).
3. Subsequent addition (or deletion) of any feature does not produce a better subset.
4. A sufficiently good subset is selected (e.g., a subset may be sufficiently good if its classification error rate is less than the allowable error rate for a given task).

B.0.4 Result Validation

In real-world applications, the result validation step consists in monitoring the change of classification performance with respect to the subset of features obtained after the feature selection process has stopped (that is, after the third step B.0.3). Note that result validation is useful only in the case of independent evaluation functions. In the case of dependent evaluation functions, result validation is redundant.

Appendix C

Biological and Technical Background

The purpose of Appendix C is to introduce the biological and technical concepts which are necessary to understand how a GWAS dataset is constructed. The appendix is organized as follows. The first section gives a short summary on the biology of coronary artery disease. The second and third sections introduce the concepts of DNA and SNPs from genetics, followed by a brief definition of a GWAS in section 4. Section 5 explains why the structure of the dataset OHGS-2 is as such. It is the most important section as one of the central ideas in this thesis rely on the particular structure of OHGS-2.

C.1 Coronary Artery Disease

The cardiovascular system (CVS), composed of the heart, blood, and blood vessels is an organ system that allows blood to circulate through the body. In general, oxygenated blood travels in arteries whereas deoxygenated blood flows in veins. In addition to delivery of oxygen to cells, blood is essential for nutrient delivery, metabolic waste collection, acid-base status and circulation other molecules such as immune cells and hormones. Blood is pumped through the body by the heart which gets its

vascularization through coronary arteries.

Coronary artery disease (CAD) is defined as the stenosis (narrowing) or blocking of coronary arteries. The pathophysiology of CAD is linked to atherosclerosis in coronary arteries. Atherosclerosis is defined as the accumulation of atherosclerotic plaques within the walls of arteries. The earliest but asymptomatic manifestation of these plaques, fatty streaks, can be seen by age 20. Following injury to the endothelium (vessel inner lining), an inflammatory response is triggered leading to accumulation of low density lipoproteins (a type of cholesterol) and leukocytes (white blood cells) within the vessels walls. These leukocytes eventually transform into foam cells which are prominent inflammatory participants. Over time, as these fatty streaks progress to plaques, the lipid (fat) pools are covered by a protective layer of smooth muscle cells. At a given time t , these plaques can be either “stable” or “vulnerable” when the lipid pool is large, the fibrous cap is thin and abundant inflammatory cells are present. Even though, over a patient’s life, most plaques remain asymptomatic, complications may occur. These complications can lead to different clinical pictures depending on the type of plaque and their location. As this thesis focuses on CAD, we restrict our interest to atherosclerotic plaques within the coronary arteries. In the coronary arteries, stable plaques that narrow the vessel lumen lead to intermittent shortness of breath and chest pain on exertion (stable angina pectoris). Vulnerable plaques are subject to rupture, leading to acute thrombus (clot) formation. The magnitude of the thrombotic response that follows a plaque’s rupture is extremely variable. It only occasionally evolves into a life-threatening luminal clot [8]. If immediate reperfusion is not achieved, major luminal thrombus can lead to the death of the affected tissues and severe impairment of heart function. This process is the myocardial infarction (MI), also known as a heart attack. MI can lead to cardiac arrest and death.

C.2 DNA and SNPs

Deoxyribonucleic acid (DNA) is the molecule that contains the genetic information unique to each specie. A nucleotide is the building block of DNA. It is composed of

a sugar group called deoxyribose, a phosphate group and a base, the varying unit. Four types of bases exist: Cytosine (*C*), Guanine (*G*), Adenosine (*A*) and Thymine (*T*). A DNA strand is formed by the alignment of nucleotides via covalent bonds between the sugar group of one nucleotide and the phosphate group of the following nucleotide, leading to an alternating sugar-phosphate backbone. The DNA double helix is formed by two strands of DNA coiled around each other. These two strands are held together by (hydrogen) bonds between bases. Adenine always forms a base pair with Thymine and Cytosine always binds to Guanine.

In humans, the majority of the 3.08 billion base pairs of DNA is stored in the cells' nucleus subdivided into 46 different molecules, corresponding to the 46 chromosomes. Each chromosome contains 44 to 246 million base pair, corresponding to 1.5 to 8.4 cm of DNA. Depending on the phase of the cell's cycle, DNA is compacted to different degrees to fit in a 90 μm cell nucleus. In its most compact form during cell division, condensed chromosomes measure from 1.3 to 10 μm .

Most cells of a human body are called somatic cells and contains 46 chromosomes. By contrast, germ cells contain only 23 chromosomes. A cell with 46 chromosome is called a diploid cell, whereas a cell with 23 chromosomes is an haploid cell. Diploid cells contain 2 copies of the same chromosome whereas haploid cells contain only one copy. During fecundation, when a maternal (oocyte) and paternal (spermatozoid) germ cell come together (fertilization) the resulting zygote inherits 2 copies of each chromosome, one from its mother and one from its father, making a 46 chromosome cell.

As mentioned above, DNA contains the biological information that defines a specie and within the same species, an individual. At the molecular level, a gene is a segment of DNA that codes for a specific protein (with some exceptions). Between genes, there is abundant non coding DNA, once called junk DNA. Although it does not contain information to synthesize a protein, non-coding DNA plays a key role in regulation of gene expression, cell growth, cell division etc. Between any two individuals, the genome is at least 99.5 % identical. The genetic variations between

two individuals are the consequences of germline mutations (mutations occurring in the gametes) in their respective ancestry: Embryos are derived from the fusion of an egg and a sperm hence a mutation in at least one of the parent's fused gametes is then found in each nucleated cell of their offspring. Mutations usually arise from unrepaired DNA damage, DNA replication errors or mobile genetic elements [29]. Geneticists distinguish between three major classes of mutations: single nucleotide substitutions, insertion/deletion of one or many (up to twenty) nucleotides and large genomic rearrangements. Note that there also exist rare types of mutation mechanism that do not belong to any of the three classes [25]. A variation in the DNA sequence that occurs in a population with a frequency of 1 % or higher is termed a polymorphism. Polymorphisms, like mutations, can be of one or more nucleotide changes. The commonest polymorphism is the single nucleotide polymorphism (SNP). All existing variants of a SNP in a population are referred to as alleles. The allele that is observed more frequently in the population is termed the major allele, while the less frequent variant is the minor allele. Most (over 99.6%) SNPs are bi-allelic but 271'934 tri-allelic SNPs were identified in the 1000 Genomes Phase III variation data [42]. Humans, as diploids, always have a pair of alleles. If one has two copies of the major allele, he is said to be homozygous major; if he has two copies of the minor allele, he is said to be homozygous minor and having one copy of each allele makes him heterozygous. For simplicity, geneticists often label the two alleles as A and B . Therefore, an individual's genotype for a given gene is either AA , BB or AB .

C.3 DNA Microarrays

DNA microarrays rely on hybridization, the biochemical principle that nucleotide bases bind to their complementary partners (A binds to T and C binds to G). DNA microarrays are silicon biochips on which a collection of microscopic DNA spots are attached. Each DNA spot contains picomoles of a specific DNA sequence, called probes.

There are many different types of arrays used for different applications. To detect

particular SNPs among alleles within a population, geneticists use SNP arrays. There are 2 main array manufacturers that commercially produce SNP arrays: Affymetrix and Illumina. We will concentrate on the Affymetrix technology as the data analysed in this these was produce with the Affymetrix 6.0 chip.

The Affymetrix 6.0 chips allow for the detection of 906,600 SNPs. Every SNP site is interrogated by a set of 6 to 8 probes; 3 or 4 replicates of the same probe for each of the 2 alleles. For Affymetrix SNP arrays, a probe is a short 25 base pair fragment of DNA design to be complementary to the sequence harbouring a particular SNP and featuring one of the 2 alleles of that SNP. The SNP, being centered on the probe, is base pair number 13. When using an Affymetrix 6.0 chip, the acquired DNA is first purified, then amplified. The amplification is done via repeated reverse transcription. Those multiple single-strand DNA copies are then coupled with a label. The label is typically a fluorophore, a fluorescent chemical compound. A fluorophore has the property to emit light that it has just previously absorbed. Generally, the emitted light has a longer wavelength than the absorbed light. For SNP arrays, the fluorophores absorb ultra-violet radiation and emits light in the visible region of the wavelength spectrum. Now that the single strand DNA copies are labelled, they are broken down in small fragments. These fragments are then mixed with a propriety hybridization solution and the mixture is poured in the pinholes of the microarray. The pinholes are then sealed and the microarray is mixed in order for the small fragments to be uniformly distributed over the microarray probes. By the hybridization principle, labelled single-strand fragment of DNA will bind preferentially to probes that are complementary to themselves. After a defined hybridization period, the microarray is washed off and dried. This eliminates all non-binded DNA fragments and only the strongly hybridized ones remain. The microarray is then scanned by a machine that emits ultra-violet radiation and produces a measure of the signal intensity associated with each probe set. The underlying principle is that the signal intensity depends upon the number of targeted fragments perfectly hybridized to their complementary probe. For example, suppose a patient is homozygous AA for a given SNP. Then, the

DNA fragments in the hybridization solution only contains allele A. Therefore, the fragments will only bind to the 4 probes harbouring the complementary sequence to allele A. Likewise, for a homozygous BB patient, the DNA fragment would only bind to the 4 probes harbouring the complementary sequence to allele B. For a heterozygous AB patient, DNA fragments would bind to all 8 probes.

In reality, this is an oversimplification of the facts. First, the signal intensity also depends upon the amount of targeted fragments in the hybridization mixture. Indeed, even if the perfect complimentary probes of a given target fragment exist on the microarray, the probability that they will hybridize is lower than 1 and decreases as the concentration of the targeted fragments in the hybridization mixture decreases. Also, for the 2 different allelic version of the probes, the 24 other bases remain exactly the same. Thus, the 2 versions of the probes are identical on 96% of their DNA sequence. Hence, the probability that a fragment with allele A will bind to a probe that contains allele B is rather high. Moreover, the hydrogen bounds between the DNA bases C and G are stronger than the bound between A and T. Therefore, a target sequence rich in CG bounds will be more likely to bind to the probe of the opposite allele. Despite these limitations, commercial SNP array can now genotype SNPs with an accuracy over 95% [32].

C.4 Chiamo: a Genotype Calling Algorithm

In order to yield SNP genotype inferences from the raw intensity associated with each probe set, extensive processing and analysis is required. This computational analysis is done by genotype calling algorithms. With each new version of a SNP microarray, the computational community (biostatisticians, bioinformaticians) responds with a new algorithmic development. These developments are driven both by the microarray manufacturers and academics. It is interesting to note that there has been a productive synergistic relationship between advances in biological understanding, computational methodology and technological advancements. Progress in each of these three areas has spurred progress in the other two. For example, the choice of

SNPs to include on a microarray is influenced by the genotype calling algorithm: the SNPs chosen are the SNPs for which the current algorithm performs best [32].

Genotype calling algorithms are composed of two steps. The first step is the normalisation of the raw data and the second step is the calling algorithm per se. The output of the calling algorithm is, as stated above, an inference on the individual's genotype at the SNP of interest. Such an inference is known as the genotype call. Genotype calling algorithms are applied after normalization of the raw data. Since Genotype calling algorithms examine probe intensities, it is crucial to standardize these intensities to take into account technological differences such as platform fabrication. Normalization aims to correct these biases by homogenizing the intensity distribution of each array. Currently, the consensus settles on a non-parametric method that ensures that all arrays have the identical probe intensity distribution. That method is called Quantile Normalization. Once the raw data has been normalized, it can be genotyped. In the case of the Affimetrix SNP array 6.0, the default genotype calling algorithm is Birdseed, developed by J. Korn *et al.* [31]. In 2011, Robert Davies, at the UOHI, modified the normalisation step of Chiamo, a genotype calling algorithm developed for Affimetrix 500k arrays in order to apply it to SNP array 6.0 data. Robert Davies then compared the outputs of both Chiamo and Birdseed and concluded that "Chiamo is slightly more accurate than Birdseed" [19]. No further results or explanation are given for this statement.

Here is a brief overview of Robert Davies' modified version of Chiamo. The genotype calling algorithm Chiamo [16] was developed in 2007 by Peter Donnelly, Jonathan Marchini, Chris Spencer and Yik Ying Teo, all members of The Wellcome Trust Case Control Consortium at the time. As stated above, each SNP on the SNP array 6.0 is interrogated by 6 or 8 probes - 3 or 4 replicates of the same probe for each of the two alleles. Hence, intensity data for each SNP consists of two sets of 4 repeated measurements: The raw data for the s^{th} SNP of the i^{th} array (thus individual) can be denoted as $I_{isk} = (I_{isk}^A, I_{isk}^B)$, where k represents the k^{th} probe. Hence $k = 1, \dots, K$ with $k \in \{3, 4\}$. First, the vectors I_{sk} are normalised using

the pre-processing step of quantile normalization. Second, the quantile normalized intensities are log transformed to reduce their skewness. Let $Y_{iks} = (Y_{iks}^A, Y_{iks}^B)$ denote the log transformed intensities. The third step is the probe set summarization during which the signals are combined across all k probes using an arithmetic mean to create a single pair of intensities $X_{is} = (X_{is}^A, X_{is}^B)$ for the i^{th} individual at SNP s . That is:

$$X_{is}^A = \frac{1}{n_k} \sum_{k=1}^{n_k} Y_{iks}^A \quad \text{and} \quad X_{is}^B = \frac{1}{n_k} \sum_{k=1}^{n_k} Y_{iks}^B.$$

Chiamo is applied after quantile normalization of the data from each SNP. For a given SNP s , it uses a two-dimensional four-stage-hierarchical Bayesian Gaussian mixture model to call genotypes. Note that both the component parameters and the mixture weights are estimated with a four-stage Bayesian hierarchical model. It is natural for Chiamo to be based on a mixture model since a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population. The distribution for the set of bi-variate intensity vectors X_{is} is a two-dimensional four-stage-hierarchical Bayesian Gaussian mixture distribution with four components. It is important to keep in mind that a n -dimensional Gaussian mixture distribution with m components need not be a m -stage-hierarchical Bayesian distribution. The Gaussian mixture distribution is given by a linear combination of four two-dimensional normals of mean $\mu_{k,s} \in \mathbb{R}^2$ and covariance $\Sigma_{k,s} \in \mathcal{M}_{2 \times 2}$. That is,

$$X_{is} | \theta_s \sim \sum_{k=1}^4 \lambda_{k,s} N_2(\mu_{k,s}, \Sigma_{k,s}), \quad (\text{C.4.1})$$

where θ_s is the vector of parameters (note that $\lambda_{k,s}$, μ_k and Σ_k are parameters).

To specify the identity of the mixture component of observation x_{is} , one introduces a four-dimensional binary random variable Z_{is} having a one-in-four representation in which a particular element z_k is equal to 1 and all other elements are equal to zero. The random variable Z_{is} is called a latent random variable. The values of Z_{is} thus satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$ and Z_{is} follows a multinomial distribution

of parameters $(1, \lambda_s)$ where $\lambda_s \in \mathbb{R}^4$. With the latent variable Z_{is} , we obtain the following likelihood function for X_{is} :

$$X_{is}|\theta_s, Z_{is} \sim N_2(\mu_{Z_{is},s}, \Sigma_{Z_{is},s}) \quad \text{equivalent to} \quad X_{is}|\theta_s, Z_{is} \sim \prod_{k=1}^4 N_2(\mu_{k,s}, \Sigma_{k,s})^{z_k}.$$

Donnelly *et al.* [16] use the latent variable Z_{is} to denote the genotype call for the i^{th} patient of snp s . There is four different possible calls: one for each of the genotypes $\{AA, AB, BB\}$ as well as a null call to capture the clear outliers and add robustness to the model fit of the other three genotype components. This format allows for genotype uncertainty. Statistically, the probability of each genotype call, given the pair of intensities X_{is} , is the posterior distribution $P(Z_{is}|X_{is}, \hat{\theta}_s)$, where $\hat{\theta}_s$ is the maximum a posteriori estimate of θ_s . The exact iterative method to obtain the MAP estimate of the parameters θ_s and to compute the posterior distribution is not specified in the paper by Donnelly *et al.* [16] but it is very likely that the Expectation-Maximization (EM) algorithm or one of its variants was employed. The EM algorithm seems the most plausible algorithm since it explicitly computes the posterior distribution of the latent variables in its step E. Indeed, the EM algorithm is an iterative process with the following steps:

- E step: Evaluate the posterior distribution of the latent variables $P(Z_{is}|\theta_{n-1}, X_{is})$, where θ_{n-1} are the parameter values obtain at level $n - 1$.
- M step: Determine the revised parameter estimate θ_n given by $\theta_n = \text{Argmax}_{\theta} \mathcal{Q}(\theta_{n-1}, \theta)$, where \mathcal{Q} represents the expectation of the complete-data log likelihood evaluated with respected to θ :

$$\mathcal{Q}(\theta, \theta_{n-1}) = \sum_i P(Z_{is}|X_{is}, \theta_{n-1}) \ln P(X_{is}, Z_{is}|\theta).$$

For an in-depth coverage of the EM algorithm (as well as Gaussian mixture models) refer to chapter 9 in *Pattern Recognition and Machine Learning* [11].

A short-coming of the EM algorithm is that the sequence of complete-data log likeli-

hoods may converge to a local maximum depending on the initial parameter values. To avoid that problem, Donnely *et al.* ran the EM algorithm twelve times with twelve different random starts. That is, they ran the algorithm with twelve different set of initial parameter values θ_0 .

A visual representation of the data is obtained by plotting each pair of normalized summary probe intensities (X_{is}^A, X_{is}^B) in the plane. If the given SNP has been genotyped well, one should see three clear, distinct, clusters on the plot that would correspond to the three genotypes. The next step is to colour each point to indicate how the genotype calling algorithm Chiamo classifies that individual (either as a homozygote for one of the two alleles, a heterozygote, or a null (missing) call). Such coloured scatter plot in the plane is called cluster plot. A cluster plot is therefore a graphical representation of the results of both the genotyping and genotype calling of a SNP, with each point representing one individual. See Figure 3 and Supplementary Figure 1 in [31], for example of cluster plots.

It is important to keep in mind that the output of Chiamo is not a genotype call per se but the probability of each genotypes. Hence, for a given SNP s , the output for patient i is a vector of \mathbb{R}^4 where the k^{th} component is $P(Z_{is} = z_k | X_{is}, \hat{\theta})$ with $Z_{is} = z_k$ being an abusive notation for the vector with the k^{th} component equal to 1 and all others zero. Since Z_{is} follows a 4-multinomial distribution and thus $\sum_k P(Z_{is} = z_k) = 1$, the probability $P(Z_{is} = z_4 | X_{is}, \hat{\theta})$ of a null call is not explicitly given in the output. The output of a complete genotype file from Chiamo has a one-line-per-snp format: The first 5 entries of each line should be the SNP id, rs id of the SNP, base-pair position of the SNP, the allele coded A and the allele coded B. The SNP id can be used to denote the chromosome number of each SNP. The next three numbers on the line should be the probabilities of the three genotypes AA, AB and BB at the SNP for the first individual in the cohort. The next three numbers should be the genotype probabilities for the second individual in the cohort, and so

on. Therefore, from a mathematical stand point, the output of a complete genotype file is a matrix of dimension $n \times 3m$ where n is the number of SNPs and m is the number of patients. Thus, for a row i , each triple $(a_{i,3k}, a_{i,3k+1}, a_{i,3k+2})$ represents the probabilities of patient k having genotypes AA, AB and BB for SNP i . Note that, for a patient k , the sum $a_{i,3k} + a_{i,3k+1} + a_{i,3k+2}$ is not necessarily equal to 1 since the null call is not given in the output.

C.5 Quality Control

Quality Control (QC) in the case of GWAS, is the process of ensuring that the data obtained by a genotype calling algorithm is of acceptable quality. QC removes results that are likely to be inaccurate. It is performed before any analysis on the data since errors in genotype calling have the potential to introduce “systematic biases into genetic case-control association studies, leading to an increase in the number of false positive associations” [3].

QC is a multi-layered process that removes patients or SNPs. It is important to note that the “layering” is not commutative: the order of the steps affects the end result. The different steps can be grouped in two categories: the “per-patient” QC and the “per-SNP” QC. In their paper *Data Quality Control in Genetic Case-Control Association Studies*, Anderson *et al.* lay out a very clear protocol that details all the steps typically carried out during QC for a GWAS dataset. In the context of this thesis, it is important to remember that “step 0” of QC categorizes, for each patient k and SNP i , the triple $(a_{i,3k}, a_{i,3k+1}, a_{i,3k+2})$ as either “called” or “missing” (or “uncalled”). For a fixed threshold $\alpha \in [0, 1]$, a triple $(a_{i,3k}, a_{i,3k+1}, a_{i,3k+2})$ is considered called if $a_{i,3k} + a_{i,3k+1} + a_{i,3k+2} \geq \alpha$ and missing if $a_{i,3k} + a_{i,3k+1} + a_{i,3k+2} < \alpha$. In the landmark paper *Genomewide Association Study of 14000 cases of seven common diseases and 3000 shared controls* [16], the threshold α has a value of 0.9, following an “analysis of the relationship between concordance and missing data rates”. The data for this analysis is not shown. After their QC, the called rate for their data was 99.63%. Hence, the vast majority of the triples were called. Uncalled triples were

removed from the dataset.

C.6 Information on OHGS Dataset

Genome Wide Association Studies (GWAS) are defined as observational studies of a genome wide set of genetic variants in individuals to investigate if any variant is statistically more prevalent for a given trait. The power of SNP arrays to interrogate a significant number of SNPs both rapidly and cheaply has made SNPs the preferred genetic variants for GWAS. The most common design for GWAS is the case-control setup in which a chosen trait is used to split a large sample of individuals in two groups: the case group “affected” by the trait and the control group, without the trait. For each SNP, one then investigates if the allele frequency, or the frequency of a combination of allele, is significantly different between the case and control group in order to detect evidence of association with the trait. The statistical methodology most commonly used for testing genetic association of case-control data is covered in *A Tutorial on statistical methods for population association studies* [4].

The Ottawa Heart Genomics Study (OHGS) is a GWAS based out of the University of Ottawa Heart Institute (UOHI). The OHGS was divided in two separate studies: OHGS-1 and OHGS-2. For the purpose of this thesis we will focus on OHGS-2. The OHGS-2 is of case-control design with 1929 cases and 1978 controls. Subjects were recruited from the University of Ottawa Heart Institute (UOHI) lipid clinic, catheterization laboratory, or from the Cleveland Clinic. The inclusion criteria for cases were given in R. Davies Master’s thesis [18]:

“Inclusion criteria for cases was set as having had either a myocardial infarction, coronary revascularization (coronary angioplasty/percutaneous coronary intervention or coronary artery bypass graft) or had angiography or computed tomography angiography demonstrating stenosis of at least 50% in at least one coronary artery. Age limits for cases were originally set at ≤ 55 for men and ≤ 65 for women; however, several cases were included which did not meet this criteria: 22 men aged ≥ 56 , 4 women aged ≥ 66 , 6 men of unknown age at onset, 4 women of unknown age at

onset. Cases were excluded if they had diabetes mellitus or overt hyperlipidemia and if they had nonEuropean ancestry. Two sets of inclusion criteria were set up for controls, due to different acquisition protocols. One set of controls were recruited with inclusion criteria set as being healthy and to have a lack of cardiovascular disease history. For the other set of controls, recruited through the catheterization lab at the UOHI, inclusion criteria was set as having an angiogram which showed that none of the coronary arteries had a stenosis encompassing greater than 50% of the vessel. Age cutoffs for controls were originally set as men aged 65 and older, women 70 and older; however, several controls were included which did not meet this criteria: 5 men aged ≤ 64 and 30 women aged ≤ 69 .”

For both cases and control, the genetic profile of each individual was obtained with Affimetrix 6.0 arrays. Although 6.0 arrays genotype 946,000 SNPs, the .CEL files containing the raw dataset only contains 894,240 SNPs, from chromosomes 1 to 22. Note that SNPs from the two sex chromosomes X and Y are not considered. The SNP count for each chromosome is given in the table below.

The raw dataset was then genotyped by R.W. Davies using his own modification of the genotype calling algorithm Chiamo. The genotype files were saved under OHGS_B2.i.6.0.gen (where $i = 1, \dots, 22$) and stored on the OHI_GA_DRIG, a blade server physically residing within the department of IT at UOHI.

Recall that (see section C.4), for a given SNP, Chiamo outputs a patient’s genotype as a triple representing the probability of each genotype. Consequently, the dataset is stored in two matrices. The matrix for the cases is of dimension 1929×2682720 . The matrix for the controls is of dimension 1978×2682720 (where $2682720 = 3 \times 894240$). For a given row i , each triple $(a_{i,3k}, a_{i,3k+1}, a_{i,3k+2})$, for $k = 1, \dots, 894240$, represents the probabilities of patient i having genotype *homozygous major*, *heterozygous*, and *homozygous minor* for SNP k . Since there are only 3 possible genotypes, $a_{i,3k}, a_{i,3k+1}, a_{i,3k+2} \in [0, 1]$ and $a_{i,3k} + a_{i,3k+1} + a_{i,3k+2} = 1$.

Chromosome	1	2	3	4	5	6	7	8
# of SNPs	73571	75918	62268	57582	57971	57687	48380	50026
Chromosome	9	10	11	12	13	14	15	16
# of SNPs	42785	49600	45927	43802	34979	28936	26907	28552
Chromosome	17	18	19	20	21	22		
# of SNPs	21319	27212	12422	23488	12924	11984		

Table C.1: Information on the number of SNPs, from each of Chromosomes 1 to 22, in the OHGS-2 dataset.

Appendix D

A geometrical perspective of the Kantorovich-Rubinstein Distance

In this thesis, we have often use the fact that, for a finite optimal function $f : \mathcal{X} \rightarrow [0, \Delta]$, the Kantorovich-Rubinstein distance $W_{\mathcal{X}}(\mu_1, \mu_2)$ can be expressed as the area between the distribution functions F_1 and F_2 (where $F_i(s) = \mu_i(\{x \in \mathcal{X} : f(x) \leq s\})$). In this appendix, using basic properties of product measures, we construct the geometric regions of the $[0, \Delta] \times [0, 1]$ rectangle which, added together, give the region between the distribution functions F_1 and F_2 .

Let $\gamma \in \mathbb{R}$ and $E_\gamma = \{x \in \mathcal{X}; f(x) \geq \gamma\}$. Then we have:

$$\begin{aligned} W_{\mathcal{X}}(\mu_1, \mu_2) &= \int_{E_\gamma} f \, d(\mu_1 - \mu_2) + \int_{E_\gamma^c} f \, d(\mu_1 - \mu_2) \\ &= \int_{E_\gamma} f \, d\mu_1 - \int_{E_\gamma} f \, d\mu_2 + \int_{E_\gamma^c} f \, d\mu_1 - \int_{E_\gamma^c} f \, d\mu_2. \end{aligned}$$

Now, suppose that $f : \mathcal{X} \rightarrow [0, \Delta]$ and $\gamma \in [0, \Delta]$. Thus, $W_{\mathcal{X}}(\mu_1, \mu_2) = \int_0^\Delta f \, d(\mu_2 - \mu_1)$.

We first consider $\int_{E_\gamma} f \, d\mu_i$.

Using properties of product measures (see p.162, [15]):

$$\int_{E_\gamma} f \, d\mu_i = \int_0^\gamma \mu_i(\{x \in E_\gamma; f(x) > y\}) \, dy + \int_\gamma^\Delta \mu_i(\{x \in E_\gamma; f(x) > y\}) \, dy.$$

Consider the first term: $\int_0^\gamma \mu_i(\{x \in E_\gamma; f(x) > y\}) \, dy$. The variable y is such that $y \leq \gamma$. Then:

$$\begin{aligned} \{x \in E_\gamma; f(x) > y\} &= \{x \in [0, \Delta]; f(x) > y \wedge f(x) > \gamma\} \\ &= \{x \in [0, \Delta]; f(x) > \gamma\}. \end{aligned}$$

Hence, $\int_0^\gamma \mu_i(\{x \in E_\gamma; f(x) > y\}) \, dy = \int_0^\gamma 1 - F_i(\gamma) \, dy$. Geometrically, it corresponds to the rectangle with a length between $F_i(\gamma)$ and 1, and a width between 0 and γ .

Consider the second term: $\int_\gamma^\Delta \mu_i(\{x \in E_\gamma; f(x) > y\}) \, dy$. The variable y is such that $y > \gamma$. Then:

$$\begin{aligned} \{x \in E_\gamma; f(x) > y\} &= \{x \in [0, \Delta]; f(x) > y \wedge f(x) > \gamma\} \\ &= \{x \in [0, \Delta]; f(x) > y\}. \end{aligned}$$

Hence, $\int_\gamma^\Delta \mu_i(\{x \in E_\gamma; f(x) > y\}) \, dy = \int_\gamma^\Delta 1 - F_i(y) \, dy$. Geometrically, it corresponds to the area between the curves F_i and 1 on $[\gamma, \Delta]$.

We now consider $\int_{E_\gamma^c} f \, d\mu_i$.

Using properties of product measures (see p.162 of [15])

$$\int_{E_\gamma^c} f \, d\mu_i = \int_0^\gamma \mu_i(\{x \in E_\gamma^c; f(x) > y\}) \, dy + \int_\gamma^\Delta \mu_i(\{x \in E_\gamma^c; f(x) > y\}) \, dy.$$

Consider the first term: $\int_0^\gamma \mu_i(\{x \in E_\gamma^c; f(x) > y\}) dy$. The variable y is such that $y \leq \gamma$. Then:

$$\begin{aligned} \{x \in E_\gamma^c; f(x) > y\} &= \{x \in [0, \Delta]; f(x) \leq \gamma \wedge f(x) > y\} \\ &= \{x \in [0, \Delta]; y < f(x) \leq \gamma\}. \end{aligned}$$

Hence, $\int_0^\gamma \mu_i(\{x \in E_\gamma^c; f(x) > y\}) dy = \int_0^\gamma F_i(\gamma) - F_i(y) dy$. Geometrically, it corresponds to the area between the curves F_i and the horizontal line of equation $F_i(\gamma)$ on $[0, \gamma]$.

Consider the second term: $\int_\gamma^\Delta \mu_i(\{x \in E_\gamma^c; f(x) > y\}) dy$. The variable y is such that $y > \gamma$. Then

$$\begin{aligned} \{x \in E_\gamma^c; f(x) > y\} &= \{x \in [0, \Delta]; f(x) \leq \gamma \wedge f(x) > y \text{ with } y \geq \gamma\} \\ &= \{x \in [0, \Delta]; f(x) \leq \gamma < y < f(x)\} = \emptyset. \end{aligned}$$

Hence, $\int_\gamma^\Delta \mu_i(\{x \in E_\gamma^c; f(x) > y\}) dy = 0$.

Using the fact that

$$W_{\mathcal{X}}(\mu_1, \mu_2) = \int_{E_\gamma} f d\mu_1 + \int_{E_\gamma^c} f d\mu_1 - \left(\int_{E_\gamma} f d\mu_2 + \int_{E_\gamma^c} f d\mu_2 \right),$$

We can write the K-R distance:

$$\begin{aligned} W_{\mathcal{X}}(\mu_1, \mu_2) &= \left(\int_0^\gamma 1 - F_1(\gamma) dy + \int_\gamma^\Delta 1 - F_1(y) dy \right) + \int_0^\gamma F_1(\gamma) - F_1(y) dy \\ &\quad - \left(\left(\int_0^\gamma 1 - F_2(\gamma) dy + \int_\gamma^\Delta 1 - F_2(y) dy \right) + \int_0^\gamma F_2(\gamma) - F_2(y) dy \right). \end{aligned} \tag{D.0.1}$$

Now, using the fact that $\mu_i(\{x \in \mathcal{X}; f(x) > y\}) = 1 - \mu_i(\{x \in \mathcal{X}; f(x) \leq y\})$, we have

$$\begin{aligned} \int_{E_\gamma} f \, d\mu_i &= \int_0^\Delta \mu_i(\{x \in E_\gamma; f(x) > y\}) \, dy \\ &= \Delta - \int_0^\Delta \mu_i(\{x \in E_\gamma; f(x) \leq y\}) \, dy \\ &= \Delta - \left(\int_0^\gamma \mu_i(\{x \in E_\gamma; f(x) \leq y\}) \, dy + \int_\gamma^\Delta \mu_i(\{x \in E_\gamma; f(x) \leq y\}) \, dy \right) \end{aligned}$$

Consider the first term: $\int_0^\gamma \mu_i(\{x \in E_\gamma; f(x) \leq y\}) \, dy$. The variable y is such that $y \leq \gamma$. Then:

$$\begin{aligned} \{x \in E_\gamma; f(x) \leq y\} &= \{x \in [0, \Delta]; f(x) \leq y \wedge f(x) > \gamma \text{ with } y \leq \gamma\} \\ &= \{x \in [0, \Delta]; f(x) \leq y \leq \gamma < f(x)\} = \emptyset \end{aligned}$$

Hence, $\int_0^\gamma \mu_i(\{x \in E_\gamma; f(x) \leq y\}) \, dy = 0$.

Consider the second term: $\int_\gamma^\Delta \mu_i(\{x \in E_\gamma; f(x) \leq y\}) \, dy$. The variable y is such that $y > \gamma$. Then:

$$\begin{aligned} \{x \in E_\gamma; f(x) \leq y\} &= \{x \in [0, \Delta]; f(x) \leq y \wedge f(x) > \gamma\} \\ &= \{x \in [0, \Delta]; \gamma < f(x) \leq y\}. \end{aligned}$$

Hence, $\int_\gamma^\Delta \mu_i(\{x \in E_\gamma; f(x) \leq y\}) \, dy = \int_\gamma^\Delta F_i(y) - F_i(\gamma) \, dy$.

Geometrically, it corresponds to the area between the curves F_i and the horizontal line of equation $F_i(\gamma)$ on $[\gamma, \Delta]$.

We now consider $\int_{E_\gamma^c} f \, d\mu_i$.

$$\int_{E_\gamma^c} f \, d\mu_i = \Delta - \left(\int_0^\gamma \mu_i(\{x \in E_\gamma^c; f(x) \leq y\}) \, dy + \int_\gamma^\Delta \mu_i(\{x \in E_\gamma^c; f(x) \leq y\}) \, dy \right)$$

Consider the first term: $\int_0^\gamma \mu_i(\{x \in E_\gamma^c; f(x) \leq y\}) \, dy$. The variable y is such that $y \leq \gamma$. Then:

$$\begin{aligned} \{x \in E_\gamma^c; f(x) \leq y\} &= \{x \in [0, \Delta]; f(x) \leq \gamma \wedge f(x) \leq y \text{ with } y \leq \gamma\} \\ &= \{x \in [0, \Delta]; f(x) \leq y\}. \end{aligned}$$

$$\text{Hence, } \int_0^\gamma \mu_i(\{x \in E_\gamma^c; f(x) \leq y\}) \, dy = \int_0^\gamma F_i(y) \, dy.$$

Geometrically, it corresponds to the area under the curve F_i on $[0, \gamma]$.

Consider the second term: $\int_\gamma^\Delta \mu_i(\{x \in E_\gamma^c; f(x) \leq y\}) \, dy$. The variable y is such that $y \geq \gamma$. Then:

$$\begin{aligned} \{x \in E_\gamma^c; f(x) \leq y\} &= \{x \in [0, \Delta]; f(x) \leq \gamma \wedge f(x) \leq y \text{ with } y \geq \gamma\} \\ &= \{x \in [0, \Delta]; f(x) \leq \gamma\}. \end{aligned}$$

$$\text{Hence, } \int_\gamma^\Delta \mu_i(\{x \in E_\gamma^c; f(x) \leq y\}) \, dy = \int_\gamma^\Delta F_i(\gamma) \, dy.$$

Geometrically, it corresponds to the rectangle between the x-axis and the horizontal line $F_i(\gamma)$ on $[\gamma, \Delta]$.

We can write the K-R distance:

$$\begin{aligned} W_{\mathcal{X}}(\mu_1, \mu_2) &= \Delta - \int_\gamma^\Delta F_1(y) - F_1(\gamma) \, dy + \Delta - \left(\int_0^\gamma F_1(y) \, dy + \int_\gamma^\Delta F_1(\gamma) \, dy \right) \\ &\quad - \left(\Delta - \int_\gamma^\Delta F_2(y) - F_2(\gamma) \, dy + \Delta - \left(\int_0^\gamma F_2(y) \, dy + \int_\gamma^\Delta F_2(\gamma) \, dy \right) \right) \\ &= \int_\gamma^\Delta F_2(y) - F_2(\gamma) \, dy + \int_0^\gamma F_2(y) \, dy + \int_\gamma^\Delta F_2(\gamma) \, dy \end{aligned}$$

$$- \left(\int_{\gamma}^{\Delta} F_1(y) - F_1(\gamma) dy + \int_0^{\gamma} F_1(y) dy + \int_{\gamma}^{\Delta} F_1(\gamma) dy \right). \quad (\text{D.0.2})$$

Bibliography

- [1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 3rd edition, 2013.
- [2] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iterations. *Advances in Neural Information Processing Systems*, 30:1961–1971, 2017.
- [3] C.A. Anderson, F.H. Pettersson, G.M. Clarke, L.R. Cardon, A. P. Morris, and K.T. Zondervan. Data quality control in genetic case-control association studies. *Nat Protoc.*, 5:1564–1573, 2010.
- [4] D. Balding. A tutorial on statistical methods for population association studies. *Nature Review Genetics*, 7(10):781–791, 2006.
- [5] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
- [6] M. Basu and T. K. Ho. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300, 2002.
- [7] Y. Bengio, A. Courville, and I. Goodfellow. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, 1st edition, 2016.

-
- [8] J. F. Bentzon, F. Otsuka, R. Virmani, and E. Falk. Mechanisms of plaque formation and rupture. *Circulation Research*, 114(12):1852–1866, 2014.
- [9] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 3rd edition, 1995.
- [10] P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition, 1999.
- [11] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.
- [12] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer, 2004.
- [13] O. Bousquet and U. von Luxburg. Distance-based classification with lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.
- [14] C.A. Cabrelli and U.M. Molter. The kantorovich metric for probability measures on the circle. *J. Comput. Appl. Math.*, (57):345–361, 1995.
- [15] D. L. Cohn. *Measure Theory*. Birkhäuser Basel, 1st edition, 1980.
- [16] The Wellcome Trust Case Control Consortium. Genomewide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature*, 447:661–78, 2007.
- [17] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.
- [18] R. Davies. On the generation of a classification algorithm from dna based microarray studies. Master’s thesis, Université d’Ottawa, 2010.

- [19] R. Davies. Statistical genetics on ohi-ga-drig, 2011.
- [20] P. Devroye, L. Györfi, and G. Lugosi. *A Probability Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer, 1st edition, 1996.
- [21] J. Dieudonné. *Eléments d'analyse*, volume 2 of *Cahiers Scientifiques*. Gauthier-Villars, 1968.
- [22] H. Duan. Applying supervised learning algorithms and a new feature selection method to predict coronary artery disease. Master's thesis, Université d'Ottawa, 2014.
- [23] D.A. Edwards. On the kantorovich-rubinstein theorem. *Expositiones Mathematicae*, (29):387–398, 2011.
- [24] A. Frank, D. Geiger, and Z. Yakhini. A distance-based branch and bound feature selection algorithm. In *Proceedings of the Nineteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 241–248. Morgan Kaufmann, 2003.
- [25] N. Hanna, B. Parfait, D. Vidaud, and M. Vidaud. Mécanismes et conséquences des mutations. *M/S : médecine sciences*, 21(11):969–980, 2015.
- [26] S. Hatko. Borel isomorphic dimensionality reduction of data and supervised learning, 2013.
- [27] A.K. Jain, R.P.W. Duin, and J.C. Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.
- [28] A. Jovic, K. Brkic, and N. Bogunovic. A review of feature selection methods with applications. In *38th International Convention on Information and Communication Technology, Electronics and Microelectronics*. IEEE, May 2015.

- [29] R. Karki, D. Pandya, R. Elston, and C. Ferlini. Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Medical Genomics*, 8(37), 2015.
- [30] A.S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, 1994.
- [31] J.M. Korn, F.G. Kuruvilla, S.A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P.J. Collins, K. Darvishi, C. Lee, M.M. Nizzari, S.B. Gabriel, S. Purcell, M.J. Daly, and D. Altshuler. Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. *Nature Genetics*, 40(11):1253–1260, 2008.
- [32] T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational, and technological advances. *Nucleic Acids Research*, 37(13):4181–4193, 2009.
- [33] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer Berlin, Heidelberg, 1st edition, 1991.
- [34] C. Lindvall. *Lectures on the Coupling Method*. Wiley Series in Probability and Mathematical Statistics. A Wiley Interscience Publication, 1992.
- [35] H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Feature selection: An ever evolving frontier in data mining. In *Proceedings of the Fourth International Workshop on Feature Selection in Data Mining*, volume 10 of *Proceedings of Machine Learning Research*, pages 4–13. PMLR, 21 Jun 2010.
- [36] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *EEE Transactions on Knowledges and Data Engineering*, 17(4):491–502, 2005.

- [37] A. Lorena, L. Garcia, J. Lehmann, M. Souto, and T. K. Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys*, 52(5):1–34, 2019.
- [38] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- [39] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The Adaptive Computations and machine Learning. MIT Press, 2nd edition, 2018.
- [40] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Trans. Comput.*, 26(9):917–922, sep 1977.
- [41] V. Pestov. Is the k-nn classifier in high dimensions affected by the curse of dimensionality? *Computers and mathematics with Applications*, 65:1427–1437, 2013.
- [42] C. Phillips, J. Amigo, A.O. Tillmar, M.A. Pecke, M. de la Puente, J. Ruiz-Ramrez, F. Bittner, Š. Idrizbegović, Y. Wang, T.J. Parsons, and M.V. Lareu. A compilation of tri-allelic snps from 1000 genomes and use of the most polymorphic loci for a large-scale human identification panel. *Forensic Science International: Genetics*, 46(102232), 2020.
- [43] D. Pollard. Asymptopia. <http://www.stat.yale.edu/~pollard/Courses/607.spring05/handouts/Totalvar.pdf>, 2005. Manuscript given as class notes.
- [44] W. Rudin. *Real and Complex Analysis*. McGraw-Hill series in higher mathematics. McGraw-Hill, 1970.
- [45] J. Saint Raymond. *Topologie, calcul différentiel et variable complexe*. Mathématiques en devenir. Calvage & Mounet, 2nd edition, 2008.
- [46] P.D. Sasieni. From genotypes to genes: Doubling the sample size. *Biometrics*, 53(4):1253–1261, 1997.

- [47] B. Schölkopf and A. Smola. *Learning with Kernels*. Adaptive Computation and Machine Learning. MIT Press, 1st edition, 2002.
- [48] P. Somol, P. Pudil, and J. Kittler. Fast branch & bound algorithms for optimal feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):900–912, 2004.
- [49] S.S. Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory Probab. Appl.*, 18:784–786, 1974.
- [50] V. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer New York, 2nd edition, 2000.
- [51] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, 1974.
- [52] V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.
- [53] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.
- [54] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.
- [55] B. Yu and B. Yuan. A more efficient branch and bound algorithm for feature selection. *Pattern Recognition*, 26(6):883 – 889, 1993.
- [56] G. Zheng, J. Joo, and Y. Yang. Pearson’s test, trend test, and max are all trend tests with different types of scores. *Ann Hum Genet*, 73:133–140, 2009.