Distances and metrics on probability measures

Bachelor thesis

written by

Alexander Stannat

at the

Universität
Konstanz

Chair of mathematics and statistics

supervised by

Prof. Dr. Michael Kupper

Konstanz
16.03.2017

Contents

1 Introduction

In mathematics the concept of distance has a very important role to play. From analysis to algebra the notion of difference between two mathematical entities can not be gone without. In order to quantify distance, mathematics provides us with the useful definition of a metric. A metric is a mapping, taking two elements of a set and assigning them to their mutual distance from each other. More specifically, a metric is a symmetric and positive semidefinite mapping from the product space of the mathematical entities in question to $\mathbb{R}_{\geq 0}$, satisfying the triangle inequality. A particular case of a metric is a norm, assigning an object to its distance from zero. In probability theory, a statistical distance quantifies the distance between two statistical objects, such as random variables, probability distributions, etc. Statistical distances endow the space of the statistical objects in question with a topology, giving way to concepts such as openness and closedness, which entail functionalanalytical principles such as convergence and density, etc.

In this transcript we will introduce the idea of probability distances, also referred to as probability metrics, if the aforementioned conditions of a metric are complied with. Note, that in this transcript we will use the terminologies of distance and metric equivalently. Given a measurable space and the set of probability measures on its σ-algebra, these metrics assign two probability measures to their mutual distance from eachother, creating some notion of distance on the space of probability measures as well as endowing it with a topology. The main value in these metrics lies in the structure that they create on the space of probability measures. In many cases the topology induced, equates to the topology of weak convergence or to an even greater one and additionally some of the structure of the underlying measurable space is transferred to the space of probability measures. There is a wide variety of probability metrics to examine and we will introduce the reader to the most important ones.

The reason behind the multitude of probability distances is best explained as follows. Imagine wanting to find out the distance between Berlin and Shanghai. This is an easy task. All we need to do is find a path connecting the two, the length of which yields one version of distance. The point is that in a case like this it has been conventionalised to take the length of the shortest path connecting the two points, which doesn't leave much room for alternative interpretations of distance. Now imagine being asked for the distance between several cities in Asia and Europe, or for that matter for the distance between Europe and Asia in general. Suddenly the concept of distance is no longer a clearly defined, tangible idea. There are an infinte number of ways of defining such distance. In this case, a useful approach, would be to interpret the spatial distributions of the towns in Europe and Asia (or Europe and Asia themselves) by two probability measures. Now we are asking for the distance between the probability measures, i.e. a probability distance.

This idea is closely related to the concept of the optimal transport problem, which we will begin with, in this transcript. After an extensive introduction to the Monge-Kantorovich problem, we move on to introducing the Wasserstein distance and the Total Variation distance as a special case as well as an upper bound of the Wasserstein distance. After a comprehensive insight into these probability metrics, we introduce the Prokhorov metric and as a special case the Lévy metric. The respective applications of these probability distances in terms of the topology that they induce on the space of probability measures as well as their relation to the topology of weak convergence, will constitute our main focus among a few bounding properties. Finally we finish by providing the reader with a superficial outlook on the most common metrics and their respective topologies.

2 Optimal transport

The field of optimal transport is a fairly old one and the idea of transporting something from one place to another, while minimising the cost is still present to this day. It dates back to the 18^{th} century, when the the French mathematician Gaspard Monge published one of his famous works entitled *Mémoire sur la théorie des déblais et des remblais* [13], which was based on the hypothetical problem of having to transport a certain amount of soil extracted from a pit to a construction site. In this case a *déblai* is an amount of mass, extracted from the ground, while a *remblai* is mass integrated in some construction site. We will broaden this concept further in the following chapter.

2.1 The Monge problem

The concept of optimal transport is best explained in the context of Monge's optimal transport problem, mentioned above. We will rephrase this problem with an analogy, involving a consortium of cafés and bakeries. Note, that this transcript is loosely based on Cedric Villani's script entitled *Optimal Transport, Old and New* [15].

Let's suppose we are given a consortium of n bakeries and m cafés. Let's also assume for the sake of the argument that the set of all bakeries $\mathcal{X}$ and the set of cafés $\mathcal{Y}$ form two disjoint subsets of the euclidean plane $\mathbb{R}^2$. One can visualise this, by picturing a map from above with points scattered across it, marking the individual bakeries and cafés. The bakeries, of course produce bread and deliver it to the cafés, that go on to sell it. Since the delivery of bread isn't free, there exists a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, where $c(x, y)$ denotes the price of delivering one unit of bread from bakery x to café y.

The cost of transportation obviously depends on the spatial distributions of the bakeries and the cafés, i.e. a function that tells us how much bread is produced by bakeries in some area of our "map" and another function to indicate how much bread is consumed by cafés in another area. This is best done by a measure, but before detailing what a measure is, we must first introduce the concept of a σ-algebra.

2.1.1 Definition

Given a set $\mathcal{X}$ we call a set of subsets $\mathcal{F}$ of $\mathcal{X}$ a σ-algebra, if it satisfies the following conditions

$$\begin{aligned} &(i)\ \mathcal{X} \in \mathcal{F} \\ &(ii)\ A \in \mathcal{F} \implies A^c \in \mathcal{F} \\ &(iii)\ A_1, A_2, A_3, \ldots \in \mathcal{F} \implies \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F} \end{aligned}$$

In that case we call $(\mathcal{X}, \mathcal{F})$ a measurable space. If $(\mathcal{X}, d)$ is a metric space, one usually restricts one's attention to the Borel σ-algebra. The Borel σ-algebra is defined over the topology, induced by d, whereby we take the smallest σ-algebra that contains all open sets. With this in mind we know that any metric space is already a measurable space. In our case we mostly focus on Borel σ-algebras on metric spaces.

With respect to the aforementioned spatial distributions of the cafés and bakeries, we introduce the idea of a measure.

2.1.2 Definition

Given a measurable space $(\mathcal{X}, \mathcal{F})$, a σ-additive function $\mu : \mathcal{F} \to [0, \infty]$ is called a measure, if it satisfies $\mu(\varnothing) = 0$.
A function μ is called σ-additive, if for a sequence of disjoint sets $A_1, A_2, A_3, \ldots \in \mathcal{F}$, it holds

$$\mu(\dot{\bigcup_{n\in\mathbb{N}}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

A measure μ is called a probability measure, if additionally it satisfies $\mu(\mathcal{X}) = 1$. We then call $(\mathcal{X}, \mathcal{F}, \mu)$ a probability space. The set of all probability measures on some measurable space $\mathcal{X}$ is symbolised by $\mathcal{P}(\mathcal{X})$.
Oftentimes we will simply write $(\mathcal{X}, \mu)$ instead of $(\mathcal{X}, \mathcal{F}, \mu)$, if $\mathcal{F}$ can be derived from the context.

Applying these definitions to the preceding "café bakery" analogy, the original Monge transport problem takes on the following form.

We're given two probability spaces $(\mathcal{X}, \mathcal{F}_\mathcal{X}, \mu)$ and $(\mathcal{Y}, \mathcal{F}_\mathcal{Y}, \nu)$, whereby $\mathcal{X}$ denotes the set of bakeries and $\mathcal{Y}$ the set of cafés. μ and ν are the respective spatial distributions of the bakeries and cafés within $\mathcal{X}$ and $\mathcal{Y}$. That means for $A \in \mathcal{F}_\mathcal{X}$, $\mu(A)$ signifies the proportion of bread that is supplied by bakeries in A. Analogously, for $B \in \mathcal{Y}$, $\nu(B)$ denotes the proportion of bread consumed by cafés in B. It is important to note here that μ and ν don't quantify the amount of bread produced/consumed itself, since probability measures are bounded by 1. Instead we define them to denote the proportion of bread, whereby the proportion of bread is simply given by the amount of units of bread produced in A or consumed in B divided by the amount of bread produced/consumed in general. For the notion of cost to remain viable we need to change the cost function to denote the cost of delivery of one unit mulitplied by the consumption/production rate. Given the fact that this has virtually no effect on the setting of the problem itself it will be silently implied and no longer mentioned.

Take a transport function $T : \mathcal{X} \to \mathcal{Y}$, with $T(x)$ being the café that bakery x delivers its bread to. Obviously T then needs to be measurable and the supply and demand of the individual cafés and bakeries need to be satisfied. For T, this implies that for any $A \in \mathcal{F}_\mathcal{X}$, the cafés in $T(A)$ consume just as much bread as bakeries in A produce. This means $\nu(T(A)) = \mu(A)$ or more generally we require

$$\mu \circ T^{-1} = \nu.$$

Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, Monge's optimal transport problem lies in finding a measurable transport function T, that minimises the total cost of transportation, given by

$$\int_\mathcal{X} c(x, T(x)) d\mu(x).$$

The problem we face now, is that for T to be well-defined, one bakery can only deliver bread to one café. It is not allowed for a bakery to split its production "mass", i.e to deliver to several cafés simultaneously. This makes finding a transport function between the two given probability spaces not always feasible.

Let's say, for instance μ is a dirac measure and ν is not, then the requirement

$$\mu \circ T^{-1} = \nu$$

can not be satisfied for any well-defined T. This can simply be interpreted as there being only one bakery and several cafés in need of supply.

We need a more general approach to the problem. With this in mind, we move on from finding a transport function T, to trying to find a coupling between the two probability measures μ and ν.

2.2 Couplings

Coupling is a very useful proof technique, that allows us to compare two unrelated random distributions.

2.2.1 Definition

Let μ and ν be two probability measures on some measurable spaces $(\mathcal{X}, \mathcal{F}_1)$ and $(\mathcal{Y}, \mathcal{F}_2)$. Coupling μ and ν means contructing a probability space $(\Omega, \mathbb{P})$ and a tuple of random variables $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$, such that

$$\begin{aligned} \text{law}(X) &:= \mathbb{P} \circ X^{-1} = \mu \\ \text{law}(Y) &:= \mathbb{P} \circ Y^{-1} = \nu. \end{aligned}$$

We call the tuple (X, Y) a coupling of (μ, ν). By abuse of language we call the law of (X, Y) a coupling of (μ, ν) as well.

It is clear, that for any two probability measures μ and ν, there always exists a coupling, since one can simply set $\Omega := \mathcal{X} \times \mathcal{Y}$ and $\mathbb{P} := \mu \otimes \nu$. The coupling of (μ, ν) given by the tuple of random variables $(\text{proj}_{\mathcal{X}}, \text{proj}_{\mathcal{Y}})$, then simply equates to the product measure, which obviously always exists. For a coupling π of (μ, ν) it holds

$$\pi(A \times \mathcal{Y}) = \mu(A) \quad \text{and} \quad \pi(\mathcal{X} \times B) = \nu(B).$$

This is known as the marginal condition and can be rephrased in the following two equivalent ways:

(*i*) For all integrable functions $\psi : \mathcal{X} \to \mathbb{R}$ and $\phi : \mathcal{Y} \to \mathbb{R}$ it holds

$$\int\limits_{\mathcal{X} \times \mathcal{Y}} \psi(x) + \phi(y) \, d\pi(x, y) = \int\limits_{\mathcal{X}} \psi(x) \, d\mu(x) + \int\limits_{\mathcal{Y}} \phi(y) \, d\nu(y).$$

(*ii*) $\pi \circ \text{proj}_{\mathcal{X}}^{-1} = \mu \quad \text{and} \quad \pi \circ \text{proj}_{\mathcal{Y}}^{-1} = \nu.$

2.2.2 Definition

A deterministic coupling of two probability measures (μ, ν) is given by a probability space $(\Omega, \mathbb{P})$ and a tuple of random variables (X, Y) as well as a measurable function $T : \mathcal{X} \to \mathcal{Y}$. In that case $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ is a called a deterministic coupling of (μ, ν), if one of the following equivalent conditions is satisfied:

(*i*)

$$\begin{gathered} \mathbb{P} \circ X^{-1} = \mu, \, \mathbb{P} \circ Y^{-1} = \nu \;\text{ and }\; T \circ X = Y, \\ \text{then }\; \mu \circ T^{-1} = \mathbb{P} \circ X^{-1} \circ T^{-1} = \mathbb{P} \circ (T \circ X)^{-1} = \mathbb{P} \circ Y^{-1} = \nu. \end{gathered}$$

(*ii*)

$$\mathbb{P} \circ (X, Y)^{-1} \text{ is concentrated on the graph of T } \; \{(x, T(x)); x \in \mathcal{X}\}.$$

(iii)

$$\text{For all integrable functions } \phi : \mathcal{Y} \to \mathbb{R} \text{ it holds}$$
$$\int\limits_{\mathcal{Y}} \phi(y) d\nu(y) = \int\limits_{\mathcal{X}} \phi(T(x)) d\mu(x).$$

(iv)

$$(\mu \otimes \nu) = \mu \circ (\mathrm{Id}, T)^{-1}.$$

An interesting example is coupling a dirac measure with another probability measure.

2.2.3 Theorem

Given two probability spaces $(\mathcal{X}, \delta_{x_0})$ and $(\mathcal{Y}, \nu)$ for an arbitrary $x_0 \in \mathcal{X}$, then the product measure $\delta_{x_0} \otimes \nu$ is the only coupling of the two measures.

Proof. We will show this by taking a coupling (X, Y) with a probability space $(\Omega, \mathbb{P})$ such that $\pi := \mathbb{P} \circ (X, Y)^{-1}$ has marginals δ_{x_0} and ν.

Assume now $\pi \neq \delta_{x_0} \otimes \nu$, then there exists $A \times B \in \mathcal{X} \times \mathcal{Y}$ such that $\pi(A \times B) \neq \delta_{x_0}(A)\nu(B)$. This leaves us with two possible cases

Case 1: $(x_0 \notin A)$

$\pi(A \times B) \leq \pi(A \times \mathcal{Y}) = \delta_{x_0}(A) = 0$, which implies $\pi(A \times B) = 0 = \delta_{x_0}(A)\nu(B)$.

Case 2: $(x_0 \in A)$

$\pi(A \times B) = \pi(\mathcal{X} \times B) - \pi(A^c \times B) = \nu(B) - 0 = \nu(B) = \delta_{x_0}(A)\nu(B)$.

Hence, we know that any coupling π of (δ_{x_0}, ν) is already equal to the product measure. □

2.3 The Monge-Kantorovich problem

In the original Monge problem we had the objective of trying to find a deterministic coupling to minimise the cost of transportation. Since this wasn't always practicable, we now change the approach to finding a coupling of the two probability measures μ and ν. In this case couplings describe transportation plans, (i.e $\pi(A, B)$ denotes the proportion of bread delivered from bakeries in a measurable set $A \subset \mathcal{X}$ to cafés in a measurable set $B \subset \mathcal{Y}$). It is now our goal to minimise the transportation cost over all possible transportation plans. This brings us to the Monge-Kantorovich problem. Recall the definition of a Polish space.

2.3.1 Definition

A topological space $\mathcal{X}$ is called complete if every Cauchy sequence converges in $\mathcal{X}$. It is called separable, if there exists a countable and dense subset. A metric space $(\mathcal{X}, d)$ is called a Polish space, if it is complete and separable.

Take two Polish spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$, whereby μ and ν are probability measures on the Borel σ-algebras, induced by the respective metrics on $\mathcal{X}$ and $\mathcal{Y}$, describing the spatial distributions of the cafés and bakeries. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a cost function, then the price of a transport plan, described by a coupling π of μ and ν, is given by

$$C := \int\limits_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y).$$

It is now our goal to find an optimal coupling π of μ and ν that minimises the transportation costs. We call couplings π of (μ, ν) transference plans, whereby the marginal condition of π can be interpreted as the fact that a workable transference plan saturates the supply of the bakeries and the demand of the cafés. Let $\Pi(\mu, \nu)$ denote the set of all couplings of (μ, ν), then it is our goal to find

$$\inf_{\pi \in \Pi(\mu,\nu)} \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y). \tag{1}$$

(1) takes on another form, given by the Kantorovich duality, that will prove itself useful in the future. This brings us to a few lemmas and theorems, which we will not prove here for the most part. The proofs to these can be found in [15].

2.3.2 Definition

Let $(\mathcal{X}, d)$ be a metric space and let $\mathcal{P}(\mathcal{X})$ be the set of all probability measures on the Borel σ-algebra of $\mathcal{X}$, induced by the metric d. We then define the topology of weak convergence as the coarsest topology on $\mathcal{P}(\mathcal{X})$ such that for all open $G \subset \mathcal{X}$, the mapping $\mathcal{P}(\mathcal{X}) \to [0,1], \mu \mapsto \mu(G)$ is lower semicontinuous.

As the name would suggest, this topology admits a convergence, namely weak convergence. We call a sequence, that converges in this topology weakly convergent, denoted $\mu_n \xrightarrow{\omega} \mu$.

2.3.3 Theorem (Portmanteau)

Given a sequence $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ and $\mu \in \mathcal{P}(\mathcal{X})$, then the following are equivalent

$$\begin{aligned}
&(i) && \mu_n \xrightarrow{\omega} \mu \\
&(ii) && \text{For every open } G \subset \mathcal{X} \text{ we have } \liminf_{n\to\infty} \mu_n(G) \geq \mu(G) \\
&(iii) && \text{For every } f \in C_b(\mathcal{X}) \text{ we have } \lim_{n\to\infty} \int\limits_{\mathcal{X}} f(x)d\mu_n(x) = \int\limits_{\mathcal{X}} f(x)d\mu(x).
\end{aligned}$$

2.3.4 Lemma

Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces then the set of all couplings $\Pi(\mu, \nu)$ of (μ, ν) is closed in the topology of weak convergence.

2.4 Existence of an optimal coupling

The given problem consists in minimising the cost of transportation. The first question that arises, pertaining to the cost of the optimal transport plan is, is there actually a transference plan that minimises the cost. It turns out that the the infimum over all couplings in (1) is attained and therefore a minimum.

2.4.1 Theorem

Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces, let

$$a : \mathcal{X} \to \mathbb{R} \cup \{-\infty\} \quad \text{and} \quad b : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$$

be two upper semicontinuous functions such that $a \in L^1(\mu), b \in L^1(\nu)$.

Given a lower semicontinuous cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{\infty\}$ with $c(x, y) \geq a(x) + b(y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, there exists a coupling $\tilde{\pi}$ of (μ, ν), such that

$$\inf_{\pi \in \Pi(\mu,\nu)} \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) = \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\tilde{\pi}(x, y).$$

It goes without saying, that this is a very useful conclusion to come to. Especially, since the conditions, on which the infimum is attained are farily mild. For instance, if one chooses a metric as the cost function, we know there exists a minimising transport plan. From this we move on to the Kantorovich duality.

3 The Kantorovich duality

Suppose now we find ourselves in the following situation:
We have come up with a transportation plan, delivering bread from bakeries in $\mathcal{X}$ to cafés in $\mathcal{Y}$. Wanting to reduce the cost of delivery, we take a bakery x_1 and reroute one unit of bread, that was originally sent to some café y_1, to a closer café y_2. We then gain $c(x_1, y_2) - c(x_1, y_1)$. Since there is now an excess of bread delivered to y_2, we reroute one unit of the bread, sent from some bakery x_2 to café y_2 to another café y_3. This process is repeated until we decide to reroute one unit of the bread sent from some bakery x_N to y_N, back to y_1, at which point the cycle is complete.
This new plan is cheaper than the old one, if and only if it holds

$$c(x_1, y_2) + c(x_2, y_3) + \ldots + c(x_N, y_1) < c(x_1, y_1) + c(x_2, y_2) + \ldots + c(x_N, y_N).$$

If one can find a cycle of rerouting such that the upper inequality is satisfied, we know that our original transference plan can't be optimal, which brings us to the definition of cyclical monotonicity.

3.1 Cyclical monotonicity

3.1.1 Definition

Let $\mathcal{X}, \mathcal{Y}$ be some sets and let $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, \infty]$ be a cost function. We call a subset $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ cyclically montone if, for any $N \in \mathbb{N}$ and any family $(x_1, y_1), \ldots, (x_N, y_N) \in \Gamma$ it holds

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}) \quad \text{(whereby } y_{N+1} = y_1).$$

We call a transference plan π between the two sets $\mathcal{X}$ and $\mathcal{Y}$ cyclically montone, if it is concentrated on some cyclically monotone subset of $\mathcal{X} \times \mathcal{Y}$.

Intuitively this means, that a transportation plan is cyclically monotone, if it can not be improved by rerouting mass as described above. It is obvious that an optimal transference plan is cyclically monotone, due to its optimality. What's interesting about this definition, is that any cyclically monotone transference plan is already optimal, which is a consequence of theorem 3.4.1.

3.2 The dual Kantorovich problem

Let's suppose now, that instead of doing the transportation ourselves, a logistics company steps in, buying bread from the bakeries and selling it to the respective cafés. Let $\psi(x)$ be the price that the company pays for one unit of bread at bakery x and let $\phi(y)$ be the price at which it sells a unit of bread to café y. Instead of the original cost $c(x, y)$ the consortium now pays

$\phi(y) - \psi(x)$ per unit of bread delivered from x to y. For the company to remain in business, it needs to set up prices such that

$$\phi(y) - \psi(x) \leq c(x,y), \qquad (x,y) \in \mathcal{X} \times \mathcal{Y},$$

otherwise the consortium could just make the delivery cheaper themselves. We call such a pair of prices competitive. In this case it is the goal of the company to maximise their revenue, given by

$$\int\limits_{\mathcal{Y}} \phi(y) d\nu(y) - \int\limits_{\mathcal{X}} \psi(x) d\mu(x)$$

over all pairs of competitive prices. At this point we impose that the prices ϕ and ψ are integrable. Later on this will follow from a bounding property of c and the fact that the prices are competitive.

Given the set $\Pi(\mu,\nu)$ of all probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. With the help of the company, the transportation of each unit of bread is not more expensive than, when the consortium was doing it themselves. We therefore obtain

$$\sup_{\phi - \psi \leq c} \left\{ \int\limits_{\mathcal{Y}} \phi(y) d\nu(y) - \int\limits_{\mathcal{X}} \psi(x) d\mu(x) \right\} \leq \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) \right\}.$$

It is, of course the company's goal to set the highest selling price ϕ and lowest buying price ψ possible, while remaining competitive.
Consider an arbitrary pair of competitive prices (ϕ, ψ). Then, of course we can obtain a higher selling price $\phi_1(y) = \inf_{x\in\mathcal{X}} (\psi(x) + c(x,y))$ while remaining competitive. ψ can be improved by replacing it with $\psi_1(x) = \sup_{y\in\mathcal{Y}} (\phi_1(y) - c(x,y))$. Now take $\phi_2(y) = \inf_{x\in\mathcal{X}} (\psi_1(x) + c(x,y))$ and $\psi_2(x) = \sup_{y\in\mathcal{Y}} (\phi_1(y) - c(x,y))$. In every step of this sequence we obtain better buying and selling prices for the company, while remaining competitive. We prove that this process is stationary.

We know that $\phi_1 \geq \phi$ and it is

$$\begin{aligned} \phi_2(y) &= \inf_{x\in\mathcal{X}} (\psi_1(x) + c(x,y)) = \inf_{x\in\mathcal{X}} \left(\sup_{z\in\mathcal{Y}} (\phi_1(z) - c(x,z)) + c(x,y) \right) \\ &\geq \inf_{x\in\mathcal{X}} (\phi_1(y) - c(x,y) + c(x,y)) \ = \phi_1(y). \end{aligned}$$

For ψ_2 we get

$$\begin{aligned} \psi_2(x) &= \sup_{y\in\mathcal{Y}} (\phi_2(y) - c(x,y)) \geq \sup_{y\in\mathcal{Y}} (\phi_1(y) - c(x,y)) = \psi_1(x), \\ \psi_2(x) &= \sup_{y\in\mathcal{Y}} (\phi_2(y) - c(x,y)) = \sup_{y\in\mathcal{Y}} \left(\inf_{z\in\mathcal{X}} (\psi_1(z) + c(z,y)) - c(x,y) \right) \\ &\leq \sup_{y\in\mathcal{Y}} (\psi_1(x) + c(x,y) - c(x,y)) = \psi_1(x). \end{aligned}$$

For ψ_1 it is

$$\begin{aligned} \psi_1(x) &= \sup_{y\in\mathcal{Y}} (\phi_1(y) - c(x,y)) \geq \sup_{y\in\mathcal{Y}} (\phi(y) - c(x,y)) = \psi(x), \\ \psi_1(x) &= \sup_{y\in\mathcal{Y}} (\phi_1(y) - c(x,y)) = \sup_{y\in\mathcal{Y}} \left(\inf_{z\in\mathcal{X}} (\psi(z) + c(z,y)) - c(x,y) \right) \\ &\leq \sup_{y\in\mathcal{Y}} (\psi(x) + c(x,y) - c(x,y)) = \psi(x). \end{aligned}$$

And for ϕ_2

$$\phi_2(y) = \inf_{x\in\mathcal{X}} (\psi_1(x) + c(x,y)) = \inf_{x\in\mathcal{X}} (\psi(x) + c(x,y)) = \phi_1(y).$$

We therefore obtain $\phi_1 = \phi_2$ and $\psi_1 = \psi_2$. This shows that, given an arbitrary pair of competitive prices one iteration of the process above will deliver a pair of prices that can not be improved without the company losing its competitivity. We call such a pair of prices (ϕ, ψ) tight. They then obviously satisfy.

$$\phi(y) = \inf_{x\in\mathcal{X}} (\psi(x) + c(x,y)), \qquad \psi(x) = \sup_{y\in\mathcal{Y}} (\phi(y) - c(x,y))$$

This means, when looking at competitive prices we can restrict our attention to ψ and reconstruct ϕ. Although taking an arbitrary buying price ψ and reconstructing an optimal selling price ϕ doesn't guarantee that the pair is in fact tight. This brings us to the definition of c-convexity.

3.3 C-convexity

3.3.1 Definition

Given two sets $\mathcal{X}, \mathcal{Y}$ and a function $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, \infty]$. Then a function $\psi : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ is called c-convex if it's not identically $+\infty$ and there exists ϕ such that

$$\forall x \in \mathcal{X} \quad \psi(x) = \sup_{y\in\mathcal{Y}} (\phi(y) - c(x,y)).$$

Then its c-transform ψ^c is defined by

$$\psi^c(y) = \inf_{x\in\mathcal{X}} (\psi(x) + c(x,y)), \quad \text{f.a. } y \in \mathcal{Y}.$$

Since c-convexity depends on the cost function c, there are particular cases, in which a c-convex function in regard to a certain cost function c has nice properties.
For instance, given a metric space $(\mathcal{X}, d)$ and the cost function $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ then for a c-convex function on $\mathcal{X}$ it obviously needs to hold

$$\forall x, y \in \mathcal{X} \quad \psi^c(y) - \psi(x) \leq d(x,y).$$

And therefore

$$\psi^c = \psi.$$

Then we know $\psi(x) - \psi(y) \leq d(x,y)$, for all $x, y \in \mathcal{X}$, i.e. ψ is Lipschitz continuous with Lipschitz constant 1.

Conversely, a 1-Lipschitz continuous function is c-convex with c-transform $\psi = \psi^c$, since

$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y} \quad \psi(y) - \psi(x) \leq d(x,y). \quad \text{This means } \psi(y) = \inf_{x\in\mathcal{X}} (\psi(x) + d(x,y)).$$

This now brings us to the Kantorovich duality, which will then lead us to the introduction of the Wasserstein metric.

3.4 The Kantorovich duality

3.4.1 Theorem

Let $(\mathcal{X},\mu)$ and $(\mathcal{Y},\nu)$ be two Polish probability spaces. Let $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ be a lower semicontinuous cost function such that there exist some upper semicontinuous, real-valued functions $a \in L^1(\mu)$ and $b \in L^1(\nu)$ satisfying

$$\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad c(x,y) \geq a(x) + b(y).$$

Then the following equality holds true

$$\begin{aligned}\min_{\pi\in\Pi(\mu,\nu)} \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) &= \sup_{\substack{(\psi,\phi)\in C_b(\mathcal{X})\times C_b(\mathcal{Y})\\ \phi-\psi\leq c}} \left(\int\limits_{\mathcal{Y}} \phi(y)d\nu(y) - \int\limits_{\mathcal{X}} \psi(x)d\mu(x)\right)\\ &= \sup_{\substack{(\psi,\phi)\in L^1(\mu)\times L^1(\nu)\\ \phi-\psi\leq c}} \left(\int\limits_{\mathcal{Y}} \phi(y)d\nu(y) - \int\limits_{\mathcal{X}} \psi(x)d\mu(x)\right)\\ &= \sup_{\psi\in L^1(\mu)} \left(\int\limits_{\mathcal{Y}} \psi^c(y)d\nu(y) - \int\limits_{\mathcal{X}} \psi(x)d\mu(x)\right).\end{aligned}$$

Here we can simply impose that ψ is c-convex.

If the transport cost associated with the optimal coupling π is finite, then there is a measurable cyclically monotone set $\Gamma \subset \mathcal{X}\times\mathcal{Y}$ such that for $\pi \in \Pi(\mu,\nu)$ the following are equivalent:

(i) π is optimal
(ii) π is cyclically monotone
(iii) There is a c-convex function ψ such that $\psi^c(y) - \psi(x) = \pi(x,y)$, $\pi-$almost surely
(iv) π is concentrated on Γ.

This is what was stated in section 3.1 (cyclical monotinicity).

If, in addition to the cost of the minimising transference plan being finite, there exist $c_{\mathcal{X}} \in L^1(\mu)$ and $c_{\mathcal{Y}} \in L^1(\nu)$, such that $c(x,y) \leq c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y)$, then the supremum on the right-hand side is attained and therefore a maximum. The minimum being attained on the left-hand side of the equation is a direct consequence of Theorem 2.4.1. This now brings us to the main focus of this bachelor thesis, the Wasserstein distance.

4 The Wasserstein distance

In the analogy above, we were given two probability measures μ,ν and an arbitrary lower semicontinuous cost function c and inspected the optimal transport cost, given by

$$C = \inf_{\pi\in\Pi(\mu,\nu)} \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y).$$

We now look at the same problem, but change the given cost function c to d^p, whereby d is a metric, inducing the σ-algebra on which the given probability measures are defined. This then introduces some notion of distance between the two probability measures, called the Wasserstein distance. Note here that d of course satisfies the conditions of theorem 2.4.1, giving us a minimum on the left-hand side of the equation. It also satisfies the conditions of the dual Kantorovich theorem 3.4.1. We will use this later on.

4.0.1 Definition

Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1, \infty)$. For two probability measures μ and ν on the Borel σ-algebra on $\mathcal{X}$, induced by the given metric d, the Wasserstein distance of order p between μ and ν is defined by

$$W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int\limits_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}}$$

$$= \inf \left\{ \left(\mathbb{E}\, d(X, Y)^p \right)^{\frac{1}{p}};\ \text{law}(X) = \mu, \quad \text{law}(Y) = \nu \right\}.$$

This can now be interpreted as the minimum cost of transportation between the bakeries and the cafés, whereby the cost function is simply given by the distance between the respective production and consumption units. The Wasserstein distance is oftentimes also referred to as the earth mover's distance, in respect to the Monge problem. We briefly introduce a proposition to show what the Wasserstein distance may look like.

4.0.2 Proposition

Take two arbitrary points $x, y \in \mathcal{X}$ and $a, b \in [0, 1]$, let

$$\mu = a\delta_x + (1 - a)\delta_y, \quad v = b\delta_x + (1 - b)\delta_y$$

be two probability measures on some measurable space $\mathcal{X}$, where the σ-algebra contains all sets with only one element. Then it holds $W_p(\mu, \nu) = |a - b|^{\frac{1}{p}} d(x, y)$.

Proof. Take a coupling π of (μ, ν), we know that π is concentrated on $\{(x, x), (x, y), (y, x), (y, y)\} \subset \mathcal{X} \times \mathcal{X}$. This yields

$$\int\limits_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi(x, y) = (\pi(x, y) + \pi(y, x))\, d(x, y)^p.$$

Due to the marginal property of π we obtain

$$\pi(x, x) + \pi(y, x) = \nu(x) = b$$
$$\pi(x, x) + \pi(x, y) = \mu(x) = a.$$

For $a \leq b$ we know that $\pi(x, y) \leq \pi(y, x)$ and therefore the infimum is attained by setting $\pi(x, y) = 0$, $\pi(x, x) = a$, $\pi(y, x) = b - a$, which implies $W_p(\mu, \nu)^p = (b - a)d(x, y)^p$.

The proof for $b \leq a$ is analogous with $\pi(y, x) = 0$, $\pi(x, x) = b$ and $\pi(x, y) = a - b$. □

4.1 Axiomatic properties of the Wasserstein distance

While W_p isn't necessarily finite, due to the fact, that $(\mathcal{X}, d)$ doesn't need to be bounded, we can show that the Wasserstein distance satisfies all other axioms of a metric on $\mathcal{P}(\mathcal{X})$. We introduce a Lemma without proof for this.

4.1.1 Gluing Lemma

Let $(\mathcal{X}_i, \mu_i), i = 1, 2, 3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3), then one can construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (Y_2, Y_3). It is easy to see why it's called gluing Lemma. If the law of (X_1, X_2) is given by π_{12} and the law of (X_2, X_3) by π_{23} then we construct π_{123} by simply "gluing" together π_{12} and π_{23} along their common marginal μ_2.

4.1.2 Theorem

The Wasserstein distance W_p satisfies all axioms of a metric, apart from finiteness.

Proof. We know that given a measurable space $(\Omega, \mathbb{P})$, if π is a coupling of a tuple of probability measures (μ, ν), i.e $\pi = \mathbb{P} \circ (X,Y)^{-1}$ for some random variables X, Y then $\tilde{\pi} = \mathbb{P} \circ (Y,X)^{-1}$ is a coupling of (ν, μ), which yields

$$W_p(\mu,\nu)^p = \inf_{\pi \in \Pi(\mu,\nu)} \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y) = \inf_{\tilde{\pi} \in \Pi(\mu,\nu)} \int\limits_{\mathcal{X}\times\mathcal{X}} d(y,x)^p d\tilde{\pi}(y,x) = W_p(\nu,\mu)^p.$$

Since $d \geq 0$ is a metric we already know that $W_p \geq 0$.

For (μ, ν) with $\mu = \nu$, we know that $\mu \circ (\mathrm{Id}, \mathrm{Id})^{-1}$ is a coupling of (μ, ν) with

$$\mu \circ (\mathrm{Id}, \mathrm{Id})^{-1}(A \times B) = \mu(A \cap B).$$

From this we conclude that the given coupling is concentrated on the diagonal and since the Wasserstein distance is based on the metric d on $\mathcal{X}$ we obtain $W_p(\mu, \nu) = 0$.

$W_p(\mu,\nu) = 0$ implies there exists some coupling of (μ, ν) given by $\mathbb{P} \circ (X,Y)^{-1}$, that's concentrated on the diagonal. We conclude $X = Y\,\mathbb{P}$-as. For any measurable $A \subset \mathcal{X}$ it then holds

$$\mu(A) = \mathbb{P} \circ (X,Y)^{-1}(A \times \mathcal{X}) = \mathbb{P} \circ (X,X)^{-1}(A \times \mathcal{X}) = \mathbb{P} \circ (X,X)^{-1}(\mathcal{X} \times A) = \nu(A),$$

which implies $\mu = \nu$.

For the triangle inequality we use the gluing Lemma.
Let μ_1, μ_2 and μ_3 be three probability measures on $\mathcal{X}$ and let (X_1, X_2) be an optimal coupling of (μ_1, μ_2) and (Y_2, Y_3) an optimal coupling of (μ_2, μ_3). Then there exist random variables (Z_1, Z_2, Z_3) with $\mathrm{law}(Z_1, Z_2) = \mathrm{law}(X_1, X_2)$ and $\mathrm{law}(Z_2, Z_3) = \mathrm{law}(Y_2, Y_3)$. In particular (Z_1, Z_3) is a coupling of (μ_1, μ_3) and we obtain

$$\begin{aligned} W_p(\mu_1, \mu_3) &\leq (\mathbb{E}\, d(Z_1, Z_3)^p)^{\frac{1}{p}} \\ &\leq (\mathbb{E}(d(Z_1, Z_2) + d(Z_3, Z_3))^p)^{\frac{1}{p}} \\ &\leq (\mathbb{E}\, d(Z_1, Z_2)^p)^{\frac{1}{p}} + (\mathbb{E}\, d(Z_2, Z_3)^p)^{\frac{1}{p}} \\ &= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3). \end{aligned}$$

In this case the third inequality is based on the Minkowski inequality in $L^p(\mathbb{P})$. □

4.2 The Wasserstein space

We now want W_p additionally to the upper axioms to also be finite, in order for it to be useful. We reduce the space of all probability measures $\mathcal{P}(X)$ to a subspace on which W_p is finite.

4.2.1 Definition

Given an arbitrary point $x_0 \in \mathcal{X}$ we define

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

It is obvious that for d bounded, it holds $P_p^{x_0}(\mathcal{X}) = \mathcal{P}(\mathcal{X})$.

We show that one can choose $x_0 \in \mathcal{X}$ randomly, without changing $P_p^{x_0}(\mathcal{X})$.
Given $x_0, y_0 \in \mathcal{X}$ arbitrary, it holds for $\mu \in P_p^{x_0}(\mathcal{X})$

$$\int\limits_{\mathcal{X}} d(x_0, x)^p d\mu(x) < \infty.$$

With the triangle inequality for d and Jensen's inequality we obtain

$$\begin{aligned}\int\limits_{\mathcal{X}} d(y_0, x)^p d\mu(x) &\leq \int\limits_{\mathcal{X}} (d(x_0, y_0) + d(x_0, x))^p d\mu(x) \\ &\leq \int\limits_{\mathcal{X}} 2^{p-1} d(x_0, y_0)^p d\mu(x) + \int\limits_{\mathcal{X}} 2^{p-1} d(x_0, x)^p d\mu(x) \\ &< \infty,\end{aligned}$$

which means $\mu \in P_p^{y_0}(\mathcal{X})$. The inversed implication follows analogously. Therefore $P_p^{x_0}(\mathcal{X}) = P_p(\mathcal{X})$ is independant of the choice of x_0.

4.2.2 Theorem

The Wasserstein distance W_p is finite on the Wasserstein space $P_p(\mathcal{X})$ and therefore a metric.

Proof. We take two probability measures μ and ν in $P_p(\mathcal{X})$ and a coupling π of the two. Once again, the triangle inequality for d and Jensen's inequality yield

$$\begin{aligned}W_p(\mu, \nu) &\leq \left(\int\limits_{\mathcal{X}\times\mathcal{X}} d(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}} \leq \left(\int\limits_{\mathcal{X}\times\mathcal{X}} (d(x, x_0) + d(x_0, y))^p d\pi(x, y) \right)^{\frac{1}{p}} \\ &\leq \left(\int\limits_{\mathcal{X}\times\mathcal{X}} 2^{p-1} \left(d(x, x_0)^p + d(x_0, y)^p \right) d\pi(x, y) \right)^{\frac{1}{p}} \\ &= \left(2^{p-1} \int\limits_{\mathcal{X}} d(x, x_0)^p d\mu(x) + 2^{p-1} \int\limits_{\mathcal{X}} d(x_0, y)^p d\nu(y) \right)^{\frac{1}{p}} < \infty.\end{aligned}$$

□

Here is a useful consequence of the Kantorovich duality:
Seeing as a c-convex function in regard to the cost function $c = d$ is 1-Lipschitz and the fact that d satisfies all conditions of the Kantorovich duality, we obtain

$$W_1(\mu, \nu) = \sup \left\{ \int\limits_{\mathcal{X}} \psi(x) d\mu(x) - \int\limits_{\mathcal{X}} \psi(y) d\nu(y)\, ;\;\; \psi \text{ is 1-Lipschitz} \right\}.$$

This is also known as the Kantorovich-Rubinstein formula.

For $p > 1$ it holds

$$W_p(\mu, \nu)^p = \sup \left\{ \int\limits_{\mathcal{X}} \psi^c(y) d\nu(y) - \int\limits_{\mathcal{X}} \psi(x) d\mu(x)\, ;\;\; \psi \in C_b(\mathcal{X}) \right\},$$

with $\psi^c(y) = \inf_{x \in \mathcal{X}} (\psi(x) + d(x,y)^p)$.

The Wasserstein distance is a very popular distance on the space of probability measures and exhibits many nice properties.

(*i*) It is $W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} ||d||_{L^p(\pi)}$ and therefore Hölder's inequality implies that for $p \leq q$ it holds

$$W_p \leq W_q.$$

(*ii*) The Wasserstein distance is defined by an infimum, which makes it fairly easy to bound from above. Looking at the Kantorovich-Rubinstein formula, we see it is defined as a supremum aswell, which makes it nice to bound from below. The construction of any coupling will bound it from above. Finding a pair of tight price-functions ϕ and ϕ^c will bound it from below.

(*iii*) The mapping $x \mapsto \delta_x$ defines an isometry between $\mathcal{X}$ and $P_p(\mathcal{X})$, since any coupling of two dirac measures was already given by the tensor product of the two. Therefore it holds for $x_0, y_0 \in \mathcal{X}$

$$W_p(\delta_{x_0}, \delta_{y_0})^p = \inf_{\pi \in \Pi(\delta_{x_0}, \delta_{y_0})} \int_{\mathcal{X} \times \mathcal{X}} d(x,y)^p d\pi(x,y) = \int_{\mathcal{X}} \int_{\mathcal{X}} d(x,y)^p d\delta_{x_0} d\delta_{y_0} = d(x_0, y_0)^p.$$

(*iv*) For a C-Lipschitz $f : \mathcal{X} \to \mathcal{X}'$ the mapping $P_p(\mathcal{X}) \to P_p(\mathcal{X}'), \mu \to \mu \circ f^{-1}$ is also C-Lipschitz, since

$$\begin{aligned} W_p(\mu \circ f^{-1}, \nu \circ f^{-1})^p &= \inf_{\pi \in \Pi(\mu \circ f^{-1}, \nu \circ f^{-1})} \int_{\mathcal{X}' \times \mathcal{X}'} d(x,y)^p d\pi(x,y) \\ &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d(f(x), f(y))^p d\pi(x,y) \\ &\leq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} C d(x,y)^p d\pi(x,y) \\ &= C \cdot W_p(\mu,\nu)^p. \end{aligned}$$

(*v*) Given a probability space $(\Omega, \mathbb{P})$ and two random variables $X, Y : \Omega \to \mathbb{R}$ with $X, Y \in L^p(\mathbb{R})$ it is

$$W_p(\mathbb{P} \circ X^{-1}, \mathbb{P} \circ Y^{-1}) \leq \|X - Y\|_p,$$

since (X, Y) is a coupling of $\mathbb{P} \circ X^{-1}$ and $\mathbb{P} \circ Y^{-1}$. Therefore the mapping $L^p(\mathbb{R}) \to P_p(\mathbb{R})$, that projects a random variable onto its law is a contraction. Recall theorem 2.2.3, stating that a coupling of a dirac measure and any other probability measure, is given by the product measure, then for a random variable X on the probability space $(\Omega, \mathbb{P})$, it holds

$$W_p(\mathbb{P} \circ X^{-1}, \delta_0) = \|X\|_p.$$

(*vi*) There exists a continuous bijection between the Wasserstein space $L^p(\mathbb{R})$ and $P_p(\mathbb{R})$. This follows from the fact, that for any probability measure there exists a random variable, with the same law.

5 Topological properties of the Wasserstein distance

We have come up with the Wasserstein space $P_p(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$, which we now want to endow with a topology, allowing the concept of openness and closedness on $P_p(\mathcal{X})$.

5.1 Weak convergence on the Wasserstein space

Since the Wasserstein space was a subspace of the space of all probability measures, the topology describing convergence needs to be at least as great as the topology of weak convergence, while additionally retaining the finiteness of the integral, so that the limit remains in $P_p(\mathcal{X})$. We define convergence on the Wasserstein space as follows.

5.1.1 Definition

Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1, \infty)$. Let $(\mu_n)_{n\in\mathbb{N}} \subset P_p(\mathcal{X})$ be a sequence of probability measures and $\mu \in P_p(\mathcal{X})$. We then say that $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ in $P_p(\mathcal{X})$, if it converges weakly in terms of regular weak convergence and for any arbitrary $x_0 \in \mathcal{X}$ any of the following equivalent properties is satisfied

$$(i)\quad \lim_{k\to\infty} \int\limits_{\mathcal{X}} d(x_0, x)^p d\mu_k(x) = \int\limits_{\mathcal{X}} d(x_0, x)^p d\mu(x)$$

$$(ii)\quad \limsup_{k\to\infty} \int\limits_{\mathcal{X}} d(x_0, x)^p d\mu_k(x) \leq \int\limits_{\mathcal{X}} d(x_0, x)^p d\mu(x)$$

$$(iii)\quad \lim_{R\to\infty} \limsup_{k\to\infty} \int\limits_{d(x_0,x)\geq R} d(x_0, x)^p d\mu_k(x) = 0.$$

This clearly shows that, if d isn't bounded, weak convergence on $P_p(\mathcal{X})$ is stronger than regular weak convergence, due to the additional conditions, posted above.

For instance, if d isn't bounded, one can consider the sequence of probability measures on $P_p(\mathcal{X})$ given by a sequence and a point $(x_n)_{n\in\mathbb{N}}, x_0$ in $\mathcal{X}$, with $r_n = d(x_0, x_n) \to \infty$ and

$$\mu_n = (1 - r_n^{-p})\, \delta_{x_0} + r_n^{-p}\, \delta_{x_n}.$$

This sequence obviously converges weakly with weak limit $\mu = \delta_{x_0}$, but it also holds

$$W_p(\mu_n, \mu) = |1 - r_n^{-p} \; - 1|^{\frac{1}{p}}\, d(x_n, x_0) = 1$$

for all $n \in \mathbb{N}$ as was shown in proposition 4.0.2. Therefore a weakly convergent sequence in $\mathcal{P}(\mathcal{X})$ isn't necessarily weakly convergent in $P_p(\mathcal{X})$.

For d bounded we know that weak convergence on $P_p(\mathcal{X})$ is equivalent to regular weak convergence, because condition (iii) of 5.1 is automatically satisfied for any weakly convergent sequence in $\mathcal{P}(\mathcal{X})$. We then of course also know that $P_p(\mathcal{X}) = \mathcal{P}(\mathcal{X})$.

Before we go on to proving that the Wasserstein distance metrizes weak convergence on $P_p(\mathcal{X})$ we provide a few lemmas.

5.1.2 Lemma

Let $(\mathcal{X}, d)$ be a Polish space, $p \in [1, \infty)$. Then every Cauchy sequence in $(P_p(\mathcal{X}), W_p)$ is tight.

Proof. Take a Cauchy sequence $(\mu_n)_{n\in\mathbb{N}} \subset P_p(\mathcal{X})$, then

$$W_p(\mu_k, \mu_l) \to 0 \qquad\qquad (l, k \to \infty).$$

We know that for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $k \geq N$ it holds

$$W_p(\mu_N, \mu_k) \leq \varepsilon^2.$$

And given $W_1 \leq W_p$ for $p \geq 1$, we obtain $W_1(\mu_N, \mu_k) \leq \varepsilon^2$ for all $k \geq N$.

Now it is obvious that for every $k \in \mathbb{N}$, there exists $j \in \{1, \ldots, N\}$ such that $W_1(\mu_k, \mu_j) \leq \varepsilon^2$ (For $k \leq N$ choose $j = k$ and for $k \geq N$ choose $j = N$).

Since we were given a Polish space we know that every probability measure on $\mathcal{X}$ is tight and therefore every finite set of probability measures is tight, because the finite union of compact sets is compact. Hence we can find a compact $K \subset \mathcal{X}$ such that

$$\sup_{k\leq N} \mu_k(K^c) \leq \varepsilon.$$

By compactness, K can be covered in finitely many balls of radius ε,

$$K \subset U := B(x_1, \varepsilon) \cup \ldots \cup B(x_m, \varepsilon) \quad \text{with } x_1, \ldots, x_m \in \mathcal{X}.$$

This means for all $k \leq N$ it is $\mu_k(U^c) \leq \varepsilon$. We now use the fact that for all $k \in \mathbb{N}$ it holds $W_1(\mu_N, \mu_k) \leq \varepsilon^2$, to show that there exists a measurable set $U_\varepsilon \supset U$ with $\sup\limits_{k\in\mathbb{N}} \mu({U_\varepsilon}^c) \leq 2\varepsilon$. Define

$$U_\varepsilon := \Big\{x \in \mathcal{X}; d(x, U) < \varepsilon\Big\} \subset B(x_1, 2\varepsilon) \cup \ldots \cup B(x_m, 2\varepsilon)$$

and

$$\phi_\varepsilon(x) := \left(1 - \frac{d(x, U)}{\varepsilon}\right)_+.$$

Obviously it is $1_U \leq \phi_\varepsilon \leq 1_{U_\varepsilon}$. We show that ϕ_ε is Lipschitz continuous with Lipschitz constant $\frac{1}{\varepsilon}$. For all $x, y \in \mathcal{X}$ with $\phi_\varepsilon(x) \geq \phi_\varepsilon(y)$ it holds

$$\begin{aligned}
|\phi_\varepsilon(x) - \phi_\varepsilon(y)| &= \left(1 - \frac{d(x, U)}{\varepsilon}\right)_+ - \left(1 - \frac{d(y, U)}{\varepsilon}\right)_+ \\
&\leq \left(\frac{d(y, U) - d(x, U)}{\varepsilon}\right)_+ \\
&\leq \left(\frac{d(x, y)}{\varepsilon}\right)_+ \\
&= \frac{d(x, y)}{\varepsilon}.
\end{aligned}$$

For $x, y \in \mathcal{X}$ with $\phi_\varepsilon(y) \geq \phi_\varepsilon(x)$ the same inequality follows analogously.
Note that the Kantorovich duality was true for any 1-Lipschitz function ϕ. We know that ϕ_ε is Lipschitz-continuous with constant $\frac{1}{\varepsilon}$ and therefore $\varepsilon\phi_\varepsilon$ is 1-Lipschitz and we can use theorem 3.4.1. We also know that given an arbitrary $k \in \mathbb{N}$ there exists $j \in \{1, \ldots, N\}$ such that $W_1(\mu_k, \mu_j) < \varepsilon^2$ and $\mu_j(U) \geq 1 - \varepsilon$.

This yields

$$\begin{aligned}\mu_k(U_\varepsilon) &= \int\limits_{\mathcal{X}} 1_{U_\varepsilon} d\mu_k \\ &\geq \int\limits_{\mathcal{X}} \phi_\varepsilon(x) d\mu_k \\ &\geq \int\limits_{\mathcal{X}} \phi_\varepsilon(x) d\mu_j - \left(\int\limits_{\mathcal{X}} \phi_\varepsilon(x) d\mu_j - \int\limits_{\mathcal{X}} \phi_\varepsilon(y) d\mu_k \right) \\ &\geq \int\limits_{\mathcal{X}} \phi_\varepsilon(x) d\mu_j - \frac{W_1(\mu_k, \mu_j)}{\varepsilon} \\ &\geq \mu_j(U) - \frac{W_1(\mu_k, \mu_j)}{\varepsilon}.\end{aligned}$$

At this point we know that for every $\varepsilon > 0$ we can find a finite union of balls $U_\varepsilon \subset \mathcal{X}$ such that for all $k \in \mathbb{N}$ it is

$$\mu_k(U_\varepsilon) \geq 1 - \varepsilon - \frac{\varepsilon^2}{\varepsilon} = 1 - 2\varepsilon.$$

We now face the problem that the set U_ε isn't necessarily compact for every $\varepsilon > 0$. Recall that $(\mathcal{X}, d)$ is a Polish space. This means that for every $k \in \mathbb{N}$ there exists a finite family $\{x_1, \ldots, x_{m(l)}\} \subset \mathcal{X}$ such that it holds

$$\mu_k \left(\mathcal{X} \Big\backslash \bigcup_{1 \leq i \leq m(l)} B(x_i, 2^{-l+1}\varepsilon) \right) \leq 2^{-l}\varepsilon.$$

Now define $S := \bigcap\limits_{1 \leq l < \infty} \bigcup\limits_{1 \leq i \leq m(l)} \overline{B(x_i, 2^{-l+1}\varepsilon)}$ then for $k \in \mathbb{N}$ arbitrary it is

$$\begin{aligned}\mu_k(\mathcal{X} \backslash S) &= \mu_k \left(\bigcup_{l \in \mathbb{N}} \left(\bigcup_{i=1}^{m(l)} \overline{B(x_i, 2^{-l+1}\varepsilon)} \right)^c \right) \\ &\leq \sum_{l \in \mathbb{N}} \mu_k \left(\left(\bigcup_{i=1}^{m(l)} \overline{B(x_i, 2^{-l+1}\varepsilon)} \right)^c \right) \\ &\leq \sum_{l \in \mathbb{N}} 2^{-l}\varepsilon \\ &= \varepsilon.\end{aligned}$$

Furthermore it is clear that S is totally bounded and closed. Note that this implies compactness of S. To recap, we now know that for every ε there exists a compact $K \subset \mathcal{X}$ such that

$$\sup_{k \in \mathbb{N}} \mu_k(K^c) \leq \varepsilon$$

and therefore $(\mu_n)_{n \in \mathbb{N}}$ is tight. □

5.1.3 Lemma

For every $\varepsilon \geq 0$ there exists a constant C_ε such that for all $a, b \in \mathbb{R}_+$ it holds

$$(a + b)^p \leq (1 + \varepsilon)a^p + C_\varepsilon b^p.$$

The proof to this requires a differentiation of cases.

Proof. By dividing by a^p we see that the upper ineqaulity is equivalent to

$$(1+x)^p \leq 1 + \varepsilon + C_\varepsilon x^p, \text{ where } x = \frac{b}{a}.$$

Case 1: $x \leq x_\varepsilon := (1+\varepsilon)^{\frac{1}{p}}$

$$(1+x)^p \leq x^p + C_\varepsilon x^p \leq 1 + \varepsilon + C_\varepsilon x^p.$$

Case 2: $x > x_\varepsilon := (1+\varepsilon)^{\frac{1}{p}}$
Choosing $C_\varepsilon := (1 + x_\varepsilon^{-1})^p$ yields

$$1 + \varepsilon + C_\varepsilon x^p = 1 + \varepsilon + (1 + x_\varepsilon^{-1})^p x^p \geq 1 + \varepsilon + (1+x)^p \geq (1+x)^p.$$

□

5.2 Prokhorov's theorem

Let $(\mathcal{X}, d)$ be a separable metric space and $\mathcal{P}(\mathcal{X})$ denote the space of probability measures on the Borel σ-algebra $\mathcal{B}(\mathcal{X})$ induced by d.

(i) A subset $K \subset \mathcal{P}(\mathcal{X})$ is tight if and only if $\overline{K}$ is sequentially compact in the space equipped with the topology of weak convergence, i.e every infinite sequence $(\mu_k)_{k \in \mathcal{X}} \subset K$ has a subsequence that converges weakly.

(ii) The space $\mathcal{P}(\mathcal{X})$ endowed with the topology of weak convergence is metrizable.

(iii) If $(\mathcal{X}, d)$ is a Polish space, there is a complete topology on $\mathcal{P}(\mathcal{X})$ equivalent to the topology of weak convergence.

5.2.1 Lemma

Let $(\mathcal{X}, d)$ and $(\mathcal{Y}, d')$ be two Polish spaces. Given two tight sets of probability measures $\mathcal{P} \subset \mathcal{P}(\mathcal{X})$ and $\mathcal{Q} \subset \mathcal{P}(\mathcal{Y})$, let $\Pi(\mathcal{P}, \mathcal{Q})$ be the set of all couplings $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ between probability measures in $\mathcal{P}$ and $\mathcal{Q}$. Then $\Pi(\mathcal{P}, \mathcal{Q})$ is tight as well.

Proof. Take an arbitrary $\varepsilon > 0$. Knowing that $\mathcal{P}$ and $\mathcal{Q}$ are tight, we know that there are two compact sets $K \subset \mathcal{X}$ and $\tilde{K} \subset \mathcal{Y}$ such that

$$\sup_{\mu \in \mathcal{P}} \mu(K^c) < \varepsilon \qquad \text{and} \qquad \sup_{\nu \in \mathcal{Q}} \nu(\tilde{K}^c) < \varepsilon.$$

Note that for $K \subset \mathcal{X}$ and $\tilde{K} \subset \mathcal{Y}$ compact, $K \times \tilde{K}$ is compact in $\mathcal{X} \times \mathcal{Y}$.
From this we conclude

$$\begin{aligned} &\sup_{\pi \in \Pi(\mathcal{P}, \mathcal{Q})} \pi\left((K \times \tilde{K})^c\right) \leq \sup_{\pi \in \Pi(\mathcal{P}, \mathcal{Q})} \pi(K^c \times \mathcal{Y}) + \sup_{\pi \in \Pi(\mathcal{P}, \mathcal{Q})} \pi(\mathcal{X} \times \tilde{K}^c) \\ &= \sup_{\mu \in \mathcal{P}} \mu(K^c) + \sup_{\nu \in \mathcal{Q}} \nu(\tilde{K}^c) \\ &< 2\varepsilon. \end{aligned}$$

Therefore $\Pi(\mathcal{P}, \mathcal{Q})$ is tight. □

5.2.2 Lemma

Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a lower semicontinuous cost function and $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ an upper semicontinuous function with $h \leq c$.

Given a sequence of probability measures $(\pi_k)_{k\in\mathbb{N}}$ on $\mathcal{X} \times \mathcal{Y}$, that converges weakly to some $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, in such a way that $h \in L^1(\pi_k)$ for all $k \in \mathbb{N}$ and $h \in L^1(\pi)$.

If, additionally it holds

$$\lim_{k\to\infty} \int\limits_{\mathcal{X}\times\mathcal{Y}} h(x,y)d\pi_k \quad = \quad \int\limits_{\mathcal{X}\times\mathcal{Y}} h(x,y)d\pi,$$

then

$$\int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi \quad \leq \quad \liminf_{k\to\infty} \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi_k.$$

5.2.3 Lemma

Let $(\mathcal{X}, \mathcal{T})$ be a sequentially compact topological space and $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ a sequence, such that every convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$ has the same limit $x \in \mathcal{X}$.
Then $(x_n)_{n\in\mathbb{N}}$ converges to x.

Proof. We presume $(x_n)_{n\in\mathbb{N}}$ doesn't converge to x, then there exists a neighbourhood $A \in \mathcal{T}$ of x such that for every $n_0 \in \mathbb{N}$ we can find $n \in \mathbb{N}$ with $n \geq n_0$ and $x_n \notin A$.

We define a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ such that for all $k \in \mathbb{N}$, $x_{n_k} \notin A$, which due to the sequential compactness has a convergent subsequence $(x_{n_{k'}})_{k'\in\mathbb{N}}$ with limit x. This, of course is a contradiction to our assumption: $x_{n_k} \notin A$ for all $k \in \mathbb{N}$.

Consequently we conclude that no such sequence can exist, which means $x_n \to x$. □

5.3 Metrization of weak convergence on the Wasserstein space

We recall that we had defined a convergence on the Wasserstein space. This means we have endowed $P_p(\mathcal{X})$ with a topology. Since W_p is a metric on $P_p(\mathcal{X})$ it also induces a topology on the Wasserstein space. It turns out, that these two are identical.

5.3.1 Theorem

Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1, \infty)$. The Wasserstein metric induces the topology of weak convergence on the Wasserstein space $P_p(\mathcal{X})$. This means that for $(\mu_k)_{k\in\mathbb{N}} \subset P_p(\mathcal{X})$ and $\mu \in P_p(\mathcal{X})$ the two following statements are equivalent:

$$\begin{aligned} &(i) \quad \mu_k \text{ converges weakly in } P_p(\mathcal{X}) \text{ to } \mu \\ &(ii) \quad W_p(\mu_k, \mu) \to 0. \end{aligned}$$

Proof. We start off by showing, that (ii) implies (i).
Given $(\mu_k)_{k\in\mathbb{N}} \subset P_p(\mathcal{X})$ and $\mu \in P_p(\mathcal{X})$ with $W_p(\mu_k, \mu) \to 0$, we show that $(\mu_k)_{k\in\mathbb{N}}$ converges to μ in terms of regular weak convergence and additionally satisfies condition (ii) of definition 5.1.1.

Due to convergence we know that $(\mu_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $(P_p(\mathcal{X}), W_p)$ and by Lemma 5.1.2, tight. By Prokhorov's theorem we know that $\{\mu_k | k \in \mathbb{N}\}$ is relatively sequentially compact. This means there exists a weakly convergent subsequence $(\mu_{k'})_{k'\in\mathbb{N}}$ of $(\mu_k)_{k\in\mathbb{N}}$ with some weak limit $\tilde{\mu} \in \mathcal{P}(\mathcal{X})$. Of course, it still holds $W_p(\mu_{k'}, \mu) \to 0$.

Since $(\mathcal{X}, d)$ is a Polish space we know that every probability measure is tight, which means $\{\tilde{\mu}\}$ is tight. The sequence $(\mu_{k'})_{k' \in \mathbb{N}}$ is also tight, as a subsequence of a tight sequence.
We know by Lemma 5.2.1 that the set $\Pi(\{\mu_{k'} | k' \in \mathbb{N}\}, \{\tilde{\mu}\})$ of all couplings between measures $\mu_{k'}$ and $\tilde{\mu}$ is tight as well, which, again by Prokhorov's theorem means that it is relatively sequentially compact.

Now, let $(\pi_{k'})_{k' \in \mathbb{N}}$ denote the sequence of optimal couplings of $(\mu_{k'}, \mu)$. We know that there exists a convergent subsequence of $(\pi_{k'})_{k' \in \mathbb{N}}$, which we will simply denote $(\pi_{k'})_{k' \in \mathbb{N}}$ again. Let $\tilde{\pi} \in P(\mathcal{X} \times \mathcal{X})$ be the weak limit of said subsequence, then we know that $\tilde{\pi}$ is a coupling of $(\tilde{\mu}, \mu)$, because for every $f \in C_b(\mathcal{X})$ it holds

$$\int\limits_{\mathcal{X} \times \mathcal{X}} f(x) d\tilde{\pi}(x,y) = \lim_{k' \to \infty} \int\limits_{\mathcal{X} \times \mathcal{X}} f(x) d\tilde{\pi}_{k'}(x,y) = \lim_{k' \to \infty} \int\limits_{\mathcal{X}} f(x) d\tilde{\mu}_{k'}(x) = \int\limits_{\mathcal{X}} f(x) d\tilde{\mu}(x)$$

and of course

$$\int\limits_{\mathcal{X} \times \mathcal{X}} f(y) d\tilde{\pi}_{k'}(x,y) = \lim_{k' \to \infty} \int\limits_{\mathcal{X} \times \mathcal{X}} f(y) d\tilde{\pi}_{k'}(x,y) = \int\limits_{\mathcal{X}} f(y) d\mu(y).$$

Hence $\tilde{\pi}$ satisfies the marginal conditions and therefore is a coupling of $(\tilde{\mu}, \mu)$.
This yields

$$W_p(\tilde{\mu}, \mu)^p = \inf_{\pi \in \Pi(\tilde{\mu}, \mu)} \int\limits_{\mathcal{X} \times \mathcal{X}} d(x,y)^p d\pi(x,y) \leq \int\limits_{\mathcal{X} \times \mathcal{X}} d(x,y)^p d\tilde{\pi}(x,y).$$

Knowing that $(\pi_{k'})_{k' \in \mathbb{N}}$ converges weakly to $\tilde{\pi}$ we can use Lemma 5.2.2, setting $c(x,y) = d^p(x,y)$, which is obviously continuous as a composition of continuous functions, to obtain

$$\int\limits_{\mathcal{X} \times \mathcal{X}} d(x,y)^p d\tilde{\pi}(x,y) \leq \liminf_{k' \to \infty} \int\limits_{\mathcal{X} \times \mathcal{X}} d(x,y)^p d\pi_{k'}(x,y) = \liminf_{k' \to \infty} W_p(\mu_{k'}, \mu)^p = 0.$$

This implies $W_p(\tilde{\mu}, \mu) = 0$, so $\tilde{\mu} = \mu$. We now know that $\{\mu_k \mid k \in \mathbb{N}\}$ is a relatively sequentially compact sequence whereby every convergent subsequence has the same limit. Then the closure is sequentially compact an by Lemma 5.2.3, we conclude that $(\mu_k)_{k \in \mathbb{N}}$ converges weakly to μ.

At this point we have only shown that $W_p(\mu_k, \mu) \to 0$ implies $\mu_k \to \mu$ in terms of regular weak convergence in $\mathcal{P}(\mathcal{X})$. We now need to prove that $(\mu_k)_{k \in \mathbb{N}}$ converges to μ in $P_p(\mathcal{X})$, by showing that for an arbitrary $x_0 \in \mathcal{X}$ it holds

$$\limsup_{k \to \infty} \int\limits_{\mathcal{X}} d(x, x_0)^p d\mu_k(x) \leq \int\limits_{X} d(x, x_0)^p d\mu(x).$$

Like before, let $(\pi_k)_{k \in \mathbb{N}}$ be the sequence of optimal couplings of (μ_k, μ)
We know by Lemma 5.1.3 that for every $\varepsilon \geq 0$ there exists $C_\varepsilon \geq 0$ such that for all $a, b \geq 0$ it is

$$(a+b)^p \leq (1+\varepsilon) a^p + C_\varepsilon b^p.$$

We apply this inequality to $d(x_0, x)$ for $x_0 \in \mathcal{X}$ and obtain for arbitrary $\varepsilon \geq 0$

$$d(x, x_0)^p \leq \ (d(x_0, y) + d(x, y))^p \leq (1+\varepsilon) d(x_0, y)^p + C_\varepsilon d(x, y)^p.$$

Integrating over π_k yields

$$\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,x_0)^p d\pi_k(x,y) \leq \int\limits_{\mathcal{X}\times\mathcal{X}} (1+\varepsilon)d(x_0,y)^p + C_\varepsilon d(x,y)^p d\pi_k(x,y)$$

and the marginal property of π_k yields

$$\int\limits_{\mathcal{X}} d(x,x_0)^p d\mu_k(x) \leq (1+\varepsilon)\int\limits_{\mathcal{X}} d(x_0,y)^p d\mu(y) + C_\varepsilon \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_k(x,y).$$

We know that $W_p(\mu_k,\mu)\to 0$, which means

$$\limsup_{k\to\infty} \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_k(x,y) = 0.$$

Therefore

$$\limsup_{k\to\infty} \int\limits_{\mathcal{X}} d(x,x_0)^p d\mu_k(x) \leq (1+\varepsilon)\int\limits_{\mathcal{X}} d(x_0,y)^p d\mu(y).$$

Since this equality holds true for any $\varepsilon \geq 0$ it is

$$\limsup_{k\to 0} \int\limits_{\mathcal{X}} d(x,x_0)^p d\mu_k(x) \leq \int\limits_{\mathcal{X}} d(x,x_0)^p d\mu(x).$$

Hereby we have shown property (ii) in definition 5.1.1 of weak convergence in $P_p(\mathcal{X})$. This shows that $W_p(\mu_k,\mu)\to 0$ implies $\mu_k \to \mu$ in $P_p(\mathcal{X})$. In other words the topology on $P_p(\mathcal{X})$ induced by W_p is greater than the topology of weak convergence on $P_p(\mathcal{X})$.

We now show that (i) implies (ii).
Let $(\mu_k)_{k\in\mathbb{N}} \in P_p(\mathcal{X})$ be weakly convergent in $P_p(\mathcal{X})$ with limit $\mu \in P_p(\mathcal{X})$. It immediately follows that the sequence then also converges weakly in terms of regular weak convergence. By Prokhorov's theorem, the sequence is tight and so is $\{\mu\}$, since we're on a Polish space.

Let $(\pi_k)_{k\in\mathbb{N}}$ be the sequence of optimal couplings of (μ_k,μ). Then, by Lemma 5.2.1 $(\pi_k)_{k\in\mathbb{N}}$ is tight and by Prokhorov's theorem there exists a weakly convergent subsequence $(\pi_{k'})_{k'\in\mathbb{N}}$, with some limit $\pi \in P(\mathcal{X}\times\mathcal{X})$.

We know that π is in fact a coupling of (μ,μ), because for arbitrary $f \in C_b(\mathcal{X})$, π satisfies the marginal conditions

$$\int\limits_{\mathcal{X}\times\mathcal{X}} f(x)d\pi(x,y) = \lim_{k'\to\infty} \int\limits_{\mathcal{X}\times\mathcal{X}} f(x)d\pi_{k'}(x,y) = \lim_{k'\to\infty} \int\limits_{\mathcal{X}} f(x)d\mu_{k'}(x) = \int\limits_{\mathcal{X}} f(x)d\mu(x)$$

$$\int\limits_{\mathcal{X}\times\mathcal{X}} f(y)d\pi(x,y) = \lim_{k'\to\infty} \int\limits_{\mathcal{X}\times\mathcal{X}} f(y)d\pi_{k'}(x,y) = \int\limits_{\mathcal{X}} f(y)d\mu(y).$$

Since $(\pi_k')_{k'\in\mathbb{N}}$ converges weakly to π we know by lemma 5.2.2 that

$$\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y) \leq \liminf_{k'\to\infty} \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_{k'}(x,y).$$

Therefore π is an optimal coupling of (μ,μ). If π wasn't optimal there would be another coupling $\tilde{\pi}$ with

$$\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\tilde{\pi}(x,y) \leq \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y).$$

Hence one could find another sequence $(\tilde{\pi}_{k'})_{k'\in\mathbb{N}} \subset P(\mathcal{X}\times\mathcal{X})$ such that $\tilde{\pi}_{k'} \xrightarrow{\omega} \tilde{\pi}$. But, of course by 5.2.2 it would follow

$$\begin{aligned}\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y) &\leq \liminf_{k'\to\infty} \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_{k'}(x,y)\\ &\leq \liminf_{k'\to\infty} \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\tilde{\pi}_{k'}(x,y)\\ &\leq \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\tilde{\pi}(x,y).\end{aligned}$$

This means $\tilde{\pi} = \pi$, i.e π is already an optimal coupling of (μ,μ) and since $W_p(\mu,\mu) = 0$ it is

$$\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y) = 0.$$

In other words π is concentrated on the diagonal $\{(x,x)|x\in\mathcal{X}\}$, such that $\pi = \mu\circ(\text{Id,Id})^{-1}$.

Since every weakly convergent subsequence of $(\pi_k)_{k\in\mathbb{N}}$ has the same limit π and $(\pi_k)_{k\in\mathbb{N}}\cup\{\pi\}$ is sequentially compact, by Lemma 5.2.3 we know that the entire sequence converges weakly to π.

So far we have shown there exists a sequence of optimal couplings $(\pi_k)_{k\in\mathbb{N}}$ of (μ_k,μ), that converges weakly to an optimal coupling π of (μ,μ). Using this and property *(iii)* of weak convergence on $P_p(\mathcal{X})$ we will prove $\limsup\limits_{k\to\infty} W_p(\mu_k,\mu) = 0$.

Given an arbitrary $x_0\in\mathcal{X}$ and $R>0$, then $d(x,y)\geq R$ implies either

$$d(x,x_0)\geq\frac{R}{2} \quad\text{and}\quad d(x,x_0)\geq d(y,x_0)$$

or

$$d(y,x_0)\geq\frac{R}{2} \quad\text{and}\quad d(y,x_0)\geq d(x,x_0).$$

This means $1_{\{d(x,y)\geq R\}} \leq 1_{\{d(x,x_0)\geq\frac{R}{2} \text{ and } d(x,x_0)\geq d(y,x_0)\}} + 1_{\{d(y,x_0)\geq\frac{R}{2} \text{ and } d(y,x_0)\geq d(x,x_0)\}}$ and with the help of Jensen's inequality we obtain

$$\begin{aligned}&(d(x,y)^p - R^p)_+ = (d(x,y)^p - R^p)\,1_{\{d(x,y)\geq R\}}\\ &\quad\leq d(x,y)^p\,1_{\{d(x,y)\geq R\}}\\ &\quad\leq d(x,y)^p\,1_{\{d(x,x_0)\geq\frac{R}{2} \text{ and } d(x,x_0)\geq d(y,x_0)\}} + d(x,y)^p\,1_{\{d(y,x_0)\geq\frac{R}{2} \text{ and } d(y,x_0)\geq d(x,x_0)\}}\\ &\quad\leq (d(x,x_0)+d(y,x_0))^p\,1_{\{d(x,x_0)\geq\frac{R}{2} \text{ and } d(x,x_0)\geq d(y,x_0)\}}\\ &\quad+ (d(x,x_0)+d(y,x_0))^p\,1_{\{d(y,x_0)\geq\frac{R}{2} \text{ and } d(y,x_0)\geq d(x,x_0)\}}\\ &\quad\leq 2^p d(x,x_0)^p\,1_{\{d(x,x_0)\geq\frac{R}{2} \text{ and } d(x,x_0)\geq d(y,x_0)\}} + 2^p d(y,x_0)^p\,1_{\{d(y,x_0)\geq\frac{R}{2} \text{ and } d(y,x_0)\geq d(x,x_0)\}}\\ &\quad\leq 2^p d(x,x_0)^p\,1_{\{d(x,x_0)\geq\frac{R}{2}\}} + 2^p d(y,x_0)^p\,1_{\{d(y,x_0)\geq\frac{R}{2}\}}.\end{aligned}$$

Using the sequence of optimal couplings $(\pi_k)_{k\in\mathbb{N}}$ of (μ_k, μ), it follows

$$\begin{aligned}
W_p(\mu_k,\mu)^p &= \int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_k(x,y) \\
&= \int\limits_{\mathcal{X}\times\mathcal{X}} (d(x,y)\wedge R)^p + (d(x,y)^p - R^p)_+ \, d\pi_k(x,y) \\
&\leq \int\limits_{\mathcal{X}\times\mathcal{X}} (d(x,y)\wedge R)^p \, d\pi_k(x,y) + \int\limits_{d(x,x_0)\geq\frac{R}{2}} 2^p d(x,x_0)^p d\pi_k(x,y) \\
&\quad + \int\limits_{d(y,x_0)\geq\frac{R}{2}} 2^p d(y,x_0)^p d\pi_k(x,y) \\
&= \int\limits_{\mathcal{X}\times\mathcal{X}} (d(x,y)\wedge R)^p \, d\pi_k(x,y) + \int\limits_{d(x,x_0)\geq\frac{R}{2}} 2^p d(x,x_0)^p d\mu_k(x) \\
&\quad + \int\limits_{d(y,x_0)\geq\frac{R}{2}} 2^p d(y,x_0)^p d\mu(y).
\end{aligned}$$

We know that $(\pi_k)_{k\in\mathbb{N}}$ converges weakly to a probability measure π and since $(d(x,y)\wedge R)^p$ is bounded and continuous Portmanteau's theorem yields

$$\lim_{k\to\infty} \int\limits_{\mathcal{X}\times\mathcal{X}} (d(x,y)\wedge R)^p \, d\pi_k(x,y) = \int\limits_{\mathcal{X}\times\mathcal{X}} (d(x,y)\wedge R)^p \, d\pi(x,y) = 0.$$

From this it follows

$$\begin{aligned}
\limsup_{k\to\infty} W_p(\mu_k,\mu)^p \leq \limsup_{k\to\infty} & \int\limits_{d(x,x_0)\geq\frac{R}{2}} 2^p d(x,x_0)^p d\mu_k(x) \\
& + \int\limits_{d(y,x_0)\geq\frac{R}{2}} 2^p d(x_0,y)^p d\mu(y)
\end{aligned}$$

and since $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ in $P_p(\mathcal{X})$ it holds

$$\lim_{R\to\infty} \limsup_{k\to\infty} \int\limits_{d(x_0,x)\geq\frac{R}{2}} d(x_0,x)^p d\mu_k(x) = 0$$

and of course

$$\lim_{R\to\infty} \int\limits_{d(x_0,y)\geq\frac{R}{2}} d(x_0,y)^p d\mu_k(y) = 0,$$

which yields $\limsup\limits_{k\to\infty} W_p(\mu_k,\mu) = 0$.

We have now proven that weak convergence in $P_p(\mathcal{X})$ implies convergence in the topology induced by the Wasserstein metric W_p, which gives us the desired equivalence. □

At this point we know that W_p metrizes weak convergence on $P_p(\mathcal{X})$. This means for two weakly convergent sequences $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ in $P_p(\mathcal{X})$ with respective weak limits $\mu,\nu\in P_p(\mathcal{X})$, it holds

$$W_p(\mu_k,\nu_k)\to W_p(\mu,\nu).$$

This can be proven, by simply using the same argument as in the proof above.

On the other hand, if the two sequences only converge weakly in $P_p(\mathcal{X})$ in terms of regular weak convergence, we can use Lemma 5.2.1 and obtain $W_p(\mu,\nu) \leq \liminf_{k\to\infty} W_p(\mu_k,\nu_k)$ as we did above.

5.4 Properties of the Wasserstein space

When examining the Wasserstein distance, in many cases the underlying space was Polish. Some properties of the underlying space, especially in regard to its topology are transfered onto the Wasserstein space via the metric.

5.4.1 Theorem

Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1,\infty)$. Then the Wasserstein space $(P_p(\mathcal{X}), W_p)$ is a Polish space as well.

Proof. We have already proven, that $P_p(\mathcal{X})$ is a metric space and therefore restrict our attention to the separability and completeness.

We show the separability first:
Knowing that $(\mathcal{X}, d)$ is a Polish space, we know there exists a countable and dense subset $\mathcal{D} \subset \mathcal{X}$.

Now define $\mathcal{P}$ as the set of all measures, for which there exists $N \in \mathbb{N}$,

$$x_1,\dots,x_N \in \mathcal{D} \quad \text{and} \quad a_1,\dots,a_N \in \mathbb{Q} \qquad \text{with} \qquad \sum_{i=1}^N a_i = 1,$$

that are given by $\sum_{i=1}^N a_i\delta_{x_i}$.

It is obvious that $\mathcal{P} \subset P_p(\mathcal{X})$, since for arbitrary x_0 it holds

$$\int_{\mathcal{X}} d(x_0,x)^p d\mu(x) = \sum_{i=1}^N a_i d(x_0,x_i)^p < \infty$$

and that $\mathcal{P}$ is countable. We will show that $\mathcal{P}$ is dense in $P_p(\mathcal{X})$ and conclude the separability of $P_p(\mathcal{X})$.

Take $\mu \in P_p(\mathcal{X})$, then we know that for arbitrary $\varepsilon > 0$ and $x_0 \in \mathcal{D}$, there exists a compact set $K \subset \mathcal{X}$ such that

$$\int_{\mathcal{X}\setminus K} d(x_0,x)^p d\mu(x) \leq \varepsilon^p. \tag{2}$$

This follows directly from the fact, that for $\mu \in P_p(\mathcal{X})$ it holds

$$C := \int_{\mathcal{X}} d(x_0,x)^p d\mu < \infty.$$

Consequently we know that the mapping

$$\mathcal{B}(\mathcal{X}) \longrightarrow [0,1],\ A \mapsto \frac{1}{C}\int_A d(x_0,x)^p d\mu$$

is a probability measure in $\mathcal{P}(\mathcal{X})$. Knowing that every probability measure on a Polish space is tight, yields (2). Since K is compact and $\mathcal{X}$ is separable, K can be covered by finitely many balls $B(x_k, \varepsilon)$, with $x_k \in \mathcal{D}$.

Now define $B'_k = B(x_k, \varepsilon) \backslash B'_{k-1}$, with $B_1 = B(x_1, \varepsilon)$, then all B'_ks are disjoint and still cover K.

We now define a function $f : \mathcal{X} \to \mathcal{X}$, by

$$f(B'_k \cap K) = \{x_k\}, \qquad f(\mathcal{X} \backslash K) = \{x_0\}.$$

Then for $x \in K$ it obviously holds $d(x, f(x)) \leq \varepsilon$ and therefore

$$\int\limits_{\mathcal{X}} d(x, f(x))^p d\mu(x) = \int\limits_{K} d(x, f(x))^p d\mu(x) + \int\limits_{\mathcal{X}\backslash K} d(x, x_0)^p d\mu(x) \leq \varepsilon^p \mu(K) + \varepsilon^p = 2\varepsilon^p.$$

We can see that f is a measurable function, since we're given the Borel σ-algebra. Then the tuple of random variables (Id, f) is a coupling of μ and $\mu \circ f^{-1} = \sum\limits_{i=1}^{N} \mu(B'_i)\delta_{x_i}$, since

$$\mu \circ (\mathrm{Id}, f)^{-1}(\mathcal{X} \times A) = \mu(\{x \in \mathcal{X} | (x, f(x)) \in \mathcal{X} \times A\}) = \mu(f^{-1}(A))$$

and

$$\mu \circ (\mathrm{Id}, f)^{-1}(A \times \mathcal{X}) = \mu(\{x \in \mathcal{X} | (x, f(x)) \in A \times \mathcal{X}\}) = \mu(A).$$

The Wasserstein distance between these two is bounded by

$$W_p(\mu, \mu \circ f^{-1})^p \leq \int\limits_{\mathcal{X}\times\mathcal{X}} d(x, y)^p d\mu \circ (\mathrm{Id}, f)^{-1}(x, y) = \int\limits_{\mathcal{X}} d(x, f(x))^p d\mu(x) \leq 2\varepsilon^p.$$

What we have shown now, is that for any $\varepsilon > 0$ and any $\mu \in P_p(\mathcal{X})$ one can find $a_1, \ldots, a_N \in \mathbb{R}$ and $x_1, \ldots, x_N \in \mathcal{D}$ with $\nu = \sum\limits_{i=1}^{N} a_i \delta_{x_i}$, whereby $a_i = \mu(B'_i)$, such that $W_p(\mu, \nu) \leq 2\varepsilon^p$.

Since $\mathbb{Q}$ is dense in $\mathbb{R}$ the a_i can be approximated with arbitrary precision by $b_1, \ldots, b_N \in \mathbb{Q}$. Take $N + 1$ sequences $(b_i^{(k)})_{k\in\mathbb{N}} \subset \mathbb{Q}$ $(0 \leq i \leq N)$ such that $b_i^{(k)} \nearrow a_i$. Then define

$$\nu_k := \sum_{i=0}^{N} b_i^{(k)} \delta_{x_i} \in \mathcal{P},$$

which yields for $A \subset \mathcal{X}$ closed

$$\limsup_{k\to\infty} \nu_k(A) = \limsup_{k\to\infty} \sum_{i=0}^{N} b_i^{(k)} \delta_{x_i}(A) \leq \sum_{i=0}^{N} a_i \delta_{x_i}(A) = \nu(A).$$

This means $(\nu_k)_{k\in\mathbb{N}}$ converges weakly to ν and because for arbitrary $\tilde{x} \in \mathcal{X}$ it holds

$$\limsup_{k\to\infty} \int\limits_{\mathcal{X}} d(\tilde{x}, x)^p d\nu_k(x) = \limsup_{k\to\infty} \sum_{i=0}^{N} b_i^{(k)} d(\tilde{x}, x_i)^p \leq \sum_{i=0}^{N} a_i d(\tilde{x}, x_i)^p = \int\limits_{\mathcal{X}} d(\tilde{x}, x)^p d\nu.$$

Hence we know that $(\nu_k)_{k\in\mathbb{N}}$ converges weakly to ν in $P_p(\mathcal{X})$ and therefore $W_p(\nu_k, \nu) \to 0$. What this has shown now is that any $\mu \in P_p(\mathcal{X})$ can be approximated by probability measures in $\mathcal{P}$ of the form given above, thereby proving the separability of $P_p(\mathcal{X})$.

We now move on to the completeness of $P_p(\mathcal{X})$:

Let $(\mu_k)_{k\in\mathbb{N}} \subset P_p(\mathcal{X})$ be a W_p-Cauchy sequence. Then by Lemma 5.1.2 we know that $(\mu_k)_{k\in\mathbb{N}}$ is tight. This means by Prokhorov's theorem there exists a weakly convergent subsequence $(\mu_{k'})_{k'\in\mathbb{N}}$ of $(\mu_k)_{k\in\mathbb{N}}$ with some weak limit $\mu \in \mathcal{P}(\mathcal{X})$. Given the fact that $(\mu_{k'})_{k'\in\mathbb{N}} \subset (\mu_k)_{k\in\mathbb{N}}$ is also a Cauchy sequence in terms of the Wasserstein metric, we know that for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for every $k \in \mathbb{N}_{\geq N}$ it holds $W_p(\mu_N, \mu_k) \leq \varepsilon$. This means

$$\begin{aligned}\int\limits_{\mathcal{X}} d(x_0,x)^p d\mu(x) = W_p(\delta_{x_0},\mu)^p \\ &\leq (W_p(\delta_{x_0},\mu_N) + W_p(\mu_N,\mu))^p \\ &\leq 2^p \left(W_p(\delta_{x_0},\mu_N)^p + W_p(\mu_N,\mu)^p\right) \\ &= 2^p \left(\int\limits_{\mathcal{X}} d(x_0,x)^p d\mu_N(x)\right)^p + 2^p\varepsilon^p.\end{aligned}$$

It is $\int\limits_{\mathcal{X}} d(x_0,x)^p d\mu_N(x) < \infty$ and therefore $\int\limits_{\mathcal{X}} d(x_0,x)^p d\mu(x) < \infty$, which means $\mu \in P_p(\mathcal{X})$.

The sequence $(\mu_{k'})_{k'\in\mathbb{N}}$ is weakly convergent with weak limit μ, therefore we know that for every $l' \in \mathbb{N}$ it holds $W_p(\mu,\mu_{l'}) \leq \liminf\limits_{k'\to\infty} W_p(\mu_{k'},\mu_{l'}) \leq \limsup\limits_{k'\to\infty} W_p(\mu_{k'},\mu_{l'})$. Adding another $\limsup$ yields

$$\limsup_{l'\to\infty} W_p(\mu,\mu_{l'}) \leq \limsup_{l'\to\infty}\limsup_{k'\to\infty} W_p(\mu_{k'},\mu_{l'}).$$

Since $(\mu_{k'})_{k'\in\mathbb{N}}$ is a Cauchy sequence in terms of the Wasserstein metric it is

$$\limsup_{l'\to\infty}\limsup_{k'\to\infty} W_p(\mu_{k'},\mu_{l'}) = 0.$$

And therefore

$$\limsup_{l'\to\infty} W_p(\mu,\mu_{l'}) = 0$$

which means $(\mu_{k'})_{k'\in\mathbb{N}}$ is actually weakly convergent in terms of weak convergence in $P_p(\mathcal{X})$. Thereby we have a Cauchy sequence with a convergent subsequence and hence $(\mu_k)_{k\in\mathbb{N}}$ converges weakly in $P_p(\mathcal{X})$. Subsequently every Cauchy sequence in $P_p(\mathcal{X})$ converges, i.e $P_p(\mathcal{X})$ is complete. □

5.4.2 Remark

If $(\mathcal{X}, d)$ is compact, then the Wasserstein space $(P_p(\mathcal{X}), W_p(\mathcal{X}))$ is also compact.

Proof. We know that if $(\mathcal{X}, d)$ is compact, it holds $\sup\limits_{x\in\mathcal{X}} d(x_0,x) < \infty$ for all $x_0 \in \mathcal{X}$. From this we conclude that

$$\forall x_0 \in \mathcal{X} \quad \int\limits_{\mathcal{X}} d(x,x_0)d\mu(x) < \infty.$$

This implies $P_p(\mathcal{X}) = \mathcal{P}(\mathcal{X})$, whereby we recall that $\mathcal{P}(\mathcal{X})$ denotes the set of all probability measures on $\mathcal{X}$.

For any $\mu \in \mathcal{P}(\mathcal{X})$ it obviously holds $\mu(\mathcal{X}\backslash\mathcal{X}) = \mu(\varnothing) = 0$. Since $\mathcal{X}$ is compact we know that for every $\varepsilon > 0$ there exists a compact $K \subset \mathcal{X}$ sucht that $\sup\limits_{\mu\in\mathcal{P}(\mathcal{X})} \mu(K^c) < \varepsilon$ and therefore $\mathcal{P}(\mathcal{X})$ is tight. By Prokhorov's theorem $\mathcal{P}(\mathcal{X})$ is relatively sequentially compact and since $(\mathcal{P}(\mathcal{X}), W_p(\mathcal{X}))$ is a closed metric space it is compact. □

For instance, the Wasserstein space over the euclidean space $\mathbb{R}^n$ is a Polish space, but not compact, whereas the Wasserstein space over a closed intervall $[a,b]$ is Polish and compact. It is then of course simply given by $\mathcal{P}([a,b])$.

5.4.3 Remark

$(\mathcal{X}, d)$ locally compact does not necessarily imply $(P_p(\mathcal{X}), W_p(\mathcal{X}))$ locally compact.

Proof. A good example of this is given by $\mathbb{N}_0$ endowed with the discrete metric. This is a locally compact topological space, since the sets, given by $\{n\}$ are open and compact. Because the discrete metric is bounded by 1, we know that $P_p(\mathbb{N}_0) = \mathcal{P}(\mathbb{N}_0)$. Then $(\mathcal{P}(\mathbb{N}_0), W_1)$ is not locally compact. To show this we note, that on a Hausdorff space local compactness is equivalent to every point having a precompact neighbourhood. By Prokhorov this is equivalent to every point having a tight neighbourhood. For every $\varepsilon > 0$ the open Ball $B(\delta_0, \varepsilon)$ contains the sequence $(\mu_n)_{n\in\mathbb{N}}$, given by

$$\mu_n := (1 - \tilde{\varepsilon})\delta_0 + \tilde{\varepsilon}\delta_{x_n}$$

for a $\tilde{\varepsilon} < \varepsilon$, because

$$W_1(\delta_0, \mu_n) = \int\limits_{\mathcal{X}} d(0,x) d\mu_n = \tilde{\varepsilon}\underbrace{d(0,n)}_{=1} + (1-\tilde{\varepsilon})\underbrace{d(0,0)}_{=0} < \varepsilon.$$

It is obvious that the sequence $(\mu_n)_{n\in\mathbb{N}}$ isn't tight. Hence δ_0 does not have a tight neighbourhood and $(\mathcal{P}(\mathbb{N}_0), W_1)$ isn't locally compact. □

Now let's take a look at the Wasserstein space over the euclidean space $\mathbb{R}$.

6 The Wasserstein metric on $\mathbb{R}$

We have now understood most properties of the Wasserstein distance, especially in regard to the topology it induces on the Wasserstein space over Polish spaces. Probably the most well-known Polish space is the euclidean space $\mathbb{R}^d$, endowed simply with the standard metric. We know that the Wasserstein space over $\mathbb{R}^d$ is a Polish space.

6.1 The distance between distribution functions

6.1.1 Theorem

Let $(\mathbb{R}, d)$ be the one-dimensional euclidean space, endowed with the standard metric. Given $\mu, \nu \in P_p(\mathbb{R})$ we define the distribution functions F and G of μ and ν, by $F(x) = \mu((-\infty, x])$ and $G(x) = \nu((-\infty, x])$.

Then the Wasserstein distance between μ and ν is given by

$$W_1(\mu,\nu) \overset{(i)}{=} \int\limits_{-\infty}^{\infty} |F(x) - G(x)|\, dx \overset{(ii)}{=} \int\limits_{0}^{1} |F^{-1}(t) - G^{-1}(t)|\, dt.$$

Proof. Let's start with (i).
Recall the Kantorovich-Rubinstein formula for W_1

$$W_1(\mu,\nu) = \sup_{\|\phi\|_{\mathrm{Lip}}\le 1} \left(\int\limits_{\mathbb{R}} \phi(x) d\mu(x) - \int\limits_{\mathbb{R}} \phi(y) d\nu(y) \right) = \sup_{\|\phi\|_{\mathrm{Lip}}\le 1} \int\limits_{\mathbb{R}} \phi(x) d(\mu - \nu)(x).$$

Note that a 1-Lipschitz function ϕ is differentiable almost everywhere as well as integrable. We define f_η as the density function of $\eta := \mu - \nu$. Then the antiderivative of f_η is given by $h(r) = \eta((-\infty, r])$. Partial integration and the use of the funtion f_η, then yields

$$\int\limits_{\mathbb{R}} \phi(x) d\eta(x) = \int\limits_{\mathbb{R}} \phi(r) f_\eta(r) dr = [\phi(r)h(r)]_{r=-\infty}^{\infty} - \int\limits_{-\infty}^{\infty} \phi'(r)h(r)dr.$$

For $h(x) = F(x) - G(x)$ define $g : \mathbb{R} \to [-1, 1]$ by

$$g(x) := \begin{cases} 1, & F(x) > G(x) \\ 0, & F(x) = G(x) \\ -1, & F(x) < G(x) \end{cases}.$$

Now define

$$f(x) := \int\limits_{-\infty}^{x} g(r)dr.$$

Then f is Lipschitz continuous with Lipschitz constant 1, since for $x \geq y$ it holds

$$|f(x) - f(y)| = \left| \int\limits_{-\infty}^{x} g(r)dr - \int\limits_{-\infty}^{y} g(r)dr \right| = \left| \int\limits_{y}^{x} g(r)dr \right| \leq \int\limits_{y}^{x} |g(r)|dr \leq \int\limits_{x}^{y} 1dr = |x - y|.$$

Analogously for $x \leq y$.

Using partial integration, we obtain

$$\begin{aligned} W_1(\mu, \nu) &\geq \int\limits_{-\infty}^{\infty} f(x)d\mu(x) - \int\limits_{-\infty}^{\infty} f(x)\nu(x) \\ &= \int\limits_{-\infty}^{\infty} f(x)d\eta(x) \\ &= \int\limits_{-\infty}^{\infty} f(r)f_\eta(r)dr \\ &= \int\limits_{-\infty}^{\infty} f(r)h'(r)dr \\ &= [f(r)h(r)]_{r=-\infty}^{\infty} - \int\limits_{-\infty}^{\infty} f'(r)h(r)dr. \end{aligned}$$

Recall that $h(\infty) = \mu(\mathbb{R}) - \nu(\mathbb{R}) = 0$ and $h(-\infty) = \mu(\varnothing) - \nu(\varnothing) = 0$, which gives us

$$[f(r)h(r)]_{r=-\infty}^{\infty} = h(\infty) \int\limits_{-\infty}^{\infty} g(r)dr - h(-\infty) \int\limits_{-\infty}^{-\infty} g(r)dr = h(\infty) \int\limits_{-\infty}^{\infty} g(r)dr \leq \int\limits_{-\infty}^{\infty} h(\infty)dr = 0.$$

We therefore obtain

$$W_1(\mu, \nu) \geq \int\limits_{\mathbb{R}} f'(x)h(x)dx = \int\limits_{\mathbb{R}} g(x)h(x)dx = \int\limits_{\mathbb{R}} |h(x)|dx = \int\limits_{\mathbb{R}} |F(x) - G(x)|\, dx.$$

For the inversed inequality, note that for a 1-Lipschitz continuous function ϕ it holds $|\phi'| \leq 1$ η-a.e. and we obtain

$$\begin{aligned} W_1(\mu, \nu) = \sup_{\|\phi\|_{\mathrm{Lip}} \leq 1} \int\limits_{\mathbb{R}} \phi(x)d\eta(x) = \sup_{\|\phi\|_{\mathrm{Lip}} \leq 1} \left| \int\limits_{\mathbb{R}} \phi(x)d\eta(x) \right| &\leq \sup_{\|\phi\|_{\mathrm{Lip}} \leq 1} \left| \int\limits_{\mathbb{R}} \phi'(x)h(x)dx \right| \\ \leq \sup_{\|\phi\|_{\mathrm{Lip}} \leq 1} \int\limits_{\mathbb{R}} |\phi'(x)||h(x)|dx \leq \sup_{\|\phi\|_{\mathrm{Lip}} \leq 1} \int\limits_{\mathbb{R}} |h(x)|dx &= \int\limits_{\mathbb{R}} |F(x) - G(x)|\, dx \end{aligned}$$

and we conclude

$$W_1(\mu,\nu) = \int\limits_{-\infty}^{\infty} |F(x) - G(x)| dx.$$

For (ii) we define the inverted distribution function as

$$F^{-1}(t) := \inf\{x \in \mathbb{R} \,:\, F(x) \geq t\}.$$

We also define $a(x) := \min\{F(x), G(x)\}$ and $b(x) := \max\{F(x), G(x)\}$ and gain

$$\begin{aligned}\int\limits_{\mathbb{R}} |F(x) - G(x)| dx = \int\limits_{\mathbb{R}} b(x) - a(x) dx &= \int\limits_{\mathbb{R}} \int\limits_{0}^{b(x)} 1\, dt - \int\limits_{0}^{a(x)} 1\, dt dx \\ &= \int\limits_{\mathbb{R}} \int\limits_{0}^{1} 1_{[t \leq b(x)]}(t) - 1_{[t \leq a(x)]}(t) dt dx \\ &\overset{\text{Fubini}}{=} \int\limits_{0}^{1} \int\limits_{\mathbb{R}} 1_{[t \leq b(x)]}(t) - 1_{[t \leq a(x)]}(t) dx dt.\end{aligned}$$

We know that it holds

$$\begin{aligned}\max\{F(t), G(t)\}^{-1} &= \inf\{x \in \mathbb{R} \,:\, \max\{F(x), G(x)\} \geq t\} \\ &= \max\{\inf\{x \in \mathbb{R} \,:\, F(x) \geq t\}, \inf\{x \in \mathbb{R} \,:\, G(x) \geq t\}\} \\ &= \max\{F^{-1}(t), G^{-1}\}\end{aligned}$$

and of course it is

$$\{t \leq \max\{F(x), G(x)\}\} = \{\max\{F(t), G(t)\}^{-1} \geq x\}.$$

Using the monotonicity of F and G it follows directly from this

$$\begin{aligned}\int\limits_{\mathbb{R}} |F(x) - G(x)|\, dx &= \int\limits_{0}^{1} \int\limits_{\mathbb{R}} 1_{[b^{-1}(t) \geq x]}(t) - 1_{[a^{-1}(t) \geq x]}(t) dx dt \\ &= \int\limits_{0}^{1} \int\limits_{a^{-1}(t)}^{b^{-1}(t)} 1\, dx dt \\ &= \int\limits_{\mathbb{R}} b^{-1}(t) - a^{-1}(t) dt \\ &= \int\limits_{0}^{1} |F^{-1}(t) - G^{-1}(t)| dt.\end{aligned}$$

This yields the two desired equations. □

As mentioned before this metric is also referred to as the Kantorovich-Rubinstein metric. As a corollary one should add that for $p \geq 1$ the same equation also holds

$$W_p(\mu,\nu)^p = \int\limits_{-\infty}^{\infty} |F(x) - G(x)|^p dx = \int\limits_{0}^{1} |F^{-1}(x) - G^{-1}(x)|^p dx.$$

6.2 Particular examples of the Wasserstein distance

This perfectly expemplifies how the Wasserstein distance measures the distance between probability measures. Note, that it is not only limited to measures, but also quantifies the distance between probability distributions and therefore random variables too. Before we move on to the next probability distance, we look at a few examples of the Wasserstein distance between probability distributions.

6.2.1 Example

One of the most common distributions in probability theory is the normal distribution. Given a probability measure $\mu \in \mathcal{P}(\mathbb{R})$ with distribution function $F(x) = \mu((-\infty, x])$ it might be interesting to gauge the discrepancy between F and the normal distribution. Take the density function of the standard normal distribution $\phi(x) := \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$ and the distribution function given by $\Phi(x) := \int\limits_{-\infty}^{x} \phi(t)dt$, then the space of normal probability distributions, with expected value $\kappa \in \mathbb{R}$ and standard deviation $\sigma > 0$ could be denoted

$$\mathcal{H} := \left\{ H;\ H(x) = \Phi(\frac{x-\kappa}{\sigma}), \kappa \in \mathbb{R}, \sigma > 0 \right\}.$$

The distance between the distribution function F of μ and $\mathcal{H}$ with respect to the Wasserstein distance is then given by

$$\begin{aligned} W_p(F, \mathcal{H})^p :=& \inf \left\{ W_p(F,H)^p;\ H \in \mathcal{H} \right\} \\ =& \inf_{H \in \mathcal{H}} \int\limits_{-\infty}^{\infty} |F(x) - H(x)|^p dx \\ =& \inf_{H \in \mathcal{H}} \int\limits_{0}^{1} |F^{-1}(t) - H^{-1}(t)|^p dt \\ =& \inf_{\substack{\kappa \in \mathbb{R} \\ \sigma > 0}} \int\limits_{0}^{1} |F^{-1}(t) - \sigma\Phi^{-1}(t) - \kappa|^p dt. \end{aligned}$$

This means for two normal distributions F and G with standard deviations $\sigma_1 \geq \sigma_2$ and expected values $\kappa_1 \geq \kappa_2$, the Wasserstein distance between the two is given by

$$\begin{aligned} W_1(F,G) =& \int\limits_{0}^{1} |\sigma_1\Phi^{-1}(t) + \kappa_1 - \sigma_2\Phi^{-1}(t) - \kappa_2| dt \\ =& (\sigma_1 - \sigma_2) \int\limits_{0}^{1} |\Phi^{-1}(t) + \frac{\kappa_1 - \kappa_2}{\sigma_1 - \sigma_2}| dt \\ =& (\sigma_1 - \sigma_2) \int\limits_{0}^{1} |\Phi^{-1}(t)| dt + \kappa_1 - \kappa_2. \end{aligned}$$

6.2.2 Example

Another interesting particular example of the Wasserstein distance over $\mathbb{R}$, is given by a sequence of normal distributions $(F_n)_{n \in \mathbb{N}}$ with constant expected value κ and a sequence of standard deviations $(\sigma_n)_{n\in\mathbb{N}}$ such that $\lim\limits_{n\to\infty} \sigma_n = 0$. The limit of a sequence like that would be given by the distribution function

$$H(x) := \begin{cases} 0 & x < \kappa \\ 1 & x \geq \kappa \end{cases}.$$

Without loss of generality we assume $\kappa = 0$, then the Wasserstein distance yields

$$\begin{aligned} W_1(F_n, F) &= \int\limits_{-\infty}^{\infty} |\Phi(\frac{x}{\sigma_n}) - 1_{[x\geq 0]}(x)|dx = \int\limits_{-\infty}^{0} \Phi(\frac{x}{\sigma_n})dx + \int\limits_{0}^{\infty} 1 - \Phi(\frac{x}{\sigma_n})dx \\ &= 2\int\limits_{0}^{\infty} 1 - \Phi(\frac{x}{\sigma_n})dx = 2\sigma_n \int\limits_{0}^{\infty} 1 - \Phi(x)dx \longrightarrow 0. \end{aligned}$$

Hence a sequence of probability measures related to the normal distribution converges weakly to the dirac measure, concentrated in 0, if the standard deviation converges to 0. The distance between μ and other probability distributions in terms of the Wasserstein distance is computed analogously.

6.2.3 Example

Of course, if for a sequence $(\mu_n)_{n\in\mathbb{N}}$ the upper distance $W_p(F_n, \mathcal{H})$ goes to zero, we see that the given sequence converges weakly to a normal distribution.
For instance, take a sequence of iid random variables $(X_n)_{n\in\mathbb{N}}$ on $\mathbb{R}$ with expected value κ and standard deviation σ with respect to some probability measure $\mathbb{P}$, whereby the sequence of push-forward measures $(\mu_n)_{n\in\mathbb{N}}$, given by $\mu = \mathbb{P} \circ X_n^{-1}$ is in the Wasserstein space. Then the central limit theorem states that for

$$\nu_n := \mathbb{P} \circ \left(\frac{\sum\limits_{i=1}^{n} X_i - n\kappa}{\sigma\sqrt{n}} \right)^{-1}$$

it holds

$$W_p(\nu_n, \mathcal{H}) \longrightarrow 0.$$

One can approach the law of large numbers in a similar fashion.

This concludes the segment about the Wasserstein distance. While being one of the most commonly used probability metrics, it is by far not the only one. There are a variety of other probability distances with similar topological properties. It is easy to see why the Wasserstein distance is so popular, though. The metrization of weak convergence and the transference of properties of the underlying metric space onto the Wasserstein space are two very useful aspects of this probability metric. The only downside is the restriction of $\mathcal{P}(\mathcal{X})$ to $P_p(\mathcal{X})$. It would be very useful to find a probability metric that metrizes weak convergence on the space of all probability measures and not just a subspace of it. This brings us to the next important probability metric, the Total Variation distance.

7 Total Variation

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space.

We now introduce another distance on the space of all probability measures $\mathcal{P}(\mathcal{X})$ on $\mathcal{F}$. Given two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, then we define the Total Variation distance as

$$\|\mu - \nu\|_{TV} := \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} 1_{[x \neq y]}(x,y) d\pi(x,y) \right),$$

whereby we take the infimum over all couplings π of (μ, ν). Note the similarity to the Wasserstein distance and the Kantorovich duality. In this case we're given the cost function $c = 1_{[x \neq y]}$. This is of course equal to

$$\|\mu - \nu\|_{TV} = \inf \left\{ \mathbb{P}\left(X \neq Y\right) \middle| \ (X,Y) \text{ coupling of } (\mu,\nu) \right\}.$$

Here $\mathbb{P}$ is a probability measure on some space $(\Omega, \mathbb{P})$, such that the coupling (X, Y) satisfies

$$\mathbb{P} \circ X^{-1} = \mu \qquad \& \qquad \mathbb{P} \circ Y^{-1} = \nu.$$

One should add, that the given cost function satisfies the requirements of Theorem 2.4.1 and thus the infimum is attained and therefore a minimum. We see that our cost function satisfies the conditions of theorem 3.4.1 and therefore we know it can be rewritten as

$$||\mu - \nu||_{TV} = \sup_{\substack{(\psi,\phi) \in L^1(\mu)\times L^1(\nu) \\ \psi - \phi \leq 1_{[x\neq y]}}} \int_{\mathcal{X}} \psi(x) d\mu(x) - \int_{\mathcal{X}} \phi(y) d\nu(y)$$

$$= \sup_{\psi \in L^1(\mu)} \int_{\mathcal{X}} \psi(x) - \psi^c(x) d(\mu - \nu)(x).$$

For ψ c-convex with c-transform ψ^c.
Let's take a closer look at what c-convexity pertaining to the given cost function $1_{[x \neq y]}$ means. For ψ c-convex with c-transform ψ^c it holds

$$\psi^c(x) - \psi(x) \leq 1_{[x\neq y]}(x,x) = 0 \qquad \text{f.a. } x \in \mathcal{X}$$

and therefore $\psi^c = \psi$. This means for arbitrary $x, y \in \mathcal{X}$ we obtain

$$|\psi(x) - \psi(y)| \leq 1.$$

Hence the set of c-convex functions is given by all

$$\psi : \mathcal{X} \to \mathbb{R}$$

satisfying

$$\psi(x) - \psi(0) \leq 1 \quad \text{for all } x \in \mathcal{X}.$$

Additionally we know that

$$\mu(\mathcal{X}) - \nu(\mathcal{X}) = 0$$

and conclude that for an arbitrary function ψ and $a \in \mathbb{R}$ it holds

$$\int_{\mathcal{X}} \psi(x) + a d\mu(x) - \int_{\mathcal{X}} \psi(y) + a d\nu(y) = \int_{\mathcal{X}} \psi(x) d\mu(x) - \int_{\mathcal{X}} \psi(y) d\nu(y).$$

With this in mind we can restrict our attention to the supremum over all c-convex functions with $\psi(0) = 0$. Thus we obtain

$$||\mu - \nu||_{TV} = \sup_{\|\psi\|_\infty \leq 1} \int_{\mathcal{X}} \psi(x) d(\mu - \nu)(x).$$

Recall the additional assertion of the Kantorovich duality (theorem 3.4.1), stating that if the given cost function c was bounded two functions $c_1 \in L^1(\mu)$ and $c_2 \in L^1(\nu)$, then the given supremum was attained. Since $1_{[x \neq y]}$ is bounded from above by 1, we know that the upper supremum is a maximum.

$$||\mu - \nu||_{TV} = \max_{\|\psi\|_\infty \leq 1} \int_{\mathcal{X}} \psi(x) d(\mu - \nu)(x)$$

Before we acquaint ourselves with another not quite so trivial identity of the Total Variation distance, we introduce the Hahn-Jordan decomposition of signed measures.

7.1 Hahn and Jordan decomposition of signed measures

Recall that a measure was a nonnegative, real-valued and σ-additive function defined on a measurable space $(\mathcal{X}, \mathcal{F})$. We broaden this concept to include so called signed measures.

A signed measure is a σ-additive function $\mu : \mathcal{F} \to \overline{\mathbb{R}}$ that satisfies

$$\mu(\varnothing) = 0.$$

We additionally impose that μ assumes at most one of the values $-\infty$ and ∞. This is to prevent the case where $\infty - \infty$ occurs. And given a sequence of disjoint measurable sets $(A_n)_{n \in \mathbb{N}}$ for which it holds

$$\mu\left(\dot{\bigcup_{n \in \mathbb{N}}} A_n\right) = \sum_{n \in \mathbb{N}} \mu(A_n) < \infty,$$

we require that the series $\sum_{n \in \mathbb{N}} \mu(A_n)$ converges absolutely. In that case $(\mathcal{X}, \mathcal{F}, \mu)$ is called a signed measure space.

We say that a measurable set $A \in \mathcal{F}$ is positive with respect to a signed measure μ if for all measurable subsets B of A it holds

$$\mu(B) \geq 0.$$

Similarly, A is called negative if for all measurable subsets B of A it holds

$$\mu(B) \leq 0.$$

A set is called nullset if it is both positive and negative.

7.1.1 Lemma

Let A be a measurable set, with $\mu(A) > 0$ and $\mu(A) < \infty$, then there exists a positive set $B \subset A$, with $\mu(B) > 0$.

Proof. If A is itself positive then we are done. Suppose A is not positive, then it has a measurable subset E with $\mu(E) < 0$.

Now take $n_1 \in \mathbb{N}$ whereby n_1 is the smallest integer greater than zero, such that there exists a measurable subset of A with $\mu(E_1) < -\frac{1}{n_1}$, whereby we let E_1 be the greatest of subsets, satisfying the given inequality.

For such an E_1 it holds

$$\mu(A \backslash E_1) = \mu(A) - \mu(E_1) > \mu(A) > 0.$$

If $A \backslash E_1$ is positive, then we are done. If $A \backslash E_1$ is not, then we continue by taking the smallest $n_2 \geq n_1$ such that there exists $E_2 \subset A \backslash E_1$ with

$$\mu(E_2) < -\frac{1}{n_2},$$

whereby we let E_2 be the largest of subsets of $A \backslash E_1$ satisfying the upper condition.
If $A \backslash (E_1 \cup E_2)$ is not positive either we continue inductively.

If we reach a $K \in \mathbb{N}$ such that $A \backslash \bigcup\limits_{k=1}^{K-1} E_{n_k}$ is positive, then we are of course done.

However if we never stop, we take

$$B := A \backslash \bigcup_{k \in \mathbb{N}} E_{n_k}$$

and show that B is positive. Notice that the sequence $(E_{n_k})_{k \in \mathbb{N}}$ is pairwise disjoint and of course B and $\bigcup\limits_{k \in \mathbb{N}} E_{n_k}$ are disjoint. We obtain

$$\mu(A) = \mu(B) + \mu\left(\dot{\bigcup_{k \in \mathbb{N}}} E_{n_k}\right) = \mu(B) + \sum_{k \in \mathbb{N}} \mu(E_{n_k}) < \infty.$$

Due to the finiteness of $\mu(A)$ we conclude that the upper series $\sum\limits_{k \in \mathbb{N}} \mu(E_{n_k})$ must converge absolutely, which of course implies

$$\lim_{k \to \infty} \mu(E_{n_k}) = 0 \qquad \text{hence} \qquad \lim_{k \to \infty} n_k = 0.$$

Take an arbitrary $\varepsilon > 0$ then we know that there exists $K \in \mathbb{N}$ such that $\frac{1}{n_K} < \varepsilon$, hence $-\varepsilon < -\frac{1}{n_K}$.

If there was a $E \subset B$ such that $\mu(B) < -\varepsilon$, then E_{n_K} could not have been the greatest subset of $A \backslash \bigcup\limits_{k=1}^{K-1} E_{n_k}$ with measure less than $-\frac{1}{n_K}$, because $E_{n_K} \cup E$ would have been bigger with

$$\mu(E \cup E_{n_K}) < -\frac{1}{n_K}.$$

This, of course is a contradiction to the definition of B. Since $\varepsilon > 0$ was arbitrary we now know that every subset of B has value greater or equal to zero and thus B is positive.

It also holds

$$\mu(B) = \mu(A) - \sum_{n \in \mathbb{N}} \underbrace{\mu(E_{n_k})}_{<0} > \mu(A) > 0.$$

□

7.1.2 Theorem (Hahn Decomposition)

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a signed measure space, then there is a positive set $P \in \mathcal{F}$ and a negative set $N \in \mathcal{F}$, such that

$$P \cup N = \mathcal{X} \qquad \text{and} \qquad P \cap N = \varnothing.$$

This is known as a Hahn Decomposition of signed measures.

Proof. Without loss of generality we may assume that $\mu < \infty$. Define $p := \sup_{P\,:\,P \text{ is positive}} \mu(P)$.

Given that $\varnothing$ is positive , we know $p \geq 0$. By definition of the supremum, there exists a sequence $(P_n)_{n\in\mathbb{N}}$ of positive sets $P_n \subset \mathcal{X}$, such that

$$p = \lim_{n\to\infty} \mu(P_n).$$

We define $P := \bigcup_{n\in\mathbb{N}} P_n$ then P is obviously positive and therefore it holds $\mu(P) \leq p$.

Since it holds for any $n \in \mathbb{N}$

$$\mu(P) = \mu(P_n) + \mu(P \backslash P_n) \geq \mu(P_n),$$

we know that $p = \mu(P) < \infty$.

Now let $N = \mathcal{X} \backslash P$. We claim that N is negative. Assume there exists a $B \subset N$ with $\mu(B) > 0$, then B and P are disjoint and by lemma 7.1.1 there exists a positive $\tilde{B} \subset B$ with $\mu(\tilde{B}) > 0$. Then of course $\tilde{B} \cup P$ is positive as the union of two positive sets. We deduce

$$p \geq \mu(P \dot{\cup} \tilde{B}) = \mu(P) + \mu(\tilde{B}) \geq \mu(P) = p.$$

This shows $\mu(\tilde{B}) = 0$, which was a contradiction to our assumption. Therefore N is negative and we have our decomposition. □

We now know that every signed measure has a Hahn decomposition. While the existence is now guaranteed, the uniqueness is not. Note that a positive/negative set can be united with a nullset and remain positive/negative. Hence the Hahn decomposition is only unique up to nullsets.

This brings us to the Jordan decomposition of signed measures.

7.1.3 Jordan Decomposition

Given a signed measure μ with a Hahn decomposition P and N, we define

$$\mu^+ := \mu(\,\cdot \cap P) \qquad \text{and} \qquad \mu^- := -\mu(\,\cdot \cap N).$$

It is $\mu = \mu^+ - \mu^-$. This is known as a Jordan decomposition of μ and we define

$$|\mu| = \mu^+ + \mu^- \geq 0.$$

Note that while the Hahn decomposition of a signed measure was only unique up to nullsets, the Jordan decomposition is unique and independent of the underlying Hahn decomposition, since the only difference between N and P and another Hahn decomposition A and B are nullsets, which implies

$$\mu(\,\cdot \cap P) = \mu(\,\cdot \cap A)$$

and

$$\mu(\,\cdot \cap N) = \mu(\,\cdot \cap B).$$

Hence μ^+ and μ^- are unique.

This brings us to another not quite so trivial identity of the Total Variation distance.

7.1.4 Theorem

The Total Variation distance is given by

$$||\mu - \nu||_{TV} = \max_{\psi:\mathcal{X}\to[-1,1]} \int\limits_{\mathcal{X}} \psi(x)d\mu(x) - \int\limits_{\mathcal{X}} \psi(x)d\nu(x) = 2 \sup_{A\subset\mathcal{X}} |\mu(A) - \nu(A)|.$$

Proof. Given two measures μ and ν, the difference $\mu - \nu$ is a signed measure. Hence there is a Hahn decomposition N and P of $\mu - \nu$. It also holds $0 = \mu(\mathcal{X}) - \nu(\mathcal{X}) = \mu(P) + \mu(N) - \nu(P) - \nu(N)$. And therefore $\mu(P) - \nu(P) = \nu(N) - \mu(N)$. Then the supremum on the right-hand side is achieved by either P or N. Therefore we are granted a maximum on the right side.

The maximum in the middle is attained by $1_P - 1_N$.

We obtain

$$\begin{aligned} ||\mu - \nu||_{TV} &= \sup_{\psi:\mathcal{X}\to[-1,1]} \int\limits_{\mathcal{X}} \psi(x)d(\mu-\nu)(x) \\ &= \int\limits_{\mathcal{X}} 1_P - 1_N d(\mu-\nu)(x) \\ &= \mu(P) - \nu(P) - \mu(N) + \nu(N) \\ &= 2|\mu(P) - \nu(P)\} \\ &= 2 \sup_{A\subset\mathcal{X}} |\mu(A) - \nu(A)|. \end{aligned}$$

□

It is obvious that for any $\mu \in \mathcal{P}(\mathcal{X})$, $||\mu||_{TV}$ is bounded by 1. Unlike we did with the Wasserstein distance, we don't need to restrict the space of probability measures for Total Variation to be bounded.

7.1.5 Theorem

Total Variation is a metric on $\mathcal{P}(\mathcal{X})$.

Proof. The proof of this is analogous to the proof of theorem 4.1.2. □

The upper identity of the Total Variation distance can be used to prove a bounding property of the Wasserstein metric.

7.2 Topological properties of Total Variation

7.2.1 Theorem

Let μ and ν be two probability measures on a Polish space $(\mathcal{X}, d)$ and $p \in [1, \infty)$. Then it holds for an arbitrary $x_0 \in \mathcal{X}$

$$W_p(\mu, \nu) \leq 2^{\frac{1}{p'}} \left(\int\limits_{\mathcal{X}} d(x_0, x)^p d|\mu - \nu|(x) \right)^{\frac{1}{p}}, \quad \frac{1}{p} + \frac{1}{p'} = 1.$$

Proof. We show this by contriving a coupling of (μ, ν) such that the upper inequality is satisfied.

We know that $\mu - \nu$ is a signed measure and there is a Hahn decomposition P and N of $(\mu - \nu)$. This means

$$(\mu - \nu)^+(A) = \mu(A \cap P) - \nu(A \cap P) \quad \text{and} \quad (\mu - \nu)^-(A) = -\mu(A \cap N) + \nu(A \cap N).$$

Then, of course it holds $(\mu - \nu)^\pm \geq 0$.

We also know that $a =: (\mu - \nu)^+(\mathcal{X}) = (\mu - \nu)^-(\mathcal{X})$, since $(\mu - \nu)(\mathcal{X}) = 0$.

Now define $(\mu \wedge \nu) := \mu - (\mu - \nu)^+$. Take $\pi_1 = (\mu \wedge \nu) \circ (\text{Id,Id})^{-1}$, defined by

$$\begin{aligned}\pi_1(A \times B) &= (\mu \wedge \nu) \circ (\text{Id,Id})^{-1}(\{(x, y) \in \mathcal{X} \times \mathcal{X} \,|\, x \in A, y \in B\}) \\ &= (\mu \wedge \nu)(\{x \in \mathcal{X} \,|\, x \in A, x \in B\}) \\ &= (\mu \wedge \nu)(A \cap B) \\ &= \mu(A \cap B) - (\mu - \nu)^+(A \cap B).\end{aligned}$$

Take $\pi_2 := a^{-1}(\mu - \nu)^+ \otimes (\mu - \nu)^-$, which simply denotes the product measure of the Jordan decomposition.

Now $\pi := \pi_1 + \pi_2$ is obviously a probability measure on $\mathcal{X} \times \mathcal{X}$, since

$$\begin{aligned}\mu(\mathcal{X} \times \mathcal{X}) = \pi_1(\mathcal{X} \times \mathcal{X}) + \pi_2(\mathcal{X} \times \mathcal{X}) &= \mu(\mathcal{X}) - a + \frac{1}{a}a^2 \\ &= \mu(\mathcal{X}) = 1.\end{aligned}$$

The σ-additivity follows from the fact that $(\mu - \nu)^\pm$ are σ-additiv.
In fact, π is a coupling of (μ, ν), since it satisfies the marginal conditions. Let $A, B \subset \mathcal{X}$ be measurable sets, then it holds

$$\begin{aligned}\pi(A \times \mathcal{X}) &= \pi_1(A \times \mathcal{X}) + \pi_2(A \times \mathcal{X}) \\ &= \mu(A \cap \mathcal{X}) - (\mu - \nu)^+(A \cap \mathcal{X}) + a^{-1}(\mu - \nu)^+(A)(\mu - \nu)^-(\mathcal{X}) \\ &= \mu(A) - \mu(A \cap P) + \nu(A \cap P) + a^{-1}a(\mu - \nu)(A \cap P) \\ &= \mu(A) - \mu(A \cap P) + \nu(A \cap P) + \mu(A \cap P) - \nu(A \cap P) \\ &= \mu(A)\end{aligned}$$

and

$$\begin{aligned}\pi(\mathcal{X} \times B) &= \pi_1(\mathcal{X} \times B) + \pi_2(\mathcal{X} \times B) \\ &= \mu(\mathcal{X} \cap B) - (\mu - \nu)^+(\mathcal{X} \cap B) + a^{-1}(\mu - \nu)^+(\mathcal{X})(\mu - \nu)^-(B) \\ &= \mu(B) - \mu(B \cap P) + \nu(B \cap P) + a^{-1}a(\nu - \mu)(B \cap N) \\ &= \mu(B) - \mu(B \cap P) + \nu(B \cap P) + \nu(B \cap N) - \mu(B \cap N) \\ &= \nu(B).\end{aligned}$$

Therefore π denotes a transference plan between μ and ν. We conclude

$$W_p(\mu, \nu)^p \leq \int\limits_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi(x, y) = \int\limits_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi_1(x, y) + \int\limits_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi_2(x, y).$$

Since $\pi_1 = (\mu \wedge \nu) \circ (\text{Id,Id})^{-1}$ it is $\pi_1(A \times B) = (\mu \wedge \nu)(A \cap B)$ and therefore π_1 is concentrated on the diagonal. This means it holds

$$\int\limits_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi_1(x, y) = 0.$$

Recall our arbitrarily given $x_0 \in \mathcal{X}$. We are now left with

$$\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_2(x,y) = \frac{1}{a}\int\limits_{\mathcal{X}}\int\limits_{\mathcal{X}} d(x,y)^p d(\mu-\nu)_+(x) d(\mu-\nu)_-(y).$$

Using Jensen's inequality $(A+B)^p \leq 2^{p-1}(A^p+B^p)$ on $(d(x,x_0)+d(y,x_0))^p$ yields

$$\begin{aligned}
&\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_2(x,y) \\
&\leq \frac{1}{a}\int\limits_{\mathcal{X}}\int\limits_{\mathcal{X}} (d(x,x_0)+d(y,x_0))^p d(\mu-\nu)_+(x) d(\mu-\nu)_-(y) \\
&\leq \frac{2^{p-1}}{a}\int\limits_{\mathcal{X}}\int\limits_{\mathcal{X}} d(x,x_0)^p + d(y,x_0)^p d(\mu-\nu)_+(x) d(\mu-\nu)_-(y) \\
&= \frac{2^{p-1}}{a}\left(\int\limits_{\mathcal{X}}\int\limits_{\mathcal{X}} d(x,x_0)^p d(\mu-\nu)_+(x) d(\mu-\nu)_-(y) + \int\limits_{\mathcal{X}}\int\limits_{\mathcal{X}} d(y,x_0)^p d(\mu-\nu)_+(x) d(\mu-\nu)_-(y)\right) \\
&= \frac{2^{p-1}}{a}\left(a\int\limits_{\mathcal{X}} d(x,x_0)^p d(\mu-\nu)_+(x) + a\int\limits_{\mathcal{X}} d(y,x_0)^p d(\mu-\nu)_-(y)\right) \\
&= 2^{p-1}\left(\int\limits_{\mathcal{X}} d(x,x_0)^p d(\mu-\nu)_+(x) + \int\limits_{\mathcal{X}} d(y,x_0)^p d(\mu-\nu)_-(y)\right) \\
&= 2^{p-1}\int\limits_{\mathcal{X}} d(x,x_0)^p d|\mu-\nu|(x).
\end{aligned}$$

The last equality follows from the fact that $|\mu-\nu| = (\mu-\nu)^+ + (\mu-\nu)^-$. Note, that taking the p-th square root yields the factor $2^{\frac{p-1}{p}}$ and of course we know it holds $\frac{p-1}{p}+\frac{1}{p}=1$. Thus, choosing $p' := \frac{p}{p-1}$ yields the desired inequality. □

7.2.2 Corollary

Let $(\mathcal{X}, d)$ be a metric space. If the diameter of $\mathcal{X}$ is bounded by D, then we obtain

$$W_p(\mu,\nu)^p \leq 2^{p-1}D^p \underbrace{|\mu-\nu|(\mathcal{X})}_{\|\mu-\nu\|_{TV}} = 2^{p-1}D^p\,\|\mu-\nu\|_{TV}$$

and

$$W_p(\mu,\nu) \leq 2^{\frac{p-1}{p}} D \sqrt[p]{\|\mu-\nu\|_{TV}}.$$

In particular, we know

$$W_1 \leq \operatorname{diam}(\mathcal{X})\|\cdot\|_{TV}.$$

This means for a bounded Polish space $(\mathcal{X}, d)$, the topology induced by $\|\cdot\|_{TV}$ is greater than the topology induced by W_p on the Wasserstein space $P_p(\mathcal{X})$. Knowing that a bounded Polish space implies $P_p(\mathcal{X}) = \mathcal{P}(\mathcal{X})$ and that the topology on $\mathcal{P}(\mathcal{X})$ is simply the topology of regular weak convergence, we now conclude that Total Variation induces a topology greater than that of weak convergence on $\mathcal{P}(\mathcal{X})$, if (X, d) is a bounded Polish space.

7.2.3 Corollary

Let $(\mathcal{X}, d)$ be a bounded Polish space and $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ a sequence. Given another probability measure $\mu \in \mathcal{X}$, it holds

$$\|\mu_n - \mu\|_{TV} \longrightarrow 0 \qquad \Longrightarrow \qquad \mu_n \overset{\omega}{\longrightarrow} \mu.$$

There is also another inversed inequality showing that W_p is an upper bound of $\|\cdot\|_{TV}$.

7.2.4 Theorem

Let $(\mathcal{X}, d)$ be a finite metric space. Then it holds

$$W_p \geq d_{\min} \sqrt[p]{\|\cdot\|_{TV}},$$

with $d_{\min} := \min\limits_{x\neq y} d(x, y)$.
This of course implies as a corollary $W_1 \geq d_{\min}\|\cdot\|_{TV}$.

Proof. The inequality follows from the fact that it holds $d(x, y) \geq d_{\min} 1_{[x\neq y]}(x, y)$ for all $x, y \in \mathcal{X}$, which yields

$$W_p(\mu, \nu)^p = \inf_{\pi\in\Pi(\mu,\nu)} \int\limits_{\mathcal{X}} d(x, y)^p d\pi(x, y) \geq \inf_{\pi\in\Pi(\mu,\nu)} \int\limits_{\mathcal{X}} d^p_{\min} \cdot 1_{[x\neq y]}(x, y) d\pi(x, y) = d^p_{\min}\|\cdot\|_{TV}.$$

□

Note that for the second inequality to hold, finiteness was needed. Otherwise $d_{\min}$ could have been 0. On an infinite set $\mathcal{X}$ one could potentially define a sequence $(x_n)_{n\in\mathbb{N}}$ that converges to some x. Then it would hold for the sequence δ_{x_n}

$$W_p(\delta_{x_n}, \delta_x) = d(x_n, x) \longrightarrow 0,$$

but also

$$||\delta_{x_n} - \delta_x||_{TV} = 1_{[x\neq y]}(x_n, x) = 1.$$

Recall that the Wasserstein distance metrized weak convergence on the Wasserstein space $P_p(\mathcal{X})$. What we have proven now, is that, given a bounded Polish space, Total Variation bounds the Wasserstein distance from above, thus metrizing a topology greater than that of weak convergence. If, additionally the Polish space is finite then the Wasserstein distance bounds Total Variation from above, creating a topology greater than the one metrized by Total Variation. Hence we know that on a finite Polish space Total Variation metrizes weak convergence. (Here the boundedness can be neglected, since finite implies bounded). Note here, that given a finite Polish space the Wasserstein space $P_p(\mathcal{X})$ is simply given by the space of probability measures $\mathcal{P}(\mathcal{X})$.

Another interesting approach to the Total Variation metric is taking the metric space $(\mathcal{X}, d)$, where $d = 1_{[x\neq y]}$ is the discrete metric. The topology induced by the discrete metric is the power set of $\mathcal{X}$, since for every $x \in \mathcal{X}$ the set $\{x\}$ is open. Then the Borel σ-algebra on $\mathcal{X}$ is given by the power set too. It is obvious that $(\mathcal{X}, d)$ is complete. $(\mathcal{X}, d)$ is separable if and only if $\mathcal{X}$ is countable. The Wasserstein distance on $(\mathcal{X}, d)$ is now given by

$$W_p(\mu, \nu)^p = \inf_{\pi\in\Pi(\mu,\nu)} \int\limits_{\mathcal{X}\times\mathcal{X}} 1_{[x\neq y]}(x, y) d\pi(x, y) = \|\mu - \nu\|_{TV}.$$

Now recall that the Wasserstein space was given by

$$\Big\{\mu \in \mathcal{P}(\mathcal{X})\,;\, \int\limits_{\mathcal{X}} 1^p_{[x\neq x_0]}(x)d\mu(x) < \infty\Big\}.$$

It is of course

$$\int\limits_{\mathcal{X}} 1^p_{[x\neq x_0]}(x)d\mu(x) \leq \mu(\mathcal{X}) = 1.$$

Hence $P_p(\mathcal{X}) = \mathcal{P}(\mathcal{X})$. At this point it is important to note, that given the discrete metric, $\mathcal{P}(\mathcal{X})$ denotes the set of all probability measures, defined on the power set of $\mathcal{X}$, which is a very limited set.

We recall that given a Polish space $(\mathcal{X}, d)$ the Wasserstein distance metrizes weak convergence on $P_p(\mathcal{X})$, which endowed with the Wasserstein distance is also a Polish space.

We now also know that on this particular metric space, the Total Variation distance and the Wasserstein distance are the same, thus inducing the same topology on $\mathcal{P}(\mathcal{X})$. Given a countable set $\mathcal{X}$ endowed with the discrete metric, the set of all probability measures $\mathcal{P}(\mathcal{X})$ defined on the power set of $\mathcal{X}$, endowed with the topology of regular weak convergence is a Polish space. Furthermore if $(\mathcal{X}, d)$ is compact, which for $d = 1_{[x\neq y]}$ is equivalent to $\mathcal{X}$ finite, then $\mathcal{P}(\mathcal{X})$ again, endowed with the topology of weak convergence is compact, which by Prokhorov's theorem, implies $\mathcal{P}(\mathcal{X})$ is tight.

To recap:
Given a finite set $\mathcal{X}$ the set of all probability measures, that are defined on every subset of $\mathcal{X}$ is a compact (and tight) Polish space. For $\mathcal{X}$ countable, it is only a Polish space.

This fairly well sums up the Total Variation distance. We move on to introducing the reader to another probability metric, namely the Prokhorov metric.

8 The Prokhorov metric

Let $(\mathcal{X}, d)$ be a metric space.

Take two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Then the Prokhorov metric between μ and ν is defined as

$$d_P(\mu, \nu) := \inf\{\epsilon > 0\,;\, \mu(B) \leq \nu(B^\epsilon) + \epsilon \text{ for all Borel sets } B \subset \mathcal{X}\},$$

where we define $B^\epsilon = \{x \in \mathcal{X}; \inf\limits_{y\in B} d(x, y) \leq \epsilon\}$.

8.0.1 Theorem

The Prokhorov metric satisfies all axioms of a metric

Proof. We start off, by proving the symmetry $d_P(\mu, \nu) = d_P(\nu, \mu)$.
Let $\varepsilon := d_P(\mu, \nu)$ and take an arbitrary Borel set $A_1 \subset \mathcal{X}$. We define $A_2 := (A_1^\varepsilon)^c$. It is $A_1 \subset (A_2^\varepsilon)^c$ from which we conclude

$$\mu(A_1^\epsilon) = 1 - \mu(A_2) \geq 1 - \nu(A_2^\epsilon) - \epsilon = \nu\left((A_2^\varepsilon)^c\right) - \epsilon \geq \nu(A_1) - \epsilon$$

and consequently every Borel set $A \subset \mathcal{X}$ satisfies the inequality

$$\nu(A) \leq \mu(A^{\epsilon}) + \epsilon.$$

Note that $B^0 = \overline{B}$ and $d_P(\mu, \nu) = 0$ then implies

$$\mu(A) \leq \nu(\overline{A}) \quad \text{and} \quad \nu(A) \leq \mu(\overline{A}).$$

This yields $\mu(A) = \nu(A)$ for all closed sets $A \subset \mathcal{X}$. From the uniqueness of measures it follows that $\mu = \nu$ on $\mathcal{X}$. The other implication is obvious.

For the triangle inequality, take three arbitrary probability measures $\mu, \nu, \lambda \in \mathcal{P}(\mathcal{X})$ and define $\epsilon := d_P(\mu, \lambda)$ and $\epsilon' := d_P(\nu, \lambda)$.

Then it holds for $A \subset \mathcal{X}$ measurable

$$\mu(A) \leq \lambda(A^{\epsilon}) + \epsilon \leq \nu(A^{\epsilon^{\epsilon'}}) + \epsilon + \epsilon' \leq \nu(A^{\epsilon + \epsilon'}) + \epsilon + \epsilon',$$

and therefore $d_P(\mu, \nu) \leq d_P(\mu, \lambda) + d_P(\lambda, \nu)$. □

We know now that d_P is, in fact a metric on $\mathcal{P}(\mathcal{X})$, which assumes values in $[0, 1]$. While not easy to compute this metric is theoretically important, because it metrizes weak convergence.

8.1 Topological properties of the Prokhorov metric

8.1.1 Theorem

For a sequence of probability measures $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ and $\mu \in \mathcal{P}(\mathcal{X})$ it holds

$$d_P(\mu_n, \mu) \longrightarrow 0 \qquad \Longrightarrow \qquad \mu_n \xrightarrow{\omega} \mu.$$

Proof. This is easily proven, using Portmanteau's theorem. We will show that for every closed set $A \subset \mathcal{X}$ it holds

$$\limsup_{n \to \infty} \mu_n(A) \leq \mu(A).$$

Since $d_P(\mu_n, \mu) \longrightarrow 0$, we know that

$$\forall \varepsilon > 0 \, \exists N \in \mathbb{N} \, \forall n \in \mathbb{N}_{\geq N} : \mu_n(A) \leq \mu(A^{\varepsilon}) + \varepsilon.$$

Letting $\varepsilon \longrightarrow 0$ yields

$$\limsup_{n \to \infty} \mu_n(A) \leq \mu(A).$$

This has shown $\mu_n \xrightarrow{\omega} \mu$. □

We have now shown that for any metric set, the Prokhorov metric metrizes a topology greater than that of weak convergence. We now show that given an additional restriction, it will also induce a topology smaller than that of weak convergence, thereby metrizing weak convergence.

8.1.2 Theorem

Let $(\mathcal{X}, d)$ be a separable metric space. Then for a sequence of probability measures $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ and some $\mu \in \mathcal{P}(\mathcal{X})$ it holds

$$\mu_n \xrightarrow{\omega} \mu \qquad \Longrightarrow \qquad d_P(\mu_n, \mu) \longrightarrow 0.$$

The proof of this is not quite as trivial as the upper one.

Proof. We know that for any open set $B \subset \mathcal{X}$ it holds $\liminf\limits_{n\to\infty} \mu_n(B) \geq \mu(B)$ and subsequently

$$\forall \epsilon > 0 \, \exists N \in \mathbb{N} \, \forall n \in \mathbb{N}_{\geq N} : \, \mu_n(B) \geq \mu(B) - \epsilon.$$

With this in mind we will try to prove that the upper inequality holds for any measurable set A and then conclude the convergence in d_p. Let $(A_n)_{n\in\mathbb{N}}$ be a partition of $\mathcal{X}$, i.e a sequence of pairwise disjoint subsets that cover $\mathcal{X}$.

Additionally we impose that the diameter of these sets must not exceed ϵ, i.e

$$\sup_{x,y\in A_n} d(x,y) \leq \epsilon \qquad \text{f.a. } n \in \mathbb{N}.$$

Note, that the existence of a partition like this follows directly from the fact that any separable metric space is second-countable, i.e its topology has a countable base.

We know that it holds

$$\mu\left(\dot{\bigcup_{n\in\mathbb{N}}} A_n\right) = 1.$$

This, of course means there exists some $k \in \mathbb{N}$ such that

$$\mu\left(\dot{\bigcup_{n>k}} A_n\right) < \epsilon.$$

We define $\mathcal{G}$ as the set of all sets of the form

$$((A_{i_1} \cup \ldots \cup A_{i_m})^\epsilon)^\circ \quad \text{with } \ 1 \leq i_1, \ldots, i_m \leq k.$$

It should be added, that we define $(A^\epsilon)^\circ := \{x \in \mathcal{X};\, d(x,A) < \epsilon\}$. Note that $\mathcal{G}$ then only consists of open sets. Using weak convergence of $(\mu_n)_{n\in\mathbb{N}}$, we know that for every $n \geq N$ and any $G \in \mathcal{G}$ it holds

$$\mu_n(G) \geq \mu(G) - \epsilon.$$

Finally we can go on to prove convergence in the topology, induced by d_P.
Take an arbitrary $A \subset \mathcal{X}$ and define A_0 as the union of all sets A_i $(i \leq k)$, that overlap with A.

It is obvious that $(A_0^\epsilon)^\circ \in \mathcal{G}$ and therefore it holds $\mu((A_0^\epsilon)^\circ) \leq \mu_n(A_0^\epsilon) + \epsilon$. Then it holds for all $n \geq N$

$$\begin{aligned}
\mu(A) = \mu\left(A \ \cap \ \dot{\bigcup_{n\in\mathbb{N}}} A_n\right) &= \mu(A \cap A_0) + \mu\left(A \ \cap \ \dot{\bigcup_{n>k}} A_n\right) \\
&\leq \mu((A_0^\epsilon)^\circ) + \mu\left(\dot{\bigcup_{n>k}} A_n\right) \leq \mu((A_0^\epsilon)^\circ) + \epsilon \\
&\leq \mu_n((A_0^\epsilon))^\circ) + \epsilon + \epsilon \ \leq \ \mu_n(A_0^\epsilon) + \epsilon + \epsilon \\
&\leq \mu_n(A^{2\epsilon}) + 2\epsilon.
\end{aligned}$$

We know now that for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ sucht that for all $n \geq N$ it holds

$$d_P(\mu_n, \mu) \leq 2\epsilon.$$

Hence, we conclude

$$d_P(\mu_n, \mu) \longrightarrow 0.$$

□

We have now shown that for a separable metric space $(\mathcal{X}, d)$ the Prokhorov metric metrizes weak convergence on $\mathcal{P}(\mathcal{X})$. Recall the theorem, stating that if $(\mathcal{X}, d)$ is a Polish space, then $(P_p(\mathcal{X}), W_p)$ is Polish as well. We will show the same for the Prokhorov metric.

8.1.3 Theorem

Let $(\mathcal{X}, d)$ be a Polish space, then $(\mathcal{P}(\mathcal{X}), d_P)$ is a Polish space as well.

Proof. We start off by showing the separability of $(\mathcal{P}(\mathcal{X}), d_P)$.

Let $\varepsilon > 0$ be arbitrary. As we did in the proof above, we take a partition of $(A_n)_{n\in\mathbb{N}}$ of $\mathcal{X}$ such that $A_n \cap A_m = \varnothing$ $(n \neq m)$, with $\operatorname{diam}(A_n) \leq \varepsilon$ and we take a sequence of points $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ such that $x_n \in A_n$. Define

$$D := \Big\{ \sum_{i=1}^{n} r_i \delta_{x_i};\ n \in \mathbb{N}, r_i \in \mathbb{Q}, x_i \in A_i \Big\}.$$

We will show that $\mathcal{D}$ is dense in $(\mathcal{P}(\mathcal{X}), d_P)$ and because it is obviously countable, this then yields the separability. We now take an arbitrary $\mu \in \mathcal{P}(\mathcal{X})$ and show that there exists a measure $\lambda \in \mathcal{D}$ with $d_P(\mu, \lambda) \leq \varepsilon$.

Since $\bigcup_{n\in\mathbb{N}} A_n = \mathcal{X}$ we know that there exists a $k \in \mathbb{N}$ such that

$$\mu\left(\bigcup_{i>k} A_i\right) < \varepsilon.$$

We now choose $r_1, \dots, r_k \in \mathbb{Q}$ such that

$$\sum_{i=1}^{k} r_i = 1 \quad \text{and} \quad \sum_{i=1}^{k} |r_i - \mu(A_i)| < \varepsilon.$$

This of course is feasible, because $\mathbb{Q}$ is dense in $\mathbb{R}$. Now define

$$\nu = \sum_{i=1}^{k} r_i \delta_{x_i}.$$

We show that for any measurable set $A \subset \mathcal{X}$ it holds $\mu(A) \leq \nu(A^\varepsilon) + \varepsilon$. As we did before we take all $A_{i_1}, \dots, A_{i_m}$ $(1 \leq i_1, \dots, i_m \leq k)$ with $A_{i_j} \cap A \neq \varnothing$ and define $A_0 = \bigcup\limits_{j=1}^{m} A_{i_j}$. With the fact that the A_i are disjoint we obtain

$$\begin{aligned}\mu(A) &= \mu(A \cap \bigcup_{i\leq k} A_i) + \mu(A \cap \bigcup_{i>k} A_i) \leq \mu(A \cap A_0) + \varepsilon \\ &\leq \sum_{j=1}^{m} \mu(A_{i_j}) + \varepsilon \leq \sum_{j=1}^{m} r_{i_j} + \varepsilon + \varepsilon = \nu(A_0) + 2\varepsilon \leq \nu(A^\varepsilon) + 2\varepsilon \\ &\leq \nu(A^{2\varepsilon}) + 2\varepsilon.\end{aligned}$$

Hence we know that for any $\varepsilon > 0$ and any measurable $A \subset \mathcal{X}$ it holds

$$\mu(A) \leq \nu(A^{2\varepsilon}) + 2\varepsilon,$$

which means $d_P(\mu, \nu) \leq 2\varepsilon$. Therefore we know that for any $\mu \in \mathcal{P}(\mathcal{X})$ and any $\varepsilon > 0$ there exists $\nu \in \mathcal{D}$ such that $d_P(\mu, \nu) \leq \varepsilon$. Consequently $\mathcal{D}$ is dense in $\mathcal{P}(\mathcal{X})$. This concludes the proof that $(\mathcal{P}(\mathcal{X}), d_P)$ is separable. Note that we didn't use the completeness of $(\mathcal{X}, d)$ in this part ,which means the separability of $(\mathcal{X}, d)$ directly implies the separability of $(\mathcal{P}(\mathcal{X}), d_P)$.

For the completeness, let $(\mu_n)_{n\in\mathbb{N}}$ be a d_P-Cauchy sequence. We will show that $(\mu_n)_{n\in\mathbb{N}}$ is tight. Tightness will then imply that there exists a convergent subsequence via Prokhorov's theorem, through which we will obtain convergence of $(\mu_n)_{n\in\mathbb{N}}$.

Take an arbitary $m \in \mathbb{N}$ and $\varepsilon > 0$. Knowing that $(\mu_n)_{n\in\mathbb{N}}$ is a d_P-Cauchy sequence, we can find find $N_m \in \mathbb{N}$ such that for all $n \in \mathbb{N}_{>N_m}$ it holds

$$d_P(\mu_n, \mu_{N_m}) \leq 2^{-m}\varepsilon.$$

Given the fact that on a Polish space every finite set of probability measures is tight we can find a compact set $K \subset \mathcal{X}$ such that

$$\sup_{j\leq N_m} \mu_j(K^c) \leq 2^{-m}\varepsilon.$$

By compactness we know that for every $m \in \mathbb{N}$ K can be covered by finitely many balls of radius $2^{-m}\varepsilon$, i.e for all $m \in \mathbb{N}$, there exists $\tilde{N}_m \in \mathbb{N}$ such that

$$K \subset \bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m}\varepsilon).$$

The fact that it holds $d_P(\mu_{N_m}, \mu_n) \leq 2^{-m}\varepsilon$ yields for all $n \in \mathbb{N}_{\geq N_m}$

$$\begin{aligned}\mu_n\left(\bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m+1}\varepsilon)\right) &\geq \mu_n\left(\left(\bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m}\varepsilon)\right)^{2^{-m}\varepsilon}\right)\\ &\geq \mu_N\left(\bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m}\varepsilon)\right) - 2^{-m}\varepsilon\\ &\geq 1 - 2^{-m}\varepsilon - 2^{-m}\varepsilon\\ &= 1 - 2^{-m+1}\varepsilon.\end{aligned}$$

Therefore it holds for all $n \in \mathbb{N}$

$$\mu_n\left(\left(\bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m+1}\varepsilon)\right)^c\right) \leq 2^{-m+1}\varepsilon.$$

The problem is that

$$\bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m+1}\varepsilon)$$

isn't compact. Define now

$$C := \bigcap_{m\in\mathbb{N}} \bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m+1}\varepsilon).$$

Then C is closed and totally bounded and therefore compact. It holds for $n \in \mathbb{N}$ arbitrary

$$\mu_n(C^c) \leq \sum_{m\in\mathbb{N}} \mu_n\left(\left(\bigcup_{i=1}^{\tilde{N}_m} \overline{B}(x_i, 2^{-m+1}\varepsilon)\right)^c\right) \leq \sum_{m\in\mathbb{N}} 2^{-m+1}\varepsilon = 2\varepsilon.$$

To recap, we know that for every $\varepsilon > 0$ there exists $C \subset \mathcal{X}$ compact, such that

$$\sup_{n\in\mathbb{N}} \mu_n(C^c) \leq 2\varepsilon.$$

Therefore $(\mu_n)_{n\in\mathbb{N}}$ is tight and by Prokhorov's theorem, there exists a convergent subsequence of $(\mu_n)_{n\in\mathbb{N}}$. Knowing that a Cauchy sequence with convergent subsequence is already convergent itself, this directly implies the convergence of $(\mu_n)_{n\in\mathbb{N}}$. Thus $(\mathcal{P}(\mathcal{X}), d_P)$ is complete. □

An interesting aspect of this proof, was the fact that every d_P-Cauchy sequence is tight, analogously to Lemma 5.1.2.

Recall that for $(\mathcal{X}, d)$ Polish, the Wasserstein metric metrized weak covergence on $P_p(\mathcal{X})$. For a bounded metric d, we also knew that the Wasserstein space $P_p(\mathcal{X})$ was equal to the space of all probability measures $\mathcal{P}(\mathcal{X})$. This means that since both, Wasserstein and Prokhorov metrize weak convergence they must be equivalent, at least for a bounded Polish space. That means there must exist some kind of lower and upper bound of Wasserstein by Prokhorov. We introduce a Lemma with which we will show exactly this.

8.1.4 Lemma

Let $(\mathcal{X}, d)$ be a separable metric space and $\mu, \nu \in \mathcal{P}(\mathcal{X})$ be two probability measures, such that $d_P(\mu, \nu) < \alpha$ for some $\alpha \in \mathbb{R}_+$.

Then there exists some probability measure π on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν (i.e. a coupling of (μ, ν)) such that

$$\pi\left((x, y) \in \mathcal{X} \times \mathcal{X} \,|\, d(x, y) \geq \alpha\right) \leq \alpha.$$

The proof of this Lemma is quite long and complicated. It can be found in [4].

Another important Lemma needed, in order to prove the bounding of the Wasserstein distance by the Prokhorov metric is Markov's inequality.

8.1.5 Lemma

Let X be a nonnegative random variable on a measurable space $(\Omega, \mathbb{P})$ and $c > 0$ then it holds

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}.$$

8.2 Bounding properties of the Prokhorov metric

This brings us to the main point, the bounding properties of the Prokhorov metric and the Wasserstrein metric.

8.2.1 Theorem

Let $(\mathcal{X}, d)$ be a bounded Polish space then the Wasserstein metric of order 1 and Lévy-Prokhorov satisfy

$$(d_P)^2 \leq W_1 \leq (\operatorname{diam}(\mathcal{X}) + 1) d_P.$$

Proof. We start with the right-hand side. Choose $\varepsilon > 0$ such that $d_P(\mu, \nu) < \varepsilon$ and choose a coupling π of (μ, ν) such that

$$\pi\left((x, y) \in \mathcal{X} \times \mathcal{X} \,|\, d(x, y) \geq \varepsilon\right) \leq \varepsilon,$$

which exists according to lemma 8.1.4. We obtain

$$\begin{aligned}\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)d\pi(x,y) &\leq \varepsilon\pi\left(d(x,y)\leq\varepsilon\right)+\mathrm{diam}(\mathcal{X})\,\pi\left(d(x,y)>\varepsilon\right)\\ &= \varepsilon+(\mathrm{diam}(\mathcal{X})-\varepsilon)\cdot\pi(d(x,y)>\varepsilon)\\ &\leq \varepsilon+(\mathrm{diam}(\mathcal{X})-\varepsilon)\varepsilon\\ &\leq (\mathrm{diam}(\mathcal{X})+1)\varepsilon.\end{aligned}$$

Taking the infimum over all ε on the right-hand side such that $d_P(\mu,\nu)\leq\varepsilon$ yields

$$W_1\leq(\mathrm{diam}(\mathcal{X})+1)d_P.$$

Now to the left inequality. We set $\varepsilon := \sqrt{W_1(\mu,\nu)}$. Then Markov's inequality over the optimal coupling of (μ,ν) yields

$$\pi(d(x,y)\geq\varepsilon)\leq\frac{1}{\varepsilon}\int\limits_{\mathcal{X}\times\mathcal{X}} d(x,y)d\pi(x,y)\leq\frac{1}{\varepsilon}\varepsilon^2=\varepsilon.$$

Knowing this, we can conclude that for any Borel set B it holds

$$\begin{aligned}\mu(B)=\pi(B\times\mathcal{X}) &\leq \pi(B\times B^{\varepsilon})+\pi(B\times\{y\in\mathcal{X};\ \inf_{x\in B}d(x,y)>\varepsilon\})\\ &\leq \pi(B\times B^{\varepsilon})+\pi(d(x,y)>\varepsilon)\\ &\leq \pi(\mathcal{X}\times B^{\varepsilon})+\varepsilon\\ &= \nu(B^{\varepsilon})+\varepsilon,\end{aligned}$$

which means $d_P(\mu,\nu)\leq\varepsilon$ and therefore $(d_P)^2\leq W_1$. □

We included something very interesting in the proof of the upper assertion. Recall lemma 8.1.4, where we stated, that if for two probability measures μ and ν it holds $d_P(\mu,\nu)<\alpha$, then we know that there exists a coupling π of (μ,ν) such that

$$\pi(d(x,y)\geq\alpha)\leq\alpha.$$

This means that

$$\inf_{\pi\in\Pi(\mu,\nu)}\pi(d(x,y)>\geq\alpha)\leq d_P(\mu,\nu)\qquad\text{for all }\alpha>d_P(\mu,\nu),$$

and therefore

$$\inf\{\alpha>0;\ \inf_{\pi\in\Pi(\mu,\nu)}(\pi(\mu,\nu)\geq\alpha)\leq\alpha\}\leq d_P(\mu,\nu).$$

Now we have shown in the upperhand proof, that if $\pi(d(x,y)\geq\alpha)\leq\alpha$, then it also holds $d_P(\mu,\nu)<\alpha$. Hence

$$d_P(\mu,\nu)\leq\inf\{\alpha>0;\ \inf_{\pi\in\Pi(\mu,\nu)}(\pi(\mu,\nu)\geq\alpha)\leq\alpha\}.$$

We obtain the equality

$$d_P(\mu,\nu)=\inf\{\varepsilon>0;\ \inf_{\pi\in\Pi(\mu,\nu)}\pi\left(d(x,y)>\varepsilon\right)\leq\varepsilon\}.$$

Before we move on to introducing the reader to another probability metric, we will examine another bounding property of Lévy-Prokhorov with Total Variation.
Recall that

$$||\mu-\nu||_{TV}=\max_{|\phi|\leq1}\int\limits_{\mathcal{X}}\phi(x)d(\mu-\nu)(x).$$

8.2.2 Theorem

Let $(\mathcal{X}, d)$ be a metric space and $\mu, \nu \in \mathcal{P}(\mathcal{X})$ two arbitrary probability measures. Then

$$d_P(\mu, \nu) \leq ||\mu - \nu||_{TV}.$$

Proof. Take $\varepsilon > 0$ such that $||\mu - \nu||_{TV} \leq \varepsilon$, then we know that for every measurable function ϕ with $|\phi| \leq 1$ it is

$$\int\limits_{\mathcal{X}} \phi(x) d(\mu - \nu)(x) \leq \varepsilon,$$

and therefore

$$\int\limits_{\mathcal{X}} 1_A(x) d\mu(x) - \int\limits_{\mathcal{X}} 1_A(y) d\nu(y) = \mu(A) - \nu(A) \leq \varepsilon \quad \text{for all measurable } A, B \subset \mathcal{X},$$

which means $\mu(A) \leq \nu(A) + \varepsilon \leq \nu(A^\varepsilon) + \varepsilon$. Consequently it holds $d_P(\mu, \nu) \leq \varepsilon$ and we conclude $d_P \leq \| \cdot \|_{TV}$. □

9 Lévy metric

The Levy metric is designed to be the equivalent of Prokhorov on $\mathbb{R}$ and is defined as follows:

Let μ and ν be two probability measures on $\mathbb{R}$ and F and G their respective distribution functions. Then

$$d_L(\mu, \nu) := d_L(F, G) := \inf\{\varepsilon > 0 \,:\, G(x - \varepsilon) \leq F(x) \leq G(x + \varepsilon), \forall x \in \mathbb{R}\}.$$

Recall that $F(x) = \mu((-\infty, x])$ and $G(x) = \nu((-\infty, x])$.

9.0.1 Theorem

The Lévy metric d_L satisfies all axioms of a metric.

Proof. $d_L \geq 0$ follows directly from the fact that we take the supremum over all $\varepsilon > 0$.
$d_L(\mu, \nu) = 0$ directly implies $F = G$ and therefore $\mu(-\infty, x]) = \nu((-\infty, x])$ for all $x \in \mathbb{R}$. We know that the set of all intervalls of the form $(-\infty, a]$ generates the Borel σ-algebra on $\mathbb{R}$. It satisfies the conditions of the uniqeness theorem of measures and since it holds $d_L(\mu, \nu) = 0$, we know that $\mu = \nu$. The other implication is obvious.

The symmetry is fairly obvious too. Let $\varepsilon := d_L(\mu, \nu)$. Then we know that

$$\mu((-\infty, x - \varepsilon]) - \varepsilon \leq \nu((-\infty, x]) \leq \mu((-\infty, x + \varepsilon]) + \varepsilon,$$

and therefore it holds

$$\begin{gathered} \nu((-\infty, x - \varepsilon]) \leq \mu((-\infty, x - \varepsilon + \varepsilon]) + \varepsilon \\ \text{and} \\ \nu((-\infty, x + \varepsilon]) \geq \mu((-\infty, x + \varepsilon - \varepsilon]) - \varepsilon. \end{gathered}$$

Hence

$$\nu((-\infty, x - \varepsilon]) - \varepsilon \leq \mu((-\infty, x]) \leq \nu((-\infty, x + \varepsilon]) + \varepsilon,$$

which means $\varepsilon = d_L(\nu, \mu)$, thereby yielding the symmetry.

For the triangle inequality, let's take three abritrary measures $\mu, \nu, \lambda \in \mathcal{P}(\mathcal{X})$ with $\varepsilon := d_L(\mu, \lambda)$ and $\varepsilon' := d_L(\lambda, \nu)$ and therefore it holds

$$\mu((-\infty, x-\varepsilon]) - \varepsilon \leq \lambda((-\infty, x]) \leq \mu((-\infty, x+\varepsilon]) + \varepsilon$$

and

$$\lambda((-\infty, x-\varepsilon']) - \varepsilon' \leq \nu((-\infty, x]) \leq \lambda((-\infty, x+\varepsilon']) + \varepsilon'.$$

We obtain

$$\begin{aligned}\mu((-\infty, x-\varepsilon-\varepsilon']) - \varepsilon - \varepsilon' &\leq \lambda((-\infty, x-\varepsilon']) - \varepsilon' \\ &\leq \nu((-\infty, x]) \\ &\leq \lambda((-\infty, x+\varepsilon']) + \varepsilon' \\ &\leq \mu((-\infty, x+\varepsilon'+\varepsilon]) + \varepsilon' + \varepsilon\end{aligned}$$

and we deduce

$$d_L(\mu, \nu) \leq \varepsilon' + \varepsilon.$$

□

We know that d_L is a metric on $\mathcal{P}(\mathbb{R})$. Note that of course $(-\infty, x]$ is a measurable set in the Borel σ-algebra on $\mathbb{R}$ and it is $(-\infty, x]^{\epsilon} = (-\infty, x+\epsilon])$. Thereupon d_L is nothing but a weaker version of d_P on $\mathbb{R}$. Ergo it holds

$$d_L \leq d_P.$$

That bounding property justifies why the topology on $\mathcal{P}(\mathbb{R})$, induced by d_L is smaller than that of weak convergence on $\mathbb{R}$, i.e for a sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ and μ in $\mathcal{P}(\mathbb{R})$ it holds

$$\mu_n \overset{\omega}{\to} \mu \qquad \Longrightarrow \qquad d_L(\mu_n, \mu) \longrightarrow 0.$$

What's interesting about this is, that in addition to being bounded by d_P, d_L metrizes weak convergence on $\mathbb{R}$.

9.0.2 Theorem

Let $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R})$ be a sequence with $d_L(\mu_n, \mu) \longrightarrow 0$ for some $\mu \in \mathcal{P}(\mathbb{R})$.

Then it holds $\mu_n \xrightarrow{\omega} \mu$.

Proof. We know that

$$\forall \varepsilon > 0\, \exists N \in \mathbb{N}\, \forall n \geq N \,:\, d_L(\mu_n, \mu) < \varepsilon$$

i.e

$$\mu((-\infty, x-\varepsilon]) - \varepsilon \;\leq\; \mu_n((-\infty, x]) \;\leq\; \mu((-\infty, x+\varepsilon]) + \varepsilon$$

for all $x \in \mathbb{R}$.

We conclude

$$\begin{aligned}\forall x \in \overline{\mathbb{R}}: \quad &\limsup_{n\to\infty} \mu_n((-\infty, x]) \leq \mu((-\infty, x]) \\ &\liminf_{n\to\infty} \mu_n((-\infty, x]) \geq \mu((-\infty, x]),\end{aligned}$$

and therefore

$$\lim_{n\to\infty} \mu_n((-\infty, x]) = \mu((-\infty, x]).$$

Taking the complement of $(-\infty, x]$ then obviously yields

$$\forall x \in \overline{\mathbb{R}}: \quad \lim_{n\to\infty} \mu_n((x, \infty)) = \mu((x, \infty)),$$

and we obtain for arbitrary $a, b \in \mathbb{R}, a \leq b$

$$\begin{aligned}
\lim_{n\to\infty} \mu_n((a, b]) &= \lim_{n\to\infty} \mu_n((-\infty, b] \cap (-\infty, a]^c) \\
&= \lim_{n\to\infty} \mu_n \left(((b, \infty)\dot{\cup}(-\infty, a])^c\right) \\
&= 1 - \lim_{n\to\infty} \mu_n \left((b, \infty)\dot{\cup}(-\infty, a]\right) \\
&= 1 - \lim_{n\to\infty} \mu_n \left((b, \infty)\right) - \mu_n((-\infty, a]) \\
&= 1 - \mu((b, \infty)) - \mu((-\infty, a]) \\
&= \mu((a, b]).
\end{aligned}$$

This yields weak convergence of the sequence $(\mu_n)_{n\in\mathbb{N}}$. □

Therefore d_L induces the topology of weak convergence on $\mathcal{P}(\mathbb{R})$. Recall that given a Polish space $(\mathcal{X}, d)$, the space $(\mathcal{P}(\mathcal{X}), d_P)$ was also Polish. Now note that d_P and d_L induce the same topology on $\mathcal{P}(\mathbb{R})$, namely the topology of weak convergence. Since $\mathbb{R}$ endowed with the regular notion of distance, is obviously a Polish space then $(\mathcal{P}(\mathbb{R}), d_P) = (\mathcal{P}(\mathbb{R}), d_L)$ is too.

10 Conclusion

Let's recall the problem we established in the introduction. We asked ourselves, what the distance between Europe and Asia might be. This concept of distance was vague and unclear. The first probability metric, we introduced was the Wasserstein distance. The Wasserstein distance was derived from Monge's optimal transport problem and quantified the price of transportation from production units in one place to consumption units in another. Of course, this was only a particular interpretation of this problem. Since the Wasserstein distance was defined using the metric as a cost function, it gives us an idea of the distance between two, by probability measures described spaces. The idea behind it was to lend weight to different places within the given spaces and to then take the distance between the two. In the upper example, this could for instance be done, by taking the distance between every square metre in Asia and in Europe and to then average out these distances. The Wasserstein distance takes the infimum over all ways to weigh these individual points. The whole point is, that this is done by couplings, which keep the spatial distributions of Europe and Asia intact. The minimising coupling of the two continents, would then be given by a probability measure that lends more weight to eastern parts of Europe and western parts of Asia. Therefore the Wasserstein metric gives us a very useful approach to distance between spatial distributions. The Wasserstein distance metrizes weak convergence. In the example above this could be interpreted as follows. If the spatial distributions converge, so does the distance between the objects that are characterised by them. The only downside to this is that one has to reduce the space of probability metrics, to a subspace. Nevertheless the Wasserstein distance has additional nice properties, such as the transference of properties of the underlying metric space onto the Wasserstein space.

The next concept of distance was given by the Total Variation distance. Total Variation is considered the least useful of the set of metrics we introduced here. Under the condition of boundedness it provides an upper bound to the Wasserstein distance and is therefore stronger than Wasserstein. For a finite space it is also a lower bound to W_p. Under these restrictive conditions it metrizes weak convergence. Another aspect of Total Variation is that it can also be interpreted as a particular case of the Wasserstein distance, using the discrete metric. An interesting conclusion to come to is that given a separable space, Total Variation metrizes weak convergence on the space of all probability measures that are defined on the power set. This conclusion isn't really that useful, since this would also follow directly from the Wasserstein distance. Therefore Total Variation is probably the least common probability distance.

The next probability metric given was the Prokhorov distance. This is arguably the most useful probability metric. In terms of the example problem, given in the introduction it can be interpreted as the smallest value, by which we can increase the radius of any place in Asia such that the weight given to said space differs from the weight of any place in Europe by less than said value. Prokhorov metrized weak convergence on the space of all probability measures on the σ-algebra of a Polish metric space. Additionally we proved that this space, endowed with the Prokhorov metric is also Polish. The useful aspect of this metric, was that it had all of the properties that the Wasserstein metric had on the Wasserstein space, without having to restrict it to a subspace. The downside to this metric is the fact, that it is not easy to compute. Hence it is theroetically important, but in practice the Wasserstein metric proves itself more useful.

One other metric we introduced was the Lévy metric, which was an equivalent version of the Prokhorov metric on $\mathbb{R}$. This metric had the seame properties as the Prokhorov metric did on $\mathbb{R}$. Hence the space of probability measures on the euclidean line is a Polish space.

11 Outlook on miscellaneous probability distances

We have synopsised some of the most important probability distances. As mentioned above, there is a large variety of others to be discovered. Before we finish, we give the reader a brief overview of some other probability metrics and their applications.

Kolmogorov metric

Given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with respective distribution functions F and G, then the Kolmogorov metric is given by

$$d_K(\mu, \nu) := \sup_{x \in \mathbb{R}} |F(x) - G(x)|.$$

Obviously it is $d_K \leq 1$ and the Kolmogorov metric bounds the Lévy metric from above, i.e

$$d_L(\mu, \nu) \leq d_K(\mu, \nu).$$

This means that convergence in the topology induced by d_K implies weak convergence and that the topology given by d_K is greater than the topology of weak convergence on $\mathcal{P}(\mathbb{R})$.

An inversed inequality exists as well, given by

$$d_K(\mu, \nu) \leq \left(1 + \sup_{x \in \mathbb{R}} |G'(x)|\right) d_L(\mu, \nu),$$

whereby absolute continuity of G with respect to the lebesgue measure is required. This would mean, restriciting the space of probability measures to the set of all probability measures with absolutely continuous distribution functions, endowed with the topology of weak convergence can be metrized by the Kolmogorov metric.

Discrepancy metric

Given a metric space $(\mathcal{X}, d)$, the discrepancy metric between two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is given by

$$d_D(\mu,\nu) := \sup \left\{ \mu(B) - \nu(B)\,;\, B \subset \mathcal{X} \text{ closed} \right\}.$$

d_D assumes values in $[0,1]$. One can see that the discrepancy metric depends on the metric d of the underlying space $\mathcal{X}$, though it is scale-invariant, since multiplying a metric by a factor yields the same topology, containing the same closed balls over which the supremum is taken. The discrepancy metric satisfies a variety of inequalities, including being bounded from above by the Wasserstein metric.

If $\mathcal{X}$ is finite, then

$$d_{\min} d_D \leq d_W,$$

where, as before $d_{\min} := \min\limits_{x \neq y} d(x,y)$ denotes the smallest distance between two distinct points in $\mathcal{X}$
Note that the finiteness was needed in order for $d_{\min} \geq 0$ to be strictly greater than zero.

For $\mathcal{X} = \mathbb{R}$ it holds

$$d_K(\mu,\nu) \leq d_D(\mu,\nu) \leq 2 d_K(\mu,\nu).$$

The discrepancy metric is also directly bounded by Total Variation, $d_D \leq \|\cdot\|_{TV}$. Another interesting and useful metric is given by the Hellinger distance.

The Hellinger distance

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Given two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ with distribution functions f and g with respect to a dominating measure λ, i.e

$$\mu(A) = \int\limits_A f(x) d\lambda(x)$$

and

$$\nu(B) = \int\limits_B g(x) d\lambda(x),$$

then the Hellinger distance is given by

$$\left(\int\limits_{\mathcal{X}} \left(\sqrt{f} - \sqrt{g} \right)^2 d\lambda \right)^{\frac{1}{2}}.$$

An important aspect of d_H is that Total Variation induces the same topology on $\mathcal{P}(\mathcal{X})$ as the Hellinger distance, which means that for a finite Polish space $(\mathcal{X}, d)$, the Hellinger distance d_H induces the topology of weak convergence on $(\mathcal{P}(\mathcal{X}), d_H)$.

References

[1] Alsion, L., Gibbs, and Edward, F. SU. (2002) *On choosing and bounding probability metrics,* pp.419-435, in *International Statistical Review,* vol. 70, no. 3.

[2] Barrio, E.D., Cuesta-Albertos, J.A., and Matran, C. (1999) *Tests of goodness of fit based on the L_2-Wasserstein distance,* pp.1230-1239, in *The Annals of Statistics,* vol. 27, no. 4.

[3] Basso, G. (2015) *A Hitchhiker's guide to Wasserstein distances.*

[4] Billingsley, P. (2013) *Convergence of probability measures,* John Wiley & Sons, Inc.

[5] Bobkov, S., and Ledoux, M. (2014) *One-dimensional empirical measures, order statistics and Kantorovich transport distances, preprint.*

[6] Bolley, F. (2008) *Separability and completeness for the Wasserstein distance,* pp.371-377, in *Séminaire de probabilités XLI,* Springer Berlin Heidelberg.

[7] Cabrelli, C. A., and Molter, U. M. (1995) *The Kantorovich metric for probability measures on the circle,* pp.345-361, in *Journal of Computational and Applied mathematics,* 57.3.

[8] den Hollander, F. (2012) *Probability Theory: The coupling method,* in *Lecture notes.*

[9] Dudley, R.M. (2010) *Distances of probability measures and random variables,* pp. 28-37, in *Selected Works of RM Dudley.* Springer New York.

[10] Givens, C. R., and Shortt, R. M. (1984) *A class of Wasserstein metrics for probability distributions,* pp.231-240, in *The Michigan Mathematical Journal,* 31.2.

[11] Kloeckner, B. (2008) *A geometric study of Wasserstein spaces: Euclidean spaces,* in *arXiv preprint arXiv: 0804.3505.*

[12] Kupper, M. (2015) *Stochastics II,* in *Lecture notes.*

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

[14] Rachev, S.T., Klebanov, L., Stoyanov, S.V., and Fabozzi, F. (2013) *The methods of distances in the theory of probability and statistics,* pp.11-31, Springer Science & Business Media.

[15] Villani,C. (2008) *Optimal transport, old and new,* pp.41-123, Springer Verlag.

Acknowledgements

This manuscript contains a Bachelor thesis, written at the University of Konstanz during the winter semester 2016/17, supervised by Michael Kupper. I would like to thank Michael Kupper and Daniel Bartl for their support and supervision in writing this manuscript.
`alexander.stannat@gmail.com`

Erklärung der Selbstständigkeit

Ich versichere hiermit, dass ich die vorliegende Arbeit zum Thema

Distances and metrics on probability measures

selbstständig verfasst und keine anderen Hilfsmittel als die angegebenen benutzt habe.
Die Stellen, die anderen Werken dem Wortlaut oder dem Sinne nach entnommen sind, habe ich in jedem einzelnen Fall durch Angaben von Quelle, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht.
Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Konstanz, 16.03.2017 Alexander Stannat