

Optimal Transport

For not-so-optimal mathematicians

Štěpán Zapadlo

Table of contents

Preface	3
1 Introduction	4
1.1 Motivation	4
1.1.1 Empirical Measures	5
2 Optimal Transport	8
2.1 Monge Formulation	8
2.2 Kantorovich Formulation	9
2.2.1 Excursion to Convex Analysis	16
2.2.2 Disintegration of Measures	16
2.2.3 c -Cyclical Monotonicity	22
3 Relevant Results	23
3.1 A Note on Weak Topology	25
3.1.1 Canonical Duality	25
3.1.2 Weak* Topology	25
References	27

Preface

These notes follow the course **Optimal Transport** (506.116) taught by Ass.Prof. Dipl.-Ing. Dr.rer.nat. Gudmund Pammer at the Institute of Statistics, University of Technology, Graz, as it appeared in the Summer semester 2026.

Most of the material comes from my hand-written notes from the lectures, however there are some additions to make the material (in my opinion) more understandable and/or complete. Use of additional material is always explicitly pointed out and cited. Nevertheless, I would like to thank Matthias Höfler, among others, for their handwritten and LaTeX notes, which I referenced at certain points.

The typesetting was done in Quarto, and hence they come in both PDF and [web](#) versions. Feel free to use whichever feels more comfortable for you. They were put together as a preparation for the final exam, and as such are bound to contain typos and errors. I will be very happy if you report and/or fix them, either by email (stepan@zapadlo.name) or by submitting an Issue/Merge Request on the [git repository](#).

1 Introduction

1.1 Motivation

Let us, for the moment, picture the Euclidean space $(\mathbb{R}^n, |\cdot|)$. Crucially, it is a complete separable metric space, i.e., a Polish space (see Definition 3.1), and it has a (simple) metric geometry together (geodesics and averages) with differential calculus. Can we do the same on probability measures?

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be from the space of Borel probability measures on $\mathbb{R}^d$.

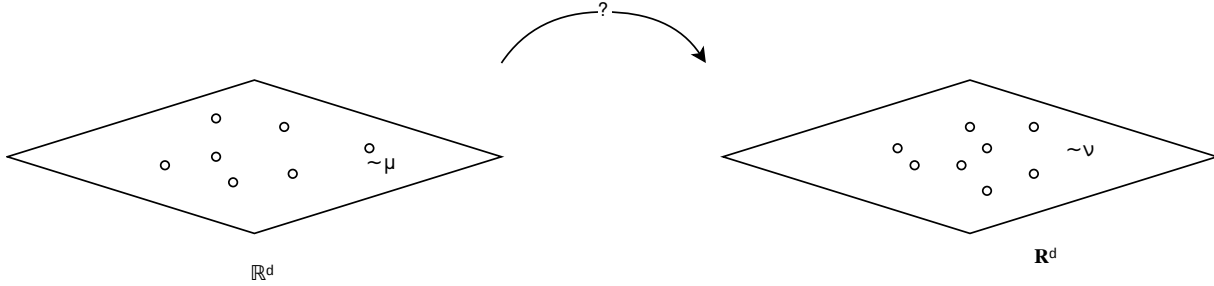


Figure 1.1: Sketch of optimal transport for discrete measures

For single-point Dirac measure (both the source and target one), one can transport it by the geodesic in $\mathbb{R}^d$. Alternatively, we could consider a linear average, see Figure 1.2.

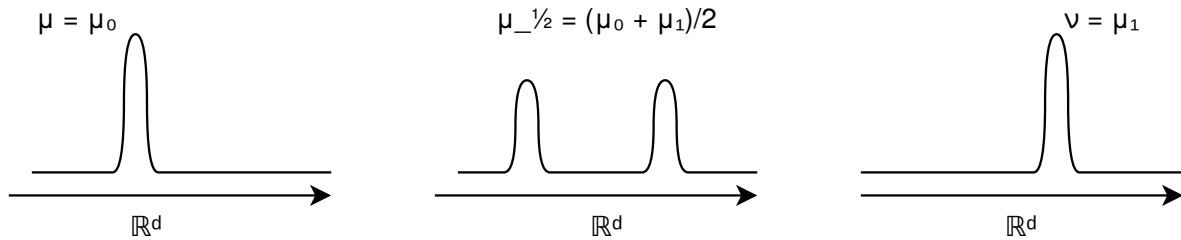


Figure 1.2: Linear average of Dirac measures

It turns out this is the geodesic with respect to the so-called *total variation distance*, see [1]. However, such transport does not take into account the geometry of the underlying base space!

Question 1: Distance of Probabilities

What is a *natural* distance on $\mathcal{P}(\mathbb{R}^d)$?

Question 2: Geometry & Calculus of Probabilities

What is an appropriate metric geometry and differential calculus on $\mathcal{P}(\mathbb{R}^d)$?

Let us start by considering the space of Dirac measures, i.e., $\mu = \delta_x, \nu = \delta_y$. The main **idea** will be to lift the geometry of $\mathbb{R}^d$ to the space of probability measures $d(\delta_x, \delta_y) := |x - y|$. Similarly, we can define the average of δ_x and δ_y as $\delta_{\frac{x+y}{2}}$. Can this be generalized?

1.1.1 Empirical Measures

Let us consider μ, ν to be convex combinations of Dirac measures, i.e.,

$$\mu = \sum_{i=1}^n \mu_i \delta_{x_i}, \quad \& \quad \nu = \sum_{j=1}^m \nu_j \delta_{y_j},$$

and denote $\mathcal{X} := \{x_1, \dots, x_n\}, \mathcal{Y} := \{y_1, \dots, y_m\}$. The first (naive) idea is to transport “mass at x_i to mass at y_j ”, i.e., define a **transport map** $T: \mathcal{X} \rightarrow \mathcal{Y}$ s.t.

$$T_{\#}\mu = \sum_{i=1}^n \mu_i \delta_{T(x_i)} = \nu,$$

where $T_{\#}\mu$ is the *pushforward* of μ with respect to T . Surely, there are often many different transport maps (among which we must choose) and there exist various visualizations, see Figure 1.3.

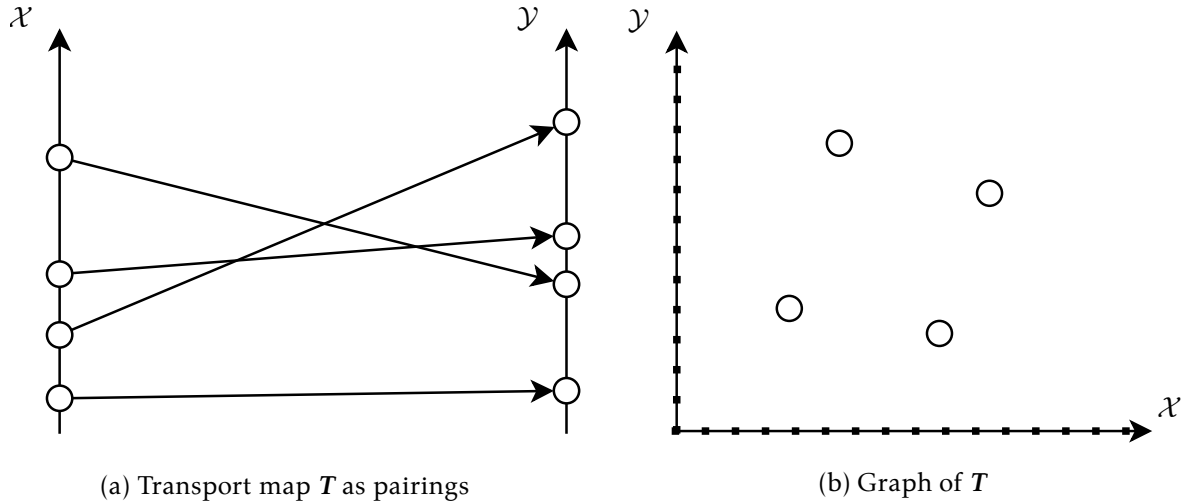


Figure 1.3: Visualizations of a transport map T

We follow up by another **idea** of selecting the “optimal” map by the “least action principle”. That is, we take a cost function $c: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$. Given a transport map, it induces the corresponding *transport cost* $\sum_{i=1}^n \mu_i c(x_i, T(x_i))$. This leads to the **Monge formulation**,

$$\inf_{T_{\#}\mu=\nu} \sum_{i=1}^n \mu_i c(x_i, T(x_i)). \quad (1.1)$$

🔥 Caution

The Monge formulation is often ill-posed!

A way to relax from the Monge formulation (and avoid the ill-posedness) is to allow for *mass-splitting*, which leads to the **Kantorovich** (standard) formulation, see Figure 1.4.

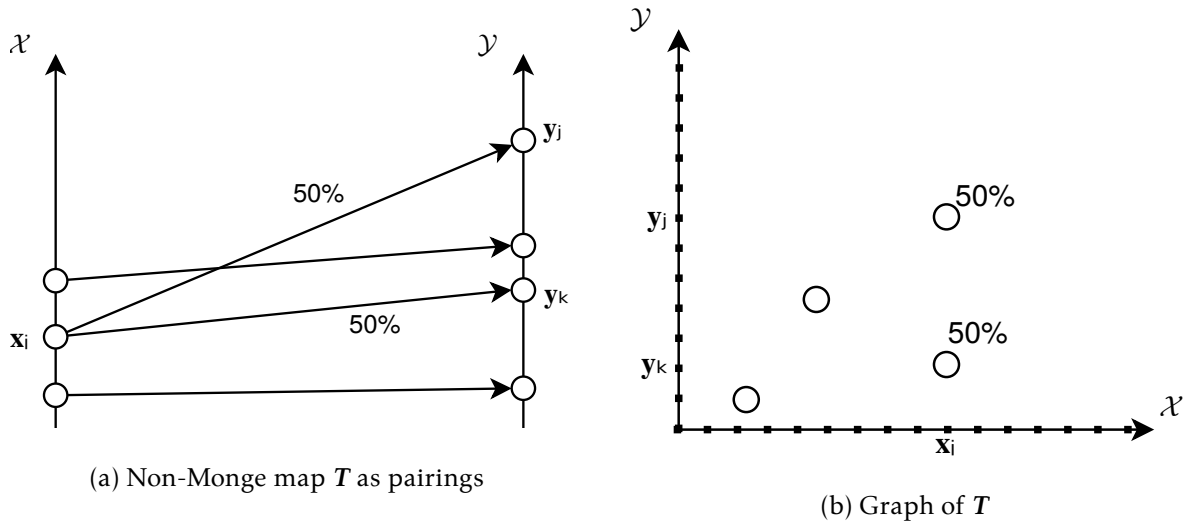


Figure 1.4: Visualizations of a non-Monge transport map T

Indeed, consider the following Kantorovich optimal transport problem,

$$\inf_{\pi} \sum_{i,j=1}^{n,m} \pi_{i,j} c(x_i, y_j), \quad \text{s.t.} \quad \begin{cases} \pi_{i,j} \geq 0, \\ \sum_{j=1}^m \pi_{i,j} = \mu_i, \\ \sum_{i=1}^n \pi_{i,j} = \nu_j. \end{cases} \quad (1.2)$$

Note that the set of constraints is convex, compact and non-empty (e.g., $\pi_{i,j} = \mu_i \cdot \nu_j$ satisfy the constraints). Thus, we have a standard linear programming problem! Moreover, one can make the following observations:

1. The cost is linear in π , hence there exists an optimizer¹;

¹Hence the cost is also continuous over a compact convex feasible set, thus we have existence.

2. *Strong duality*: (1.2) = (1.3), where

$$\sup_{(f_i), (g_j)} \sum_{i=1}^n \mu_i f_i + \sum_{j=1}^m \nu_j g_j \quad \text{s.t.} \quad f_i + g_j \leq c_{ij} = c(\mathbf{x}_i, \mathbf{y}_j) \quad (1.3)$$

and f_i, g_j are Lagrange multipliers of the marginal constraints. Assume (f, g) and π are admissible, then also *weak-duality* holds

$$\sum_{i=1}^n f_i \mu_i + \sum_{j=1}^m g_j \nu_j = \sum_{i,j=1}^{n,m} \pi_{i,j} \underbrace{(f_i + g_j)}_{\leq c_{ij}} \leq \sum_{i,j=1}^{n,m} \pi_{i,j} c_{i,j};$$

3. *Complementary slackness*: (f, g) and π are optimal if and only if $\pi_{i,j} > 0 \implies f_i + g_j = c_{ij}$;
 4. *c-cyclical monotonicity*: Let π be optimal, $\Gamma = \text{supp } \pi := \{(\mathbf{x}_i, \mathbf{y}_j) \in \mathcal{X} \times \mathcal{Y} \mid \pi_{i,j} > 0\}$. Then taking $(\hat{\mathbf{x}}_1, \hat{\mathbf{y}}_1), \dots, (\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k) \in \Gamma$ also gives

$$\sum_{i=1}^k c(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i) \leq \sum_{i=1}^k c(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_{i+1}) \quad (1.4)$$

with $\hat{\mathbf{y}}_{k+1} = \hat{\mathbf{y}}_1$. In fact, π is optimal if and only if it is c-cyclically monotone, i.e., (1.4) holds. Let us show “ $\implies$ ”. Indeed, for π optimal we also have f, g optimal for (1.3) and (3.) yields equality $f(\hat{\mathbf{x}}_i) + g(\hat{\mathbf{y}}_i) = c(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)$ on the support. Thus summing over i gives us

$$\begin{aligned} \sum_{i=1}^k c(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i) &= \sum_{i=1}^k f(\hat{\mathbf{x}}_i) + g(\hat{\mathbf{y}}_i) \\ &= \sum_{i=1}^k f(\hat{\mathbf{x}}_i) + g(\hat{\mathbf{y}}_{i+1}) \\ &\leq \sum_{i=1}^k c(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_{i+1}). \end{aligned}$$

2 Optimal Transport

Let us fix the following notation and convention: Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be Polish 3.1 metric spaces, and

- $\mathcal{B}(\mathcal{X})$ — Borel sets in $\mathcal{X}$, i.e., a σ -algebra generated by open sets;
- $\mathcal{M}(\mathcal{X})$ — finite signed Borel measures on $\mathcal{X}$;
- $\mathcal{P}(\mathcal{X})$ — Borel (probability) measures on $\mathcal{X}$;
- $M_b(\mathcal{X})$ — bounded measurable functions $\mathcal{X} \rightarrow \mathbb{R}$;
- $C_b(\mathcal{X})$ — continuous, bounded measurable functions $\mathcal{X} \rightarrow \mathbb{R}$.

Typically, $\mathbb{R}^\infty := \mathbb{R} \cup \{+\infty\}$, $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and $c: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^\infty$ is a measurable cost function.

2.1 Monge Formulation

Definition 2.1 (Pushforward). Given a Borel map $T: \mathcal{X} \rightarrow \mathcal{Y}$ the *pushforward* measure $T_{\#}\mu \in \mathcal{P}(\mathcal{Y})$ of $\mu \in \mathcal{P}(\mathcal{X})$ through T is given by

$$\forall A \in \mathcal{B}(\mathcal{Y}) : T_{\#}\mu(A) = \mu(T^{-1}(A))$$

or, equivalently, in the language of random variables $X \sim \mu \implies T(X) \sim T_{\#}\mu$.

Definition 2.2 (Transport map). A measurable map $T: \mathcal{X} \rightarrow \mathcal{Y}$ is called a **transport map** from μ to ν if $T_{\#}\mu = \nu$, equivalently $X \sim \mu \implies T(X) \sim \nu$.

Example 2.1. Assume $\mu \in \mathcal{P}(\mathbb{R})$ continuous and $\nu \in \mathcal{P}(\mathbb{R})$. Now $T = F_\nu^{-1} \circ F_\mu$, where F_μ is the cumulative distribution function $\mu((-\infty, x])$ and F_ν^{-1} is the *quantile* function of ν , i.e., $F_\nu^{-1}(u) = \inf \{x \in \mathbb{R} \mid F_\nu(x) > u\}$, is the *monotone transport map* from μ to ν .

Indeed, for μ continuous and $X \sim \mu$ one has $F_\mu(X) \sim \text{Unif}(0, 1)$. Further, $U \sim \text{Unif}(0, 1)$ implies $F_\nu^{-1}(U) \sim \nu$. Hence,

$$X \sim \mu \implies F_\mu(X) \sim \text{Unif}(0, 1) \implies (F_\nu^{-1} \circ F_\mu)(X) = T(X) \sim \nu.$$

Then the first (and naive) way of defining optimal transport problem is again the Monge formulation

$$V_c^{\text{Monge}}(\mu, \nu) := \inf_{T_{\#}\mu = \nu} \int_{\mathcal{X}} c(x, T(x)) d\mu(x). \quad (2.1)$$

However, the set $\{T \mid T_{\#}\mu = \nu\}$ might be empty, non-convex or non-compact.

2.2 Kantorovich Formulation

Let us now fix the further notation for projections:

$$\begin{array}{ccc} \text{pr}^{\mathcal{X}}: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X} & & \text{pr}^{\mathcal{Y}}: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Y}. \\ (x, y) \mapsto x & \& & (x, y) \mapsto y \end{array}$$

Definition 2.3 (Coupling). A probability measure $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is called a *coupling* on μ and ν if $\text{pr}_{\#}^{\mathcal{X}} \pi = \mu$ and $\text{pr}_{\#}^{\mathcal{Y}} \pi = \nu$. Equivalently,

$$(X, Y) \sim \pi \implies \begin{cases} X \sim \mu, \\ Y \sim \nu. \end{cases}$$

A set of all couplings of μ and ν is denoted by $\text{CPL}(\mu, \nu)$.

Example 2.2.

1. Let T be a transport map from μ to ν , then

$$\underbrace{(\text{id}, T)_{\#} \mu}_{\text{Law}(X, T(X)) \text{ if } X \sim \mu} = \pi \in \text{CPL}(\mu, \nu);$$

2. Take $U \sim \text{Unif}(0, 1) = \lambda$, then

$$(F_{\mu}^{-1}, F_{\nu}^{-1})_{\#} \lambda = \text{Law}(F_{\mu}^{-1}(U), F_{\nu}^{-1}(U)) \in \text{CPL}(\mu, \nu)$$

is called a **monotone coupling**;

3. Lastly, consider independent random variables $X \sim \mu$ and $Y \sim \nu$, then the product measure $\mu \otimes \nu = \text{Law}(X, Y) \in \text{CPL}(\mu, \nu)$ is the independent (product) coupling.

This finally leads to the **Kantorovich formulation** of the optimal transport problem,

$$V_c^{\text{OT}}(\mu, \nu) = \inf_{\pi \in \text{CPL}(\mu, \nu)} \int c(x, y) d\pi(x, y). \quad (2.2)$$

Naturally, let our goal now be to study the attainment and properties of V_c^{OT} . In principle, we would like to have a lower semi-continuous functional to be optimizer over a compact set. More precisely, we want

1. $\text{CPL}(\mu, \nu)$ to be non-empty (trivially satisfied), convex¹ and compact (to be shown), and
2. $\pi \mapsto \int c d\pi =: \pi(c)$ to be lower semi-continuous.

¹Convex combinations of probability measures are still probability measures and the marginals are conserved.

Indeed, the above 2 properties yield attainment of V_c^{OT} , as we could pick a minimizing sequence $(\pi^k)_{k \in \mathbb{N}} \in \text{CPL}(\mu, \nu)$ and by compactness extract a converging subsequence with the limit π . Lastly, by lower semi-continuity we get

$$V_c^{\text{OT}}(\mu, \nu) = \lim_k \pi^k(c) \geq \pi(c) \geq V_c^{\text{OT}}(\mu, \nu)$$

for $(\pi^k)_{k \in \mathbb{N}}$ the minimizing sequence. As such, we need a topology on $\mathcal{P}(\mathcal{X})$ (to study compactness).

Definition 2.4. The weak topology of measures on $\mathcal{M}(\mathcal{X})$ is the initial topology² with respect to $\mathcal{M}(\mathcal{X}) \ni \mu \mapsto \mu(f)$ for $f \in C_b(\mathcal{X})$.

We will endow $\mathcal{P}(\mathcal{X}) \subseteq \mathcal{M}(\mathcal{X})$ with this topology.

Remark 2.1. Let us state some (easy to show) facts concerning the weak topology on $\mathcal{P}(\mathcal{X})$:

1. $x \mapsto \delta_x$ is the *topological embedding*³ $\mathcal{X} \hookrightarrow \mathcal{P}(\mathcal{X})$;
2. this topology is *Hausdorff*⁴, i.e., for $\mu, \nu \in \mathcal{M}(\mathcal{X})$ with $\mu(f) = \nu(f) \forall f \in C_b(\mathcal{X})$ implies by [monotone class argument](#) $\mu(f) = \nu(f) \forall f \in \mathcal{M}_b(\mathcal{X})$, which in turn means $\mu = \nu$;
3. it is *countably generated*, i.e., there exists a countable family $(f_n)_{n \in \mathbb{N}} \subseteq C_b(\mathcal{X})$ such that the family $\mu \mapsto \mu(f_n)$ generates the weak topology (which also means that for testing convergence a countable set suffices) — in particular, the weak topology is [second countable](#);
4. it is *completely metrizable*, e.g., by **Prokhorov metric**

$$d_{\text{Prokhorov}}(\mu, \nu) := \inf \left\{ \varepsilon > 0 \mid \forall A \in \mathcal{B}(\mathcal{X}) : \begin{cases} \mu(A) \leq \nu(A_\varepsilon) + \varepsilon, \\ \nu(A) \leq \mu(A_\varepsilon) + \varepsilon \end{cases} \right\} \quad \text{where } A_\varepsilon := \bigcup_{x \in A} B(x, \varepsilon); \quad (2.3)$$

5. lastly, it is a *separable space*, i.e., let $(x_k)_{k \in \mathbb{N}} \subseteq \mathcal{X}$ be dense (such sequence exists since $\mathcal{X}$ is Polish), then

$$\left\{ \sum_{k=1}^n \alpha_k \delta_{x_k} \mid n \in \mathbb{N} \text{ \& } \alpha_k \in \mathbb{Q} \cap [0, 1] \text{ \& } \sum_{k=1}^n \alpha_k = 1 \right\}$$

is countable and even dense in $\mathcal{P}(\mathcal{X})$.

Proposition 2.1. Let $\mathcal{X}$ be a Polish space. Then the space $\mathcal{P}(\mathcal{X})$ of probability measures over $\mathcal{X}$ is also Polish.

Proof. Above we show that $\mathcal{P}(\mathcal{X})$ is completely metrizable separable topological space, i.e., by Definition 3.1 it is Polish. $\square$

²Also known as *induced topology*, as is the topology on the set X with respect to a family of functions on X such that it is the coarsest topology on X that makes those functions *continuous*.

³Namely, every embedding is injective and continuous.

⁴Hausdorff topology is that such distinct points have disjoint neighborhoods — that is why one goes to the indicators of such neighborhoods being in $\mathcal{M}_b(\mathcal{X})$ by a monotone class argument.

Corollary 2.1. For $\mathcal{X}$ and $\mathcal{Y}$ Polish spaces, both $\mathcal{X} \times \mathcal{Y}$ and the space $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ of probability measures over their product are also Polish.

Proof. Easy but omitted. □

Remark 2.2. Let $\mathcal{X}$ be a compact metric space, then $\mathcal{P}(\mathcal{X})$ is compact.

Proof. By the Riesz-Markov 3.9 representation one has $(C(\mathcal{X}), \|\cdot\|_\infty)' \cong (\mathcal{M}(\mathcal{X}), \text{TV})$, where $\text{TV}(\mu) = \sup_{f \in C(\mathcal{X})} \mu(f)$.
 $\|f\|_\infty \leq 1$

Further, denote $\mathcal{M}_{\leq 1}(\mathcal{X}) := \{\mu \mid \text{TV}(\mu) \leq 1\}$ a closed unit ball in the dual of $C(\mathcal{X})$. By Banach-Alaoglu 3.10 $\mathcal{M}_{\leq 1}(\mathcal{X})$ is compact in the **weak* topology** (i.e., in the weak convergence 3.5 of measures). Now, taking 1 to be the constant-1 function on the compact space $\mathcal{X}$,

$$\mathcal{M}_{\leq 1}(\mathcal{X}) \supseteq \mathcal{P}(\mathcal{X}) = \bigcap_{\substack{f \in C(\mathcal{X}) \\ f \geq 0}} \{\mu \mid \mu(f) \geq 0\} \cap \{\mu \mid \mu(1) = 1\},$$

where the first set can be replaced, by monotone class argument, by the set of all non-negative measures. Now both are closed sets with respect to the weak* topology. Hence $\mathcal{P}(\mathcal{X})$ is weak* closed subset of a weak* compact set, thus itself is weak* compact in $(\mathcal{M}(\mathcal{X}), C(\mathcal{X}))$. As $\mathcal{X}$ is compact, $C(\mathcal{X})$ coincides with $C_b(\mathcal{X})$ which gives us compactness in the sense of Definition 2.4. □

Definition 2.5 (Tightness). A family $\mathcal{E} \subseteq \mathcal{P}(\mathcal{X})$ is called **tight**, if for all $\varepsilon > 0$ there exists $K \subseteq \mathcal{X}$ such that

$$\sup_{\mu \in \mathcal{E}} \mu(K^c) \leq \varepsilon \quad \text{or, equivalently,} \quad \inf_{\mu \in \mathcal{E}} \mu(K) \geq 1 - \varepsilon.$$

Theorem 2.1 (Prokhorov's). A family $\mathcal{E} \subseteq \mathcal{P}(\mathcal{X})$ is tight if and only if $\mathcal{E}$ is relatively compact⁵.

A sketch of proof. “ $\implies$ ”: The more difficult direction, but loosely speaking one uses tightness to restrict themselves to large compacts, then employ Remark 2.2 and use a diagonalization argument. For more information refer to [2].

“ $\impliedby$ ”: Choose $\varepsilon > 0$ arbitrary. Our aim is now to construct a compact $K \subseteq \mathcal{X}$ such that $\inf_{\mu \in \mathcal{E}} \mu(K) \geq 1 - \varepsilon$. As $\mathcal{X}$ is separable (by being Polish), there exists a countable family $(x_k)_{k \in \mathbb{N}} \subseteq \mathcal{X}$ that is dense in $\mathcal{X}$. Fix $n \in \mathbb{N}$, then by relative compactness of $\mathcal{E}$, the closure

$$\bar{\mathcal{E}} \subseteq \bigcup_{N \in \mathbb{N}} \underbrace{\left\{ \mu \mid \mu \left(\bigcup_{i=1}^N B \left(x_i, \frac{1}{n} \right) \right) > 1 - \frac{\varepsilon}{2^n} \right\}}_{\text{open}}$$

is compact, and using this with continuity of measures gives existence of $N_n \in \mathbb{N}$ such that $\inf_{\mu \in \mathcal{E}} \mu(K_n) \geq 1 - \frac{\varepsilon}{2^n}$, where $K_n := \bigcup_{i=1}^{N_n} B(x_i, \frac{1}{n})$ (not necessarily compact at the moment). Define

⁵A set X is relatively compact when $\bar{X}$ is compact.

$K := \bigcap_{n \in \mathbb{N}} \overline{K_n}$, which is closed and totally bounded (i.e., we can cover it with finitely many balls of any given radius). Completeness of $\mathcal{X}$ further yields K is compact, hence $\inf_{\mu \in \mathcal{E}} \mu(K) = 1 - \sup_{\mu \in \mathcal{E}} \mu(K^c)$ where $K^c = \bigcup_{n \in \mathbb{N}} \overline{K_n}^c \subseteq \bigcup_{n \in \mathbb{N}} K_n^c$. Lastly, by subadditivity of measures

$$\inf_{\mu \in \mathcal{E}} \mu(K) \geq 1 - \sup_{\mu \in \mathcal{E}} \sum_{n \in \mathbb{N}} \underbrace{\mu(K_n^c)}_{\leq \frac{\varepsilon}{2^n}} \geq 1 - \varepsilon.$$

□

Interestingly, we can make the following observations:

- The direction *tightness* $\implies$ *relative compactness* holds even if $\mathcal{X}$ is just a metric space.
- On the other hand, “ $\impliedby$ ” requires both separability and completeness, i.e., $\mathcal{X}$ to be Polish.

Corollary 2.2. *The set of all couplings $C_{PL}(\mu, \nu)$ is compact.*

Proof. First, we prove that $C_{PL}(\mu, \nu)$ is a closed set. For this recall that $\text{pr}^{\mathcal{X}}$ is a continuous function. Then its pushforward $\text{pr}_{\#}^{\mathcal{X}}: \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \rightarrow \mathcal{P}(\mathcal{X})$ defined as $\pi \mapsto \text{pr}_{\#}^{\mathcal{X}} \pi$ is also continuous since for any $f \in C_b(\mathcal{X})$ we have

$$\int f \, d\text{pr}_{\#}^{\mathcal{X}} \pi = \int \underbrace{(f \circ \text{pr}^{\mathcal{X}})}_{\in C_b(\mathcal{X} \times \mathcal{Y})}(x, y) \, d\pi(x, y).$$

Now, we can write

$$C_{PL}(\mu, \nu) = \left\{ \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \text{pr}_{\#}^{\mathcal{X}} \pi = \mu \right\} \cap \left\{ \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \text{pr}_{\#}^{\mathcal{Y}} \pi = \nu \right\},$$

where both of the “partial” sets are closed by continuity of $\text{pr}^{\mathcal{X}}$ and $\text{pr}^{\mathcal{Y}}$, respectively. Thus, set $C_{PL}(\mu, \nu)$ is also closed.

Second, we show that $C_{PL}(\mu, \nu)$ is a tight set. Again, choose $\varepsilon > 0$ arbitrarily. By Theorem 2.1 on (trivially compact) $\mathcal{E} = \{\mu\}$ and $\{\nu\}$, respectively, there exist compact subsets $K_{\mathcal{X}} \stackrel{c}{\subseteq} \mathcal{X}$ and $K_{\mathcal{Y}} \stackrel{c}{\subseteq} \mathcal{Y}$, see Figure 2.1, such that

$$\mu(K_{\mathcal{X}}) \geq 1 - \frac{\varepsilon}{2} \quad \& \quad \nu(K_{\mathcal{Y}}) \geq 1 - \frac{\varepsilon}{2}$$

and that $K := K_{\mathcal{X}} \times K_{\mathcal{Y}} \stackrel{c}{\subseteq} \mathcal{X} \times \mathcal{Y}$. Then for $\pi \in C_{PL}(\mu, \nu)$,

$$\begin{aligned} \pi(K^c) &= \pi(K_{\mathcal{X}}^c \times \mathcal{Y} \cup \mathcal{X} \times K_{\mathcal{Y}}^c) \\ &\stackrel{\text{subadditivity}}{\leq} \underbrace{\pi(K_{\mathcal{X}}^c \times \mathcal{Y})}_{\text{pr}_{\mathcal{X}}^{-1}(K_{\mathcal{X}}^c)} + \pi(\mathcal{X} \times K_{\mathcal{Y}}^c) = \underbrace{\mu(K_{\mathcal{X}}^c)}_{\leq \frac{\varepsilon}{2}} + \underbrace{\nu(K_{\mathcal{Y}}^c)}_{\leq \frac{\varepsilon}{2}} \leq \varepsilon. \end{aligned}$$

Hence $C_{PL}(\mu, \nu)$ is tight.

Finally, by Prokhorov’s theorem 2.1 we get relative compactness of $C_{PL}(\mu, \nu)$, which yields compactness (under the weak* topology) by $C_{PL}(\mu, \nu)$ being closed. □

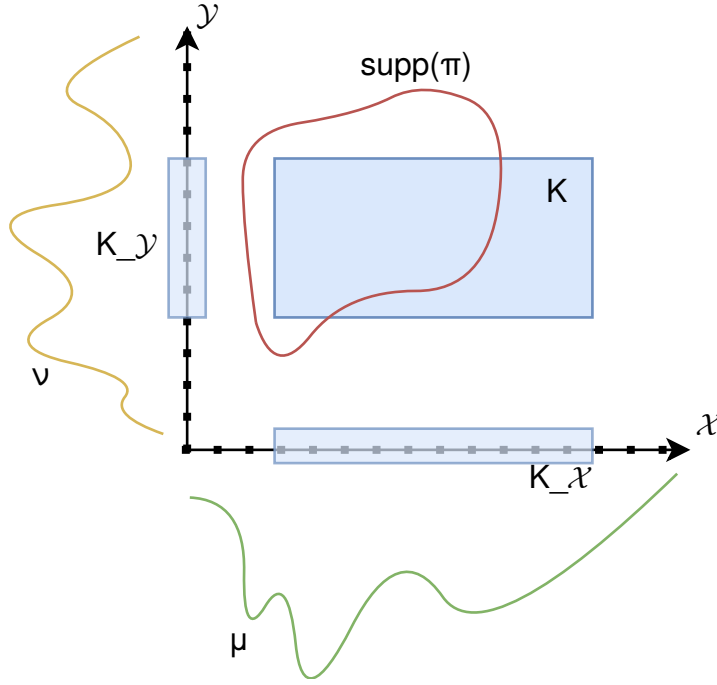


Figure 2.1: Role of K^X and K^Y

Remark 2.3. If $c \in C_b(X \times Y)$, then by the Definition 2.4 of weak topology/convergence (see also Definition 3.5), the map $\mathcal{P}(X \times Y) \rightarrow \mathbb{R}$ defined as $\pi \mapsto \pi(c)$ is continuous.

As a reminder, let us recall the definition of lower semi-continuity.

Definition 2.6 (Lower semi-continuity). A function $f: X \rightarrow \mathbb{R}^\infty$ is called **lower semi-continuous** at a point $x_0 \in X$ if $\liminf_{x \rightarrow x_0} f(x) \geq f(x_0)$.

Equivalently, f is lower semi-continuous at x_0 if for every $\alpha < f(x_0)$ there exists neighborhood $O(x_0)$ of x_0 such that $f(x) > \alpha$ for all $x \in O(x_0)$.

Lemma 2.1. Let $f: X \rightarrow \mathbb{R}^\infty$ be lower bounded, i.e., $\exists C \in \mathbb{R}$ so that $f \geq C$, then f is lower semi-continuous if and only if there exists a family $(f_k)_{k \in \mathbb{N}} \subset C_b(X)$ such that $f_k \uparrow f$.

Moreover, (f_k) can be chosen k -Lipschitz.

Proof. “ $\Leftarrow$ ”: Take a sequence $x_n \rightarrow x$. Then

$$f(x) = \sup_{k \in \mathbb{N}} f_k(x) = \sup_{k \in \mathbb{N}} \lim_{n \rightarrow \infty} f_k(x_n) \leq \liminf_{n \rightarrow \infty} \sup_{k \in \mathbb{N}} f_k(x_n) = \liminf_{n \rightarrow \infty} f(x_n).$$

“ $\implies$ ”: Let us define $f_k(x) = \inf_{y \in \mathcal{X}} f(y) + kd_{\mathcal{X}}(x, y)$, then $C \leq f_k \leq f_{k+1} \leq f$. Moreover, f_k are k -Lipschitz (thus continuous) as for any $\tilde{y} \in \mathcal{X}$

$$\begin{aligned}
f_k(x) &\leq f(\tilde{y}) + kd_{\mathcal{X}}(x, \tilde{y}) \stackrel{\text{triangle}}{\leq} f(\tilde{y}) + k(d_{\mathcal{X}}(x, \tilde{x}) + d_{\mathcal{X}}(\tilde{x}, \tilde{y})) \\
&\Downarrow \\
f_k(x) &\leq kd_{\mathcal{X}}(x, \tilde{x}) + \inf_{\tilde{y} \in \mathcal{X}} (f(\tilde{y}) + kd_{\mathcal{X}}(\tilde{x}, \tilde{y})) \\
&= kd_{\mathcal{X}}(x, \tilde{x}) + f_k(\tilde{x}) \\
&\Downarrow \\
f_k(x) - f_k(\tilde{x}) &\leq kd_{\mathcal{X}}(x, \tilde{x}) \stackrel{\text{symmetry in } x, \tilde{x}}{\implies} |f_k(x) - f_k(\tilde{x})| \leq kd_{\mathcal{X}}(x, \tilde{x}).
\end{aligned}$$

Finally, by lower semi-continuity we get point-wise convergence. Indeed, for every $\alpha < f(x_0)$ we have a ball neighborhood $B(x_0, \delta) = U$ such that for $y \in U$ we immediately get $f(y) + kd_{\mathcal{X}}(x_0, y) > \alpha$. On the other hand, for $y \notin U$, we can see

$$\underbrace{f(y)}_{\geq C} + \underbrace{kd_{\mathcal{X}}(x_0, y)}_{\geq \delta > 0} \geq C + k\delta \geq \alpha$$

for large enough values of k . In total, $f_k(x_0) \geq \alpha$ for large values of k . As α was chosen arbitrarily, we can combine it with the (trivial) inequality $f_k \leq f$ to get the desired point-wise limit. Finally, one can take $g_k := \min\{f_k, k\} \in C_b(\mathcal{X})$ which also converge monotonically to f from below and are k -Lipschitz, as $|g_k(x) - g_k(\tilde{x})|$ is either 0 (both are truncated), $|f_k(x) - f_k(\tilde{x})|$ (see above) or, without loss of generality assume $f_k(x) > f_k(\tilde{x})$ by symmetry,

$$|g_k(x) - g_k(\tilde{x})| = |k - f_k(\tilde{x})| \leq |f_k(x) - f_k(\tilde{x})| \leq kd_{\mathcal{X}}(x, \tilde{x}).$$

□

Corollary 2.3. *If $c: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^\infty$ is lower semi-continuous and lower bounded, then $\pi \mapsto \pi(c)$ is also lower semi-continuous for $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.*

Proof. By Lemma 2.1, there exists $(c_k)_{k \in \mathbb{N}} \subseteq C_b(\mathcal{X} \times \mathcal{Y})$ such that $c_k \uparrow c$. By Remark 2.3 $\pi \mapsto \pi(c_k)$ is continuous. Lastly, by Theorem 3.1 (possibly using offset by the lower bound of c), we get $\pi(c) = \sup_{k \in \mathbb{N}} \pi(c_k)$, and that $\pi \mapsto \pi(c)$ lower semi-continuous as for sequence $\pi_j \rightarrow \pi$ it holds

$$\pi(c) = \sup_{k \in \mathbb{N}} \pi(c_k) = \sup_{k \in \mathbb{N}} \lim_{j \rightarrow \infty} \pi_j(c_k) \leq \liminf_{j \rightarrow \infty} \sup_{k \in \mathbb{N}} \pi_j(c_k) = \liminf_{j \rightarrow \infty} \pi_j(c).$$

□

Proposition 2.2. *The Kantorovich optimal transport cost $V_c^{\text{OT}}: \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \rightarrow \mathbb{R}^\infty$, where c is lower semi-continuous and lower bounded, is lower semi-continuous and convex.*

Proof. Problem 3 of Exercise sheet 1.

□

💡 Tip

Our goal will now be to find the **dual formulation** of V_c^{OT} .

Remark 2.4. Assume $\mathcal{X}, \mathcal{Y}$ are compact. Using

$$C_{\text{PL}}(\mu, \nu) = \bigcap_{(f,g) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y})} \{ \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \pi(f \oplus g) = \mu(f) + \nu(g) \},$$

where $(f \oplus g)(x, y) := f(x) + g(y)$ we get

$$V_c^{\text{OT}}(\mu, \nu) = \inf_{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \sup_{(f,g) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y})} \pi(c - f \oplus g) + \mu(f) + \nu(g).$$

Intuitively, the supremum finds the “most significant violation” of the intersection used to express $C_{\text{PL}}(\mu, \nu)$ above.

By the [minimax theorem](#) since we are taking inf over a compact set and sup over a vector space and

- $\pi \mapsto \pi(c - f \oplus g)$ is lower semi-continuous and linear,
- $(f, g) \mapsto \pi(c - f \oplus g) + \mu(f) + \nu(g)$ is linear,

we get

$$V_c^{\text{OT}} = \sup_{(f,g)} \inf_{\pi} \pi(c - f \oplus g) + \mu(f) + \nu(g).$$

Setting $\pi = \delta_{(x,y)}$ yields

$$\inf_{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \pi(c - f \oplus g) \leq \inf_{\pi = \delta_{(x,y)}} \pi(c - f \oplus g) = \inf_{(x,y) \in \mathcal{X} \times \mathcal{Y}} (c(x, y) - f(x) - g(y)) =: \alpha(f, g)$$

and together with $\pi(c - f \oplus g) \geq \inf_{(x,y)} c - f \oplus g$ for every π (and diracs are admissible), we get even an equality. Thus $f(x) + g(y) + \alpha(f, g) \leq c(x, y)$ and note that $\alpha(f, g)$ is finite as f, g are bounded and c is lower-bounded. Replacing f with $\tilde{f} := f + \alpha(f, g)$ (so $\tilde{f} \oplus g \leq c$) finally produces the *duality* result:

$$V_c^{\text{OT}}(\mu, \nu) = \inf_{\pi \in C_{\text{PL}}(\mu, \nu)} \pi(c) = \sup_{\substack{(f,g) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \\ f \oplus g \leq c}} \mu(f) + \nu(g). \quad (2.4)$$

Note that we call f, g satisfying $f \oplus g \leq c$ **admissible**.

💡 Tip

Can we extend the duality result (2.4) for compact $\mathcal{X}, \mathcal{Y}$ also to other cases?

2.2.1 Excursion to Convex Analysis

Let us make the following assumptions on the setting we shall be working in:

1. $\mathcal{X}$ is **locally convex topological vector space**;
2. $\mathcal{X}'$ is the continuous dual space with weak* topology;
3. $F: \mathcal{X} \rightarrow \mathbb{R}^\infty$ is a given **proper** function.

Definition 2.7 (Convex conjugate). The *convex conjugate* of F is said to be the function $F^*: \mathcal{X}' \rightarrow \mathbb{R}^\infty$ such that

$$F^*(y) = \sup_{x \in \mathcal{X}} y(x) - F(x).$$

Theorem 2.2 (Fenchel-Moreau duality). *Let F be proper, lower semi-continuous and convex. Then*

$$F(x) = \sup_{y \in \mathcal{X}'} y(x) - F^*(y).$$

A sketch of proof.

See also Problem 4 of the Exercise sheet 1.

One always has $F^*(y) \geq y(x) - F(x)$ simply by the Definition 2.7, i.e., $F(x) \geq y(x) - F^*(y)$. For the other direction, take $(\tilde{y}, \tilde{c}) \in (\mathcal{X} \times \mathbb{R})' = \mathcal{X}' \times \mathbb{R}$ such that $\tilde{y} - \tilde{c} \leq F$. Then $F^*(\tilde{y}) \leq \tilde{c}$ and set

$$R(x) := \sup_{y \in \mathcal{X}'} y(x) - F^*(y) \implies R(x) \geq \tilde{y}(x) - \tilde{c}.$$

Put differently, $R(x)$ is greater than all affine minorants a of F , see Figure 2.2.

Further, $\text{epi } F = \{(x, y) \mid F(x) \leq y\}$ is closed, convex and non-empty since F is proper, lower semi-continuous and convex. By applying geometric Hahn-Banach theorem 3.8 we can separate $\text{epi } F$ and $(\tilde{x}, \tilde{c}) \notin \text{epi } F$ (corresponding to any affine minorant of F), which gives us the desired inequality. $\square$

2.2.2 Disintegration of Measures

Definition 2.8 (Probability Kernel). The mapping $K: \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ is a (*probability*) *kernel* if it is measurable with respect to $\mathcal{B}(\mathcal{X})$ and the Borel σ -algebra on $\mathcal{P}(\mathcal{Y})$.

Proposition 2.3. *The following are equivalent:*

1. K is measurable,
2. $\forall f \in C_b(\mathcal{Y})$ the mapping⁶ $x \mapsto \int f(y)K(x; dy)$ is measurable, and
3. $\forall f \in M_b(\mathcal{X} \times \mathcal{Y})$ the mapping $x \mapsto \int f(x, y)K(x; dy)$ is measurable.

⁶Here $K(x; dy)$ designates that x determines the measure from $\mathcal{P}(\mathcal{Y})$ used for integration in the variable y .

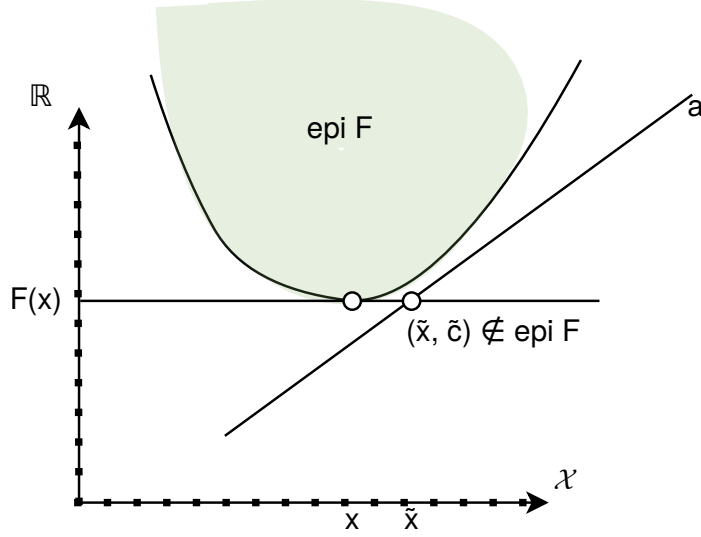


Figure 2.2: Affine minorants of $F(x)$

Proof. Firstly, the weak topology on $\mathcal{P}(Y)$ is generated by countable subsets, see Remark 2.1 (3, and possibly 5),

$$\mathcal{C} := \{\mu \mid \mu(f_k) < \alpha\} \mid \alpha \in \mathbb{Q}, k \in \mathbb{N}\}$$

with $(f_k)_{k \in \mathbb{N}} \subset C_b(Y)$. Therefore, $\sigma(\mathcal{C}) = \mathcal{B}(\mathcal{P}(Y))$.

Then K is measurable

$$\begin{aligned} &\iff \forall C \in \mathcal{C} : K^{-1}(C) \in \mathcal{B}(X) \\ &\iff \forall (k, q) \in \mathbb{N} \times \mathbb{Q} : \left\{ x \in X \mid \int f_k(y) K(x; dy) < q \right\} \in \mathcal{B}(X) \\ &\iff \forall (k, q) \in \mathbb{N} \times \mathbb{Q} : x \mapsto \int f_k(y) K(x; dy) \text{ is measurable.} \end{aligned}$$

The last remaining equivalence (2) $\iff$ (3) follows by monotone class argument. $\square$

Definition 2.9 (Gluing). Let $K: X \rightarrow \mathcal{P}(Y)$ be a kernel and $\mu \in \mathcal{P}(X)$. The **gluing** $\mu \otimes K \in \mathcal{P}(X \times Y)$ of μ and K is defined by

$$\forall f \in M_b(X \times Y) : \int f d\mu \otimes K := \iint f(x, y) K(x; dy) \mu(dx).$$

Note that the right-hand side of the gluing definition is well-defined by Proposition 2.3 (3).

Theorem 2.3 (Disintegration). Let $\pi \in \mathcal{P}(X \times Y)$, then there exists a probability kernel $K: X \rightarrow \mathcal{P}(Y)$ such that $\pi = \mu \otimes K$ with $\mu := \text{pr}_X^\# \pi$. Secondly, let $\tilde{K}: X \rightarrow \mathcal{P}(Y)$ be another disintegration kernel, i.e., it satisfies the above equality. Then $\tilde{K} = K$ μ -almost everywhere.

Proof. Omitted. □

Note

The disintegration of π with respect to μ is denoted $(\pi_x)_{x \in \mathcal{X}}$.

Example 2.3.

1. Let $T : \mathcal{X} \rightarrow \mathcal{Y}$ be measurable and $X \sim \mu$, then

$$\pi = \text{Law}(X, T(X)) = (\text{id}, T)_\# \mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \implies \pi_x = \delta_{T(x)}.$$

2. Let $(X, Y) \sim \pi$, then $\pi_x = \text{Law}(Y \mid X)$ a.s.
3. Consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $\pi(dx, dy) = f(x, y)dx dy$, then

$$\pi_x(dy) = \frac{f(x, y)}{\int_{\mathbb{R}^d} f(x, y') dy'} dy \quad \text{for } \mu\text{-a.e. } x.$$

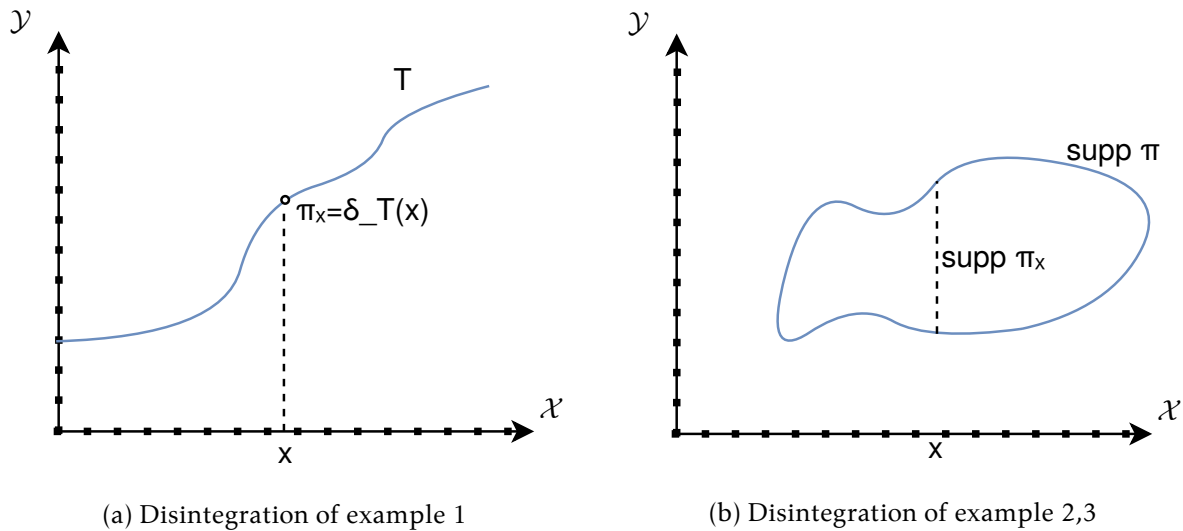


Figure 2.3: Visualizations of disintegrations of various measures.

Theorem 2.4 (Duality). Consider the usual setting of optimal transport (see beginning of this chapter). Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^\infty$ be lower semi-continuous and lower bounded cost function, then

$$V_c^{\text{OT}} = \sup_{\substack{(f,g) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \\ f \oplus g \leq c}} \mu(f) + \nu(g), \quad (2.5)$$

$$= \sup_{\substack{(f,g) \in \mathcal{L}^1(\mu) \times \mathcal{L}^1(\nu) \\ f \oplus g \leq c}} \mu(f) + \nu(g). \quad (2.6)$$

Remark 2.5.

1. *Weak duality:* let $(f, g) \in \mathcal{L}^1(\mu) \times \mathcal{L}^1(\nu)$ such that $f \oplus g \leq c$, and $\pi \in \text{CPL}(\mu, \nu)$, then

$$\mu(f) + \nu(g) = \pi(f \oplus g) \leq \pi(c),$$

which implies $V_c^{\text{OT}} \geq (2.6)$.

2. (2.6) holds even when $c \in \mathcal{M}_b(\mathcal{X} \times \mathcal{Y})$ (via Choquet capacitability theory).
3. When c is even Lipschitz, one can restrict (f, g) in (2.5) to be Lipschitz as well.
4. For $(f, g) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})$ admissible, the c -transforms

$$\begin{aligned} f^{\bar{c}}(y) &:= \inf_{x \in \mathcal{X}} (c(x, y) - f(x)) \geq g(y), \\ g^c(x) &:= \inf_{y \in \mathcal{Y}} (c(x, y) - g(y)) \geq f(x) \end{aligned} \tag{2.7}$$

satisfy $f \oplus g \leq f \oplus f^{\bar{c}} \leq c$, so replacing $f \oplus g$ by $f \oplus f^{\bar{c}}$ improves the value in (2.6).

We shall now investigate the c -transform a bit more.

Definition 2.10 (c -transform and c -concavity). For general functions f, g we call $f^{\bar{c}}, g^c$ (given by (2.7), respectively) the c -, resp. $\bar{c}$ -**transform** of f, g , if well-defined.

Further, $f: \mathcal{X} \rightarrow \mathbb{R}^\infty$ is called c -**concave** if $\exists g$ such that $f = g^c$ and $f \not\equiv -\infty$.

Recall that (f, g) are *admissible*, i.e., $f \oplus g \leq c$, then the pair $((f^{\bar{c}})^c, f^{\bar{c}})$ yields a better value, i.e.,

$$\mu(f) + \nu(g) \leq \mu((f^{\bar{c}})^c) + \nu(f^{\bar{c}})$$

if everything is measurable.

Lemma 2.2. Let $f: \mathcal{X} \rightarrow \mathbb{R}^\infty$ be proper, i.e., $f \not\equiv -\infty$. Then f is c -concave if and only if $(f^{\bar{c}})^c = f$.

Proof. See Problem 5 of Exercise sheet 1. □

Example 2.4. For simplicity assume $\mathcal{X} = \mathcal{Y}$.

1. Consider $c(x, y) = d_{\mathcal{X}}(x, y)$, which would be the Wasserstein distance in the dual, then f is c -concave $\iff f$ is 1-Lipschitz.
2. Consider $c(x, y) = \mathbb{I}_{x \neq y}$, which would be the total variation distance in the dual, then f is c -concave $\iff \sup_{x, y} f(x) - f(y) \leq 1$.

3. Consider now $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|^2$, which will lead to the Wasserstein-2 distance, then f is c-concave $\iff$ there exists g so that

$$\begin{aligned} f(\mathbf{x}) &\triangleq \inf_{\mathbf{y} \in \mathbb{R}^d} \overbrace{|\mathbf{x} - \mathbf{y}|^2}^{x^2 - 2\mathbf{x}\mathbf{y} + \mathbf{y}^2} - g(\mathbf{y}) \\ &= x^2 - \sup_{\mathbf{y} \in \mathbb{R}^d} 2\mathbf{x}\mathbf{y} - (\mathbf{y}^2 - g(\mathbf{y})) = x^2 - (|\cdot|^2 - g)^*(2\mathbf{x}), \end{aligned}$$

hence by Theorem 2.2 it holds if and only if $|\cdot|^2 - g$ is proper lower semi-continuous convex function.

Lemma 2.3. Let $h \in C(\mathcal{X} \times \mathcal{Y})$ be bounded from below. Then, $\forall \varepsilon > 0$ there exists a **selector** $S: \mathcal{X} \rightarrow \mathcal{Y}$ such that

$$h(x, S(x)) \leq \inf_{y \in \mathcal{Y}} h(x, y) + \varepsilon.$$

Proof. Consider $\tilde{h}(x) = \inf_{y \in \mathcal{Y}} h(x, y) > -\infty$ as h is lower bounded. Now $\tilde{h}$ is upper semi-continuous. Let $(y_n)_{n \in \mathbb{N}} \subseteq \mathcal{Y}$ be dense (since $\mathcal{Y}$ is Polish), and (by upper semi-continuity)

$$A_n := \left\{ x \in \mathcal{X} \mid h(x, y_n) \leq \underbrace{\tilde{h}(x)}_{\text{finite}} + \varepsilon \right\} \in \mathcal{B}(\mathcal{X}) \implies \bigcup_{n \in \mathbb{N}} A_n = \mathcal{X}.$$

This is due to h being continuous and $\tilde{h} > -\infty$. Setting $n(x) := \min \{n \in \mathbb{N} \mid x \in A_n\}$ and $S(x) = y_{n(x)}$ gives us the selector. $\square$

At long ado, we shall prove Theorem 2.4.

Proof. Consider the setting of Fenchel-Moreau theorem 2.2, i.e., $\mathcal{X} = \mathcal{M}(\mathcal{Y})$ endowed with the weak convergence of measures, and $\mathcal{X}' = C_b(\mathcal{Y})$ (which give the weak topology). Let

$$F(\nu) := \begin{cases} V_c^{\text{OT}}(\mu, \nu), & \nu \in \mathcal{P}(\mathcal{Y}), \\ \infty, & \text{otherwise.} \end{cases}$$

Observe that as $\nu \mapsto V_c^{\text{OT}}(\mu, \nu)$ is lower semi-continuous and convex together with $\mathcal{P}(\mathcal{Y}) \subseteq \mathcal{M}(\mathcal{Y})$ being a closed, convex subset yields that F is lower semi-continuous and convex. Further, if F is proper, i.e., $\exists \nu \in \mathcal{P}(\mathcal{Y})$ such that $V_c^{\text{OT}}(\mu, \nu) < \infty$, then by Theorem 2.2

$$V_c^{\text{OT}}(\mu, \nu) = \sup_{g \in C_b(\mathcal{Y})} \nu(g) - F^*(g).$$

Let us now prove the result of result in 2 steps:

1. *step*: Assume, for now, that c is Lipschitz and bounded. Importantly, then c -transforms will also be Lipschitz, and recall that in Lemma 2.1 we have shown we can approximate a lower semi-continuous, lower-bounded function by Lipschitz functions (and we shall use that here).

Let $g \in C_b(\mathcal{Y})$, then (notice that kernel induces a coupling, and vice versa)

$$\begin{aligned}
F^*(g) &= \sup_{\nu \in \mathcal{P}(\mathcal{Y})} \nu(g) - \overbrace{V_c^{\text{OT}}(\mu, \nu)}^{=\inf_{\pi \in \text{CPL}(\mu, \nu)} \pi(c)} \\
&= \sup_{\nu \in \mathcal{P}(\mathcal{Y})} \sup_{\pi \in \text{CPL}(\mu, \nu)} \underbrace{\int g(y) - c(x, y) \, d\pi(x, y)}_{=\iint g(y) - c(x, y) \, d\pi_x(y) \mu(dx)} \\
&= \sup_{K: \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y}) \text{ msb.}} \iint g - c_x \, dK_x \, d\mu(x) \quad (\text{SupKer}) \\
&\stackrel{!}{=} \int \sup_{\nu \in \mathcal{P}(\mathcal{Y})} \left\{ \int g - c_x \, d\nu \right\} \, d\mu(x) \\
&\stackrel{\text{Diracs}}{=} \int \sup_{y \in \mathcal{Y}} (g(y) - c(x, y)) \, d\mu(x) \\
&= \int -\inf_{y \in \mathcal{Y}} (c(x, y) - g(y)) \, d\mu(x) = -\mu(g^c),
\end{aligned}$$

where $\mu(g^c)$ is Lipschitz as c is assumed to be Lipschitz. However, the question remains whether $\stackrel{!}{=}$ holds. Indeed, $\leq$ is satisfied trivially. Fortunately, by Lemma 2.3 there exists a selector $S: \mathcal{X} \rightarrow \mathcal{Y}$ measurable such that

$$g(S(x)) - c(x, S(x)) + \varepsilon \geq \sup_{y \in \mathcal{Y}} g(y) - c(x, y).$$

Hence we can take $K^S: \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ with $x \mapsto \delta_{S(x)}$ such that it is measurable. Further,

$$(\text{SupKer}) \geq \iint g - c_x K_x^S \mu(dx) \geq -\varepsilon + \int \left(\sup_{y \in \mathcal{Y}} g(y) - c(x, y) \right) \mu(dx)$$

and since $\varepsilon > 0$ was arbitrary, we get the asserted equality.

2. *step*: Assume now c to only be lower bounded and lower semi-continuous. By Lemma 2.1, there exists a sequence $(c_k)_{k \in \mathbb{N}} \subseteq C(\mathcal{X} \times \mathcal{Y})$ with $c_k \uparrow c$ such that c_k is k -Lipschitz and bounded⁷. Then, by monotone convergence 3.1

$$V_c^{\text{OT}}(\mu, \nu) = \inf_{\pi \in \text{CPL}(\mu, \nu)} \sup_k \pi(c_k) \stackrel{?}{=} \sup_k \inf_{\pi \in \text{CPL}(\mu, \nu)} \pi(c_k) = \sup_k V_{c_k}^{\text{OT}}(\mu, \nu)$$

with the unjustified step “ $\stackrel{?}{=}$ ”. Surely, $\geq$ is **always true**, and for $\leq$ pick a $\pi^k \in \text{CPL}(\mu, \nu)$ with $\pi^k(c_k) = V_{c_k}^{\text{OT}}(\mu, \nu)$ — an optimizer always exists by lower semi-continuity of $\pi \mapsto \pi(c)$ from

⁷We have not shown this, but it is not difficult to do so.

Corollary 2.3 and convexity and compactness of $C_{PL}(\mu, \nu)$. Further still, by compactness of $C_{PL}(\mu, \nu)$ we can pass to a convergent subsequence $\pi^{k_j} \rightarrow \pi^\infty \in C_{PL}(\mu, \nu)$, which yields (by monotone class argument on $c_k \uparrow c$)

$$\sup_m \underbrace{\sup_k V_{c_k}^{OT}(\mu, \nu)}_{\text{independent of } m} \geq \sup_m \underbrace{\lim_{j \rightarrow \infty} \pi^{k_j}(c_m)}_{\pi^\infty(c_m)} = \pi^\infty(c) \geq V_c^{OT}(\mu, \nu)$$

proving the other inequality.

Finally, we have

$$\begin{aligned} V_c^{OT}(\mu, \nu) &= \sup_k V_{c_k}^{OT}(\mu, \nu) \stackrel{\text{Step 1}}{=} \sup_k \sup_{\substack{(f,g) \in C_b(X) \times C_b(Y) \\ f \oplus g \leq c_k \leq c}} \mu(f) + \nu(g) \\ &\leq \sup_{\substack{(f,g) \in C_b(X) \times C_b(Y) \\ f \oplus g \leq c}} \mu(f) + \nu(g) \stackrel{\text{weak duality}}{\leq} V_c^{OT}(\mu, \nu), \end{aligned}$$

which we set out to prove. □

2.2.3 c-Cyclical Monotonicity

3 Relevant Results

Definition 3.1 (Polish space). A Polish space is a separable completely metrizable topological space. In other words, complete¹ metric space that has a countable dense subset.

Theorem 3.1 (Monotone convergence [3]). Let (S, Σ, μ) be a measure space and $X \in \Sigma$. If (f_n) is a sequence of non-negative measurable functions on X such that

$$0 \leq f_1(x) \leq f_2(x) \leq \dots$$

for all $x \in X$, i.e., (f_n) is **monotone increasing**, then the point-wise supremum $f := \sup_n f_n$ is measurable and

$$\int_X f \, d\mu = \lim_{n \rightarrow \infty} \int_X f_n \, d\mu = \sup_n \int_X f_n \, d\mu.$$

Theorem 3.2 (Dominated convergence [4]). Let (f_n) be a sequence of complex-valued measurable functions on a given measure space (S, Σ, μ) . Suppose that $f_n \rightarrow f$, i.e., f_n converges point-wise to f ,

$$\lim_{n \rightarrow \infty} f_n(x) = f(x)$$

for every $x \in S$. Assume that f_n is dominated by some integrable function g (called majorant) in the sense

$$|f_n(x)| \leq g(x)$$

for all points $x \in S$ and indices n . Then f_n, f are integrable and

$$\lim_{n \rightarrow \infty} \int_S f_n \, d\mu = \int_S \lim_{n \rightarrow \infty} f_n \, d\mu = \int_S f \, d\mu.$$

Proposition 3.1 (Jensen's inequality [5]). Let $(\Omega, \mathcal{A}, \mu)$ be a probability space. Let $f : \Omega \rightarrow \mathbb{R}$ be a μ -measurable function and $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ **convex**. Then

$$\varphi\left(\int_{\Omega} f \, d\mu\right) \leq \int_{\Omega} \varphi \circ f \, d\mu,$$

and, in particular for $p \geq 1$,

$$\int_{\Omega} f(x) \, d\mu(x) \leq \left(\int_{\Omega} |f(x)|^p \, d\mu(x) \right)^{\frac{1}{p}}.$$

¹Every Cauchy sequence of points in M has a limit also in M

Lemma 3.1 (Fatou's [6]). Given a measure space $(\Omega, \mathcal{A}, \mu)$ and a set $X \in \mathcal{A}$, let (f_n) be a sequence of μ -measurable **non-negative** functions on X . Define $f(x) = \liminf_{n \rightarrow \infty} f_n(x)$ for every $x \in X$. Then f is μ -measurable, and

$$\int_X f \, d\mu \leq \liminf_{n \rightarrow \infty} \int_X f_n \, d\mu.$$

Theorem 3.3 (Riesz-Fischer [7]). For $1 \leq p < \infty$, $\mathcal{L}^p(E)$ is a Banach² space. Furthermore, if $f_n \xrightarrow{\mathcal{L}^p} f$, then $(f_n)_n$ has a subsequence that converges to f point-wise almost everywhere on E .

Theorem 3.4 (Cauchy-Schwarz inequality [8]). For all vectors $\mathbf{u}, \mathbf{v}$ of an inner product space, the following holds

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$

Theorem 3.5 (Riesz representation [9]). Let H be a Hilbert space^[^hilbert] whose inner product $\langle x, y \rangle$ is linear in its first argument and anti-linear in its second argument. For **every continuous linear functional** $\varphi \in H^*$, there exists a unique vector $f_\varphi \in H$, called the Riesz representation of φ , such that

$$\varphi(x) = \langle x, f_\varphi \rangle \quad \forall x \in H.$$

Importantly, for complex Hilbert spaces, f_φ is always located in the anti-linear coordinate of the inner product.

Theorem 3.6 (Stone-Weierstrass). Suppose X is a compact Hausdorff space³ and A is subalgebra of $\mathcal{C}(X, \mathbb{R})$ which contains a non-zero constant function. Then A is dense in $\mathcal{C}(X, \mathbb{R})$ if and only if it separates points.

Theorem 3.7 (Hahn-Banach). Let X be a vector space over the field $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Let further

- $Y \subseteq X$ be a linear subspace,
- $p: X \rightarrow \mathbb{R}$ be a sublinear⁴ function and
- $f: Y \rightarrow \mathbb{K}$ be a linear functional so that $\Re f(y) \leq p(y)$ for all $y \in Y$.

Then, there exists a linear functional $F: X \rightarrow \mathbb{K}$ so that

$$F|_Y = f \quad \& \quad \forall x \in X : \Re F(x) \leq p(x).$$

Theorem 3.8 (Geometric Hahn-Banach). Let A and B be non-empty convex subsets of a real locally convex topological vector space X . If $\text{Int } A \neq \emptyset$ and $B \cap \text{Int } A = \emptyset$ then there exists a continuous linear functional f on X such that $\sup f(A) \leq \inf f(B)$ and $f(a) < \inf f(B)$ for all $a \in \text{Int } A$ (such an f is necessarily non-zero).

²Banach space is a complete normed space.

³Hausdorff space is a topological space where distinct points have disjoint neighborhoods

⁴A function $p: X \rightarrow \mathbb{R}$ on a vector field X over $\mathbb{K}$ is called sublinear if it is *positively homogeneous*, i.e., $p(rx) = rp(x)$ for all $r \geq 0$, and subadditive, $p(x+y) \leq p(x) + p(y)$.

3.1 A Note on Weak Topology

See [10] and related pages for more information.

Suppose (X, Y, b) is a pairing (or a **dual system**) of vector spaces over a topological field $\mathbb{K}$, i.e., X and Y are vector spaces of $\mathbb{K}$ and $b: X \times Y \rightarrow \mathbb{K}$ is a bilinear map. For all $x \in X$, let $b(x, \cdot): Y \rightarrow \mathbb{K}$ denote the linear function on Y defined by $y \mapsto b(x, y)$, and similarly for fixed $y \in Y$.

Definition 3.2 (Weak topology of pairing). The *weak topology* on X induced by Y (and b) is the weakest topology on X , denoted by $\sigma(X, Y, b)$ or simply $\sigma(X, Y)$, making all maps $b(\cdot, y): X \rightarrow \mathbb{K}$ continuous as y ranges over Y .

If the field $\mathbb{K}$ has an absolute value $|\cdot|$, then the weak topology $\sigma(X, Y, b)$ on X is induced by the family of seminorms, $p_y: X \rightarrow \mathbb{R}$, defined by $p_y(x) := |b(x, y)|$ for all $y \in Y$ and $x \in X$. In particular, this shows that weak topologies are *locally convex*.

3.1.1 Canonical Duality

Consider now the space case where Y is a vector subspace of the **algebraic dual space**⁵. There is the so-called **canonical pairing** $(X, Y, \langle \cdot, \cdot \rangle)$ or simply (X, Y) whose bilinear map $\langle \cdot, \cdot \rangle$ is the **canonical evaluation map**, defined by $\langle x, x' \rangle = x'(x)$ for all $x \in X$ and $x' \in Y$.

The topology $\sigma(X, Y)$ is the **initial** topology of X with respect to Y .

3.1.2 Weak* Topology

Let X be a **topological vector space (TVS)**⁶ over $\mathbb{K}$. We call the topology X starts with the **original** (or **given**) topology. We may define a possibly different topology on X using the topological or continuous dual space X^* , which consists of all linear functionals from X into the base field $\mathbb{K}$ that are continuous with respect to the given topology.

Definition 3.3 (Weak topology). The *weak topology* on X is the weak topology on X , see Definition 3.2, with respect to the canonical pairing $\langle X, X^* \rangle$.

Definition 3.4 (Weak* topology). The *weak topology* on X^* is the weak topology on X^* with respect to the canonical pairing $\langle X, X^* \rangle$. This topology is also called the **weak* topology**.

⁵Algebraic dual space of X is the vector space of linear functionals on X .

⁶Topological vector space is a vector space equipped with a topology such that vector addition and scalar multiplication are continuous.

Definition 3.5 (Weak* convergence). Suppose that the normed space where we want to work, U , itself is the dual of some Banach⁷ space, i.e., $U = X^*$ ⁸. A sequence $(u_n) \subset U$ **converges in weak*-sense** (or in the **weak* topology**) to $u \in U$, if $u_n(x) \rightarrow u(x)$ for any $x \in X$.

Theorem 3.9 (Riesz-Markov). If K is compact Hausdorff⁹ topological space, the dual of $C(K)$ is isomorphic to the space $\mathcal{M}(K)$ of measures on K , equipped with the total variation norm.

Theorem 3.10 (Banach-Alaoglu). Let X be a Banach space. Then the closed unit ball B of X^* is compact in the weak* topology, see Definition 3.4.

Corollary 3.1. Let X be a normed vector space. Then the closed unit ball $B = \{x \in X \mid \|x\| \leq 1\}$ of the dual space is compact in weak* topology.

⁷Banach space is a complete normed space.

⁸Recall that not everything is a dual space.

⁹Hausdorff topological space separates points, i.e., every pair of distinct points have disjoint neighborhoods.

References

- [1] CONTRIBUTORS, W. (2025). [Total variation distance of probability measures](#).
- [2] GAANS, O. van. (2002). Probability measures on metric spaces.
- [3] CONTRIBUTORS, W. (2026). [Monotone convergence theorem](#).
- [4] CONTRIBUTORS, W. (2025). [Dominated convergence theorem](#).
- [5] CONTRIBUTORS, W. (2026). [Jensen's inequality](#).
- [6] CONTRIBUTORS, W. (2025). [Fatou's lemma](#).
- [7] MEZIANI, H. (2022). [Real analysis MAA 6616 lecture 25 convergence in lp spaces and completeness - Lecture25](#).
- [8] CONTRIBUTORS, W. (2026). [Cauchy–Schwarz inequality](#).
- [9] CONTRIBUTORS, W. (2025). [Riesz representation theorem](#).
- [10] CONTRIBUTORS, W. (2026). [Weak topology](#).