Optimal Transport

For not-so-optimal mathematicians

Štěpán Zapadlo

Table of contents

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

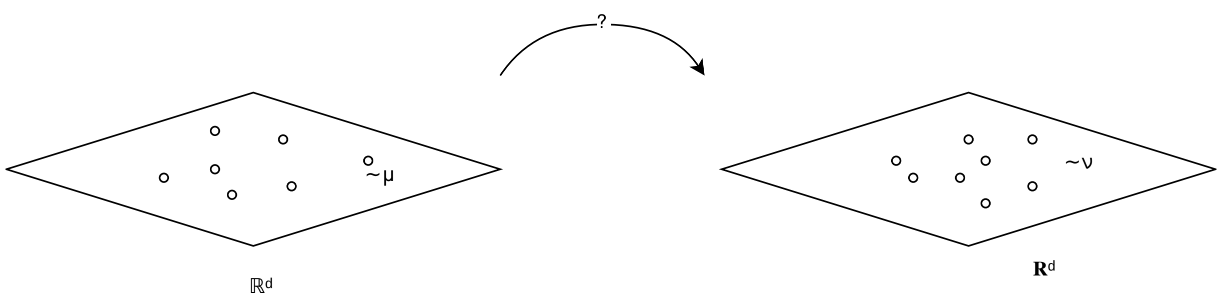


Figure 1.1: Sketch of optimal transport for discrete measures

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.

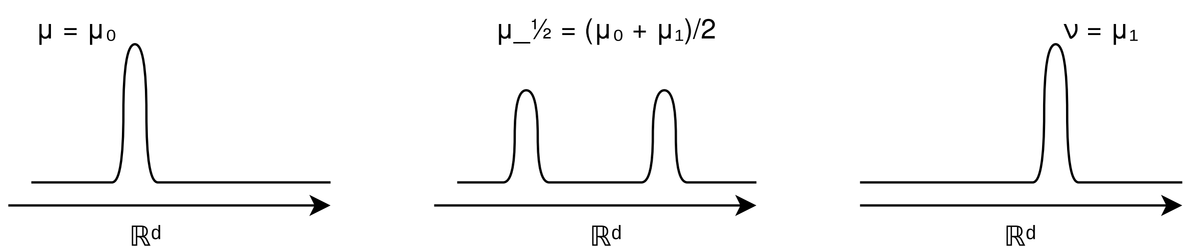


Figure 1.2: Linear average of Dirac measures

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 $\mathscr{P}(\mathbb{R}^d)$?

Question 2: Geometry & Calculus of Probabilities

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

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

1.1.1 Empirical Measures

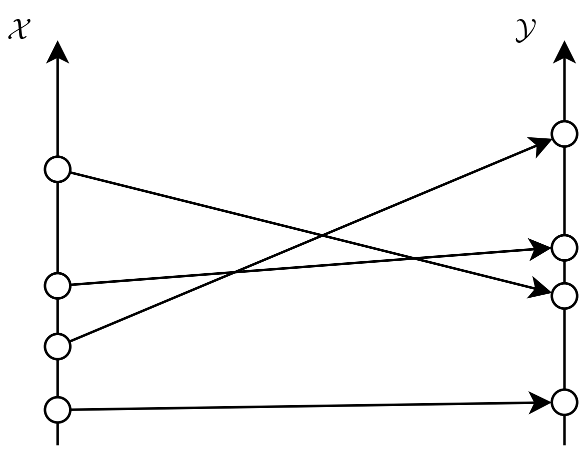
Let us consider μ, ν to be convex combinations of Dirac measures, i.e.,

$$\mu = \sum_{i=1}^{n} \mu_i \delta_{\boldsymbol{x}_i}, \quad \& \quad \nu = \sum_{j=1}^{m} \nu_j \delta_{\boldsymbol{y}_j},$$

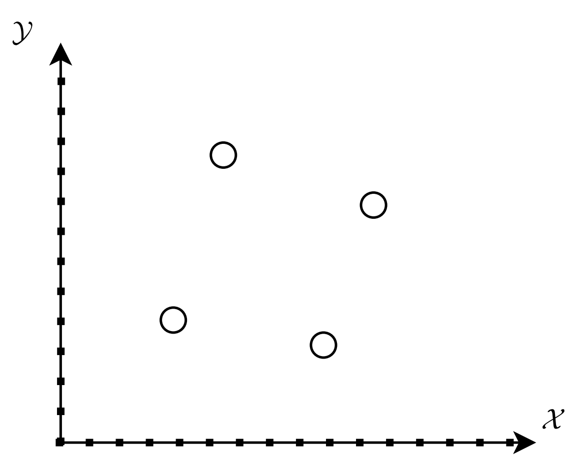
and denote $\mathcal{X} := \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, $\mathcal{Y} := \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_m\}$. The first (naive) idea is to transport "mass at $\boldsymbol{x}_i$ to mass at $\boldsymbol{y}_j$", i.e., define a **transport map** $\boldsymbol{T} \colon \mathcal{X} \to \mathcal{Y}$ s.t.

$$\boldsymbol{T}_\sharp \mu = \sum_{i=1}^{n} \mu_i \delta_{\boldsymbol{T}(\boldsymbol{x}_i)} = \nu,$$

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



(a) Transport map $\boldsymbol{T}$ as pairings



(b) Graph of $\boldsymbol{T}$

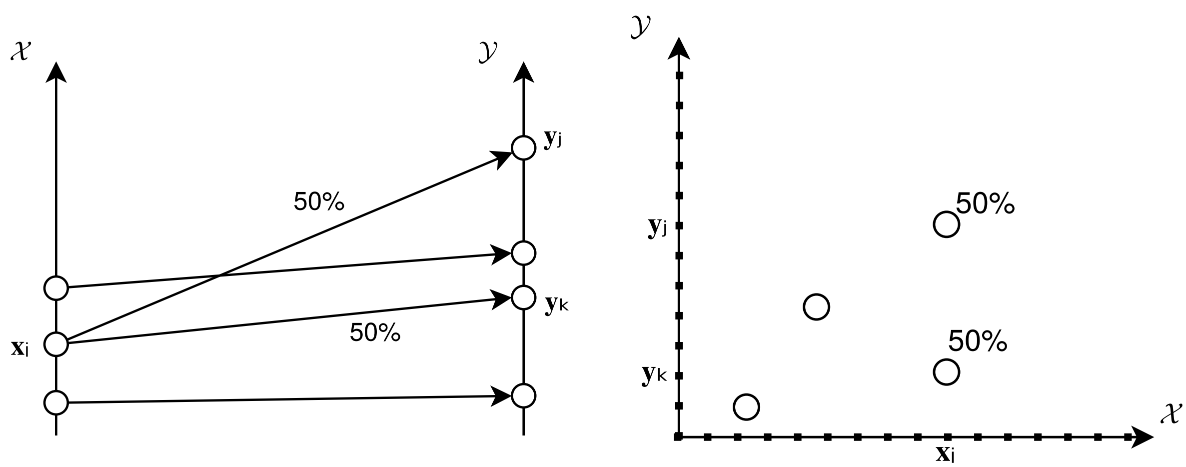
Figure 1.3: Visualizations of a transport map $\boldsymbol{T}$

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

$$\inf_{\boldsymbol{T}_\sharp\mu=\nu}\sum_{i=1}^{n}\mu_i c(\boldsymbol{x}_i, \boldsymbol{T}(\boldsymbol{x}_i)). \tag{1.1}$$

> **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.



(a) Non-Monge map $\boldsymbol{T}$ as pairings

(b) Graph of $\boldsymbol{T}$

Figure 1.4: Visualizations of a non-Monge transport map $\boldsymbol{T}$

Indeed, consider the following Kantorovich optimal transport problem,

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

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 optimizer[1];

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

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

$$\sup_{(f_i)_i,(g_j)_j} \sum_{i=1}^{n} \mu_i f_i + \sum_{j=1}^{m} \nu_j g_j \ \text{ s.t. } \ f_i + g_j \le \mathsf{c}_{ij} = \mathsf{c}(\boldsymbol{x}_i, \boldsymbol{y}_j) \tag{1.3}$$

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

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

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

$$\sum_{i=1}^{k} \mathsf{c}(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{y}}_i) \le \sum_{i=1}^{k} \mathsf{c}(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{y}}_{i+1}) \tag{1.4}$$

   with $\hat{\boldsymbol{y}}_{k+1} = \hat{\boldsymbol{y}}_1$. 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 $f(\hat{\boldsymbol{x}}_i) + g(\hat{\boldsymbol{y}}_i) = \mathsf{c}(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{y}}_i)$ on the support. Thus summing over i gives us

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

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

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

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

2.1 Monge Formulation

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

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

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

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

Example 2.1. Assume $\mu \in \mathscr{P}(\mathbb{R})$ continuous and $\nu \in \mathscr{P}(\mathbb{R})$. 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 $\mathrm{X} \sim \mu$ one has $F_{\mu}(\mathrm{X}) \sim \mathrm{Unif}(0,1)$. Further, $\mathrm{U} \sim \mathrm{Unif}(0,1)$ implies $F_{\nu}^{-1}(\mathrm{U}) \sim \nu$. Hence,

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

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

$$V_{\mathsf{c}}^{\mathrm{Monge}}(\mu, \nu) := \inf_{T_{\sharp}\mu = \nu} \int_{\mathcal{X}} \mathsf{c}(x, T(x)) \, \mathrm{d}\mu(x). \tag{2.1}$$

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

2.2 Kantorovich Formulation

Let us now fix the further notation for projections:

$$\begin{aligned} \mathrm{pr}^{\mathcal{X}} \colon \mathcal{X} \times \mathcal{Y} &\to \mathcal{X} \\ (x,y) &\mapsto x \end{aligned} \qquad \& \qquad \begin{aligned} \mathrm{pr}^{\mathcal{Y}} \colon \mathcal{X} \times \mathcal{Y} &\to \mathcal{Y}. \\ (x,y) &\mapsto y \end{aligned}$$

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

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

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

Example 2.2.

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

$$\underbrace{(\mathrm{id}, T)_{\sharp} \mu}_{\mathrm{Law}(\mathrm{X}, T(\mathrm{X})) \text{ if } \mathrm{X} \sim \mu} = \pi \in \textsc{Cpl}(\mu, \nu);$$

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

$$\left(F_{\mu}^{-1}, F_{\nu}^{-1}\right)_{\sharp} \lambda = \mathrm{Law}\left(F_{\mu}^{-1}(\mathrm{U}), F_{\nu}^{-1}(\mathrm{U})\right) \in \textsc{Cpl}(\mu, \nu)$$

   is called a **monotone coupling**;
3. Lastly, consider independent random variables $\mathrm{X} \sim \mu$ and $\mathrm{Y} \sim \nu$, then the product measure $\mu \otimes \nu = \mathrm{Law}(\mathrm{X}, \mathrm{Y}) \in \textsc{Cpl}(\mu, \nu)$ is the independent (product) coupling.

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

$$V_{\mathrm{c}}^{\mathrm{OT}}(\mu, \nu) = \inf_{\pi \in \textsc{Cpl}(\mu, \nu)} \int \mathrm{c}(x, y)\, \mathrm{d}\pi(x, y). \tag{2.2}$$

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

1. $\textsc{Cpl}(\mu, \nu)$ to be non-empty (trivially satisfied), convex[1] and compact (to be shown), and
2. $\pi \mapsto \int \mathrm{c}\, \mathrm{d}\pi =: \pi(\mathrm{c})$ to be lower semi-continuous.

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

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

$$V_{\mathsf{c}}^{\mathrm{OT}}(\mu,\nu) = \lim_k \pi^k(\mathsf{c}) \geq \pi(\mathsf{c}) \geq V_{\mathsf{c}}^{\mathrm{OT}}(\mu,\nu)$$

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

Definition 2.4. The weak topology of measures on $\mathscr{M}(\mathcal{X})$ is the initial topology[2] with respect to $\mathscr{M}(\mathcal{X}) \ni \mu \mapsto \mu(f)$ for $f \in C_{\mathrm{b}}(\mathcal{X})$.

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

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

1. $x \mapsto \delta_x$ is the *topological embedding*[3] $\mathcal{X} \hookrightarrow \mathscr{P}(\mathcal{X})$;
2. this topology is *Hausdorff*[4], i.e., for $\mu,\nu \in \mathscr{M}(\mathcal{X})$ with $\mu(f) = \nu(f)\ \forall f \in C_{\mathrm{b}}(\mathcal{X})$ implies by monotone class argument $\mu(f) = \nu(f)\ \forall f \in M_{\mathrm{b}}(\mathcal{X})$, which in turn means $\mu = \nu$;
3. it is *countably generated*, i.e., there exists a countable family $(f_n)_{n\in\mathbb{N}} \subseteq C_{\mathrm{b}}(\mathcal{X})$ 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**
$$d_{\mathrm{Prokhorov}}(\mu,\nu) := \inf\left\{\varepsilon > 0 \mid \forall A \in \mathscr{B}(\mathcal{X}) : \begin{cases}\mu(A) \leq \nu(A_\varepsilon) + \varepsilon, \\ \nu(A) \leq \mu(A_\varepsilon) + \varepsilon\end{cases} \text{ where } A_\varepsilon := \bigcup_{x\in A} B(x,\varepsilon)\right\}; \tag{2.3}$$
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
$$\left\{\sum_{k=1}^n \alpha_k \delta_{x_k} \mid n \in \mathbb{N}\ \&\ \alpha_k \in \mathbb{Q}\cap[0,1]\ \&\ \sum_{k=1}^n \alpha_k = 1\right\}$$
is countable and even dense in $\mathscr{P}(\mathcal{X})$.

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

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

[2] Also 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*.
[3] Namely, every embedding is injective and continuous.
[4] Hausdorff topology is that such distinct points have disjoint neighborhoods — that is why one goes to the indicators of such neighborhoods being in $M_{\mathrm{b}}(\mathcal{X})$ 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 $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$ 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 $\mathscr{P}(\mathcal{X})$ is compact.

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

Further, denote $\mathscr{M}_{\le 1}(\mathcal{X}) := \left\{\mu \mid \mathrm{TV}(\mu)\le 1\right\}$ a closed unit ball in the dual of $C(\mathcal{X})$. By Banach-Alaoglu 3.10 $\mathscr{M}_{\le 1}(\mathcal{X})$ 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 $\mathcal{X}$,

$$\mathscr{M}_{\le 1}(\mathcal{X}) \supseteq \mathscr{P}(\mathcal{X}) = \bigcap_{\substack{f\in C(\mathcal{X})\\ f\ge 0}} \left\{\mu \mid \mu(f)\ge 0\right\} \cap \left\{\mu \mid \mu(1) = 1\right\},$$

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 $\mathscr{P}(\mathcal{X})$ is weak* closed subset of a weak* compact set, thus itself is weak* compact in $(\mathscr{M}(\mathcal{X}), C(\mathcal{X}))$. 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 $\mathcal{E}\subseteq\mathscr{P}(\mathcal{X})$ is called **tight**, if for all $\varepsilon>0$ there exists $K \overset{c}{\subseteq} \mathcal{X}$ such that

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

Theorem 2.1 (Prokhorov's)**.** *A family $\mathcal{E}\subseteq\mathscr{P}(\mathcal{X})$ is tight if and only if $\mathcal{E}$ is relatively compact*[5].

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 $K \overset{c}{\subseteq} \mathcal{X}$ such that $\inf_{\mu\in\mathcal{E}}\mu(K)\ge 1-\varepsilon$. 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

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

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

[5] A set X is relatively compact when $\overline{X}$ 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 $\inf_{\mu\in\mathcal{E}} \mu(K) = 1 - \sup_{\mu\in\mathcal{E}} \mu(K^{\mathrm{c}})$ where $K^{\mathrm{c}} = \bigcup_{n\in\mathbb{N}} \overline{K_n}^{\mathrm{c}} \subseteq \bigcup_{n\in\mathbb{N}} K_n^{\mathrm{c}}$. Lastly, by subadditivity of measures

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

□

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* $\textsc{Cpl}(\mu, \nu)$ *is compact.*

Proof. First, we prove that $\textsc{Cpl}(\mu, \nu)$ is a closed set. For this recall that $\mathrm{pr}^{\mathcal{X}}$ is a continuous function. Then its pushforward $\mathrm{pr}^{\mathcal{X}}_\sharp \colon \mathscr{P}(\mathcal{X} \times \mathcal{Y}) \to \mathscr{P}(\mathcal{X})$ defined as $\pi \mapsto \mathrm{pr}^{\mathcal{X}}_\sharp \pi$ is also continuous since for any $f \in C_{\mathrm{b}}(\mathcal{X})$ we have

$$\int f \,\mathrm{d}\, \mathrm{pr}^{\mathcal{X}}_\sharp \pi = \int \underbrace{(f \circ \mathrm{pr}^{\mathcal{X}})}_{\in C_{\mathrm{b}}(\mathcal{X}\times\mathcal{Y})}(x, y) \,\mathrm{d}\pi(x, y).$$

Now, we can write

$$\textsc{Cpl}(\mu, \nu) = \left\{\pi \in \mathscr{P}(\mathcal{X} \times \mathcal{Y}) \mid \mathrm{pr}^{\mathcal{X}}_\sharp \pi = \mu\right\} \cap \left\{\pi \in \mathscr{P}(\mathcal{X} \times \mathcal{Y}) \mid \mathrm{pr}^{\mathcal{Y}}_\sharp \pi = \nu\right\},$$

where both of the "partial" sets are closed by continuity of $\mathrm{pr}^{\mathcal{X}}$ and $\mathrm{pr}^{\mathcal{Y}}$, respectively. Thus, set $\textsc{Cpl}(\mu, \nu)$ is also closed.

Second, we show that $\textsc{Cpl}(\mu, \nu)$ 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 $K_{\mathcal{X}} \overset{\mathrm{c}}{\subseteq} \mathcal{X}$ and $K_{\mathcal{Y}} \overset{\mathrm{c}}{\subseteq} \mathcal{Y}$, 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 $K := K_{\mathcal{X}} \times K_{\mathcal{Y}} \overset{\mathrm{c}}{\subseteq} \mathcal{X} \times \mathcal{Y}$. Then for $\pi \in \textsc{Cpl}(\mu, \nu)$,

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

Hence $\textsc{Cpl}(\mu, \nu)$ is tight.

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

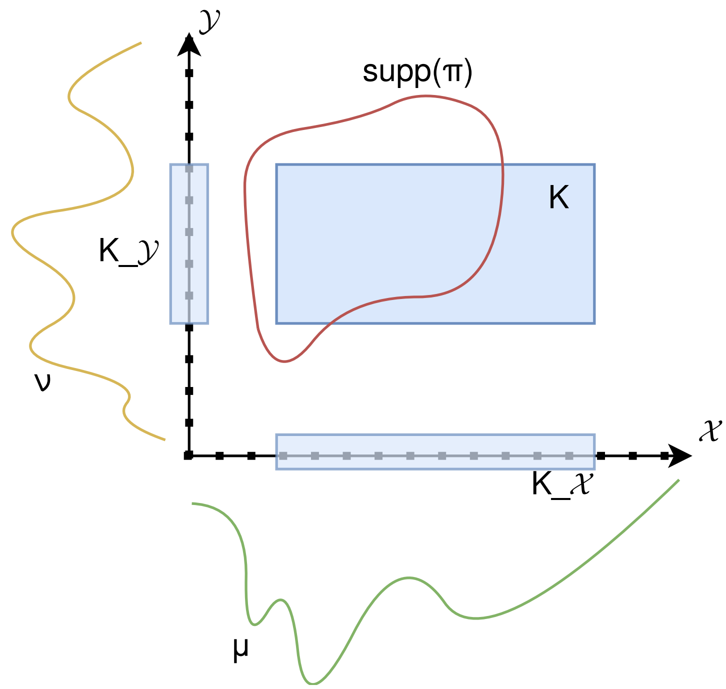


Figure 2.1: Role of $K^{\mathcal{X}}$ and $K^{\mathcal{Y}}$

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

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

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

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

Lemma 2.1. *Let $f\colon \mathcal{X} \to \mathbb{R}^{\infty}$ 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 $(f_k)_{k\in\mathbb{N}} \subset C_{\mathrm{b}}(\mathcal{X})$ such that $f_k \uparrow f$.*

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

Proof. “ $\Longleftarrow$ ”: Take a sequence $x_n \to x$. Then

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

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

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

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

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

for large enough values of k. In total, $f_k(x_0) \ge \alpha$ for large values of k. As α was chosen arbitrarily, we can combine it with the (trivial) inequality $f_k \le f$ 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 $\left|g_k(x) - g_k(\tilde{x})\right|$ is either 0 (both are truncated), $\left|f_k(x) - f_k(\tilde{x})\right|$ (see above) or, without loss of generality assume $f_k(x) > f_k(\tilde{x})$ by symmetry,

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

□

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

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

$$\pi(\mathsf{c}) = \sup_{k\in\mathbb{N}} \pi(\mathsf{c}_k) = \sup_{k\in\mathbb{N}} \lim_{j\to\infty} \pi_j(\mathsf{c}_k) \le \liminf_{j\to\infty} \sup_{k\in\mathbb{N}} \pi_j(\mathsf{c}_k) = \liminf_{j\to\infty} \pi_j(\mathsf{c}).$$

□

Proposition 2.2. *The Kantorovich optimal transport cost* $V_{\mathsf{c}}^{\mathrm{OT}}\colon \mathscr{P}(\mathcal{X})\times\mathscr{P}(\mathcal{Y}) \to \mathbb{R}^\infty$, *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 $V_{\mathsf{c}}^{\mathrm{OT}}$.

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

$$\textsc{Cpl}(\mu,\nu) = \bigcap_{(f,g)\in C_b(\mathcal{X})\times C_b(\mathcal{Y})} \left\{\pi \in \mathscr{P}(\mathcal{X}\times\mathcal{Y}) \mid \pi(f\oplus g) = \mu(f) + \nu(g)\right\},$$

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

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

Intuitively, the supremum finds the "most significant violation" of the intersection used to express $\textsc{Cpl}(\mu,\nu)$ *above.*

By the minimax theorem since we are taking inf over a compact set and sup over a vector space and

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

we get

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

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

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

and together with $\pi(\mathsf{c} - f\oplus g) \ge \inf_{(x,y)} \mathsf{c} - f\oplus g$ for every π (and diracs are admissible), we get even an equality. Thus $f(x) + g(y) + \alpha(f,g) \le \mathsf{c}(x,y)$ 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 $\tilde{f}\oplus g \le \mathsf{c}$) finally produces the *duality* result:

$$V_{\mathsf{c}}^{\mathrm{OT}}(\mu,\nu) = \inf_{\pi\in\textsc{Cpl}(\mu,\nu)} \pi(\mathsf{c}) = \sup_{\substack{(f,g)\in C_b(\mathcal{X})\times C_b(\mathcal{Y}) \\ f\oplus g\le \mathsf{c}}} \mu(f) + \nu(g). \tag{2.4}$$

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. $F\colon \mathcal{X} \to \mathbb{R}^\infty$ is a given **proper** function.

Definition 2.7 (Convex conjugate)**.** The *convex conjugate* of F is said to be the function $F^*\colon \mathcal{X}' \to \mathbb{R}^\infty$ 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, $\operatorname{epi} F = \{(x,y) \mid F(x) \leq y\}$ 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 $\operatorname{epi} F$ and $(\tilde{x}, \tilde{c}) \neq \operatorname{epi} F$ (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 $K\colon \mathcal{X} \to \mathscr{P}(\mathcal{Y})$ is a *(probability) kernel* if it is measurable with respect to $\mathscr{B}(\mathcal{X})$ and the Borel σ-algebra on $\mathscr{P}(\mathcal{Y})$.

Proposition 2.3. *The following are equivalent:*

1. *K is measurable,*
2. *$\forall f \in C_b(\mathcal{Y})$ the mapping*[6] *$x \mapsto \int f(y)K(x; \mathrm{d}y)$ is measurable, and*
3. *$\forall f \in M_b(\mathcal{X} \times \mathcal{Y})$ the mapping $x \mapsto \int f(x,y)K(x;\mathrm{d}y)$ is measurable.*

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

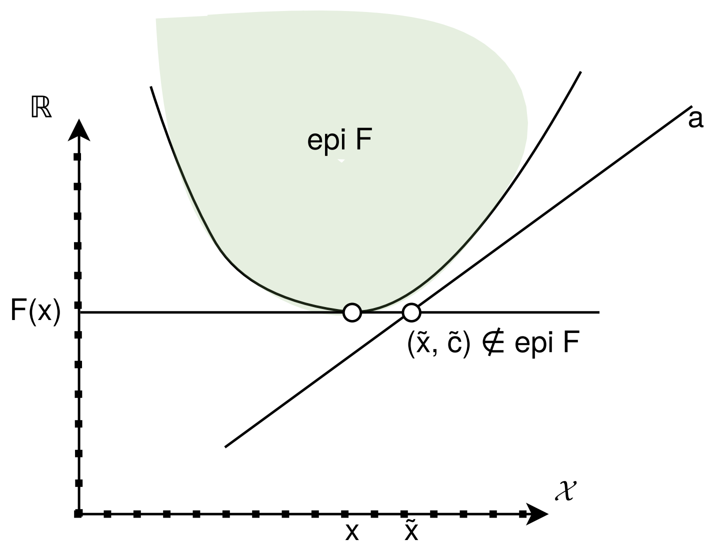


Figure 2.2: Affine minorants of $F(x)$

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

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

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

Then K is measurable

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

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

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

$$\forall f \in \mathcal{M}_b(\mathcal{X}\times\mathcal{Y}) : \ \int f \mathrm{d}\mu\otimes K := \iint f(x,y) K(x;\mathrm{d}y)\mu(\mathrm{d}x).$$

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

Theorem 2.3 (Disintegration)**.** *Let $\pi \in \mathscr{P}(\mathcal{X}\times\mathcal{Y})$, then there exists a probability kernel $K \colon \mathcal{X} \to \mathscr{P}(\mathcal{Y})$ such that $\pi = \mu\otimes K$ with $\mu := \mathrm{pr}^{\mathcal{X}}_{\sharp}\pi$. Secondly, let $\tilde{K} : \mathcal{X} \to \mathscr{P}(\mathcal{Y})$ 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 $(\pi_x)_{x\in\mathcal{X}}$.

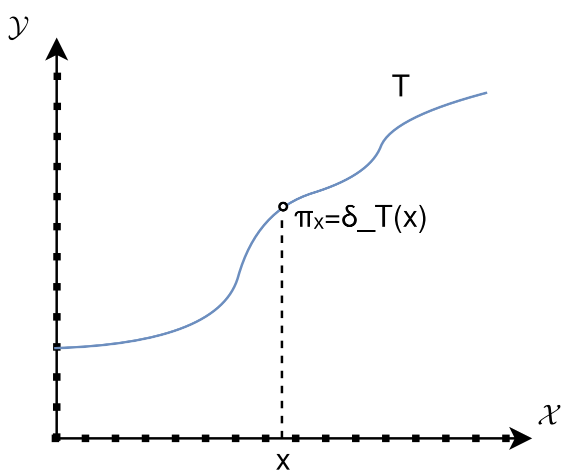
Example 2.3.

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

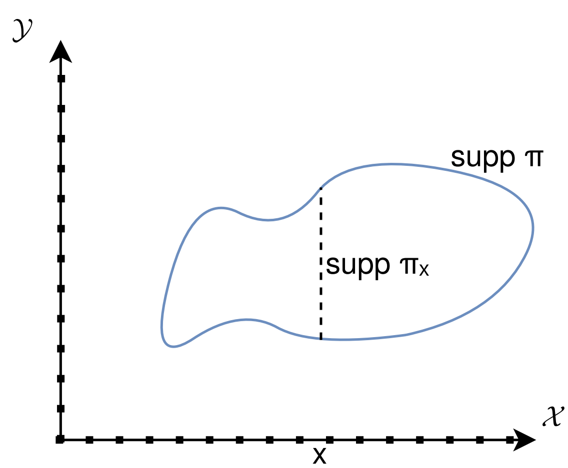
$$\pi\, \mathrm{Law}(\mathrm{X}, T(\mathrm{X})) = (\mathrm{id}, T)_\sharp \mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \implies \pi_x = \delta_{T(x)}.$$

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

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



(a) Disintegration of example 1



(b) Disintegration of example 2,3

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* $\mathsf{c}\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^\infty$ *be lower semi-continuous and lower bounded cost function, then*

$$V_{\mathsf{c}}^{\mathrm{OT}} = \sup_{\substack{(f,g)\in C_b(\mathcal{X})\times C_b(\mathcal{Y}) \\ f\oplus g \le \mathsf{c}}} \mu(f) + \nu(g), \tag{2.5}$$

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

Remark 2.5.

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

$$\mu(f) + \nu(g) = \pi(f \oplus g) \le \pi(\mathsf{c}),$$

   which implies $V_{\mathsf{c}}^{\mathrm{OT}} \ge$ (2.6).
2. (2.6) holds even when $\mathsf{c} \in \mathcal{M}_{\mathsf{b}}(\mathcal{X} \times \mathcal{Y})$ (via Choquet capacitability theory).
3. When c is even Lipchitz, one can restrict (f,g) in (2.5) to be Lipschitz as well.
4. For $(f,g) \in \mathcal{C}_{\mathsf{b}}(\mathcal{X}) \times \mathcal{C}_{\mathsf{b}}(\mathcal{Y})$ admissible, the c-transforms

$$f^{\bar{\mathsf{c}}}(y) := \inf_{x \in \mathcal{X}} (\mathsf{c}(x,y) - f(x)) \ge g(y),$$
$$g^{\mathsf{c}}(x) := \inf_{y \in \mathcal{Y}} (\mathsf{c}(x,y) - g(y)) \ge f(x)$$

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

3 Relevant Results

Definition 3.1 (Polish space)**.** A Polish space is a separable completely metrizable topological space. In other words, complete[1] 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*

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

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 $f_n \to f$, i.e., f_n converges point-wise to f,*

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

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

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

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

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

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

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

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

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

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

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

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

Theorem 3.3 (Riesz-Fischer [7]). *For $1 \leq p < \infty$, $\mathcal{L}^p(E)$ is a Banach[2] space. Furthermore, if $f_n \underset{\mathcal{L}^p}{\to} f$, 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 $\boldsymbol{u}, \boldsymbol{v}$ of an* inner product space*, the following holds*

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

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*

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

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 space[3] 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,*
- *$p \colon X \to \mathbb{R}$ be a sublinear[4] function and*
- *$f \colon Y \to \mathbb{K}$ be a linear functional so that $\mathfrak{Re}\, f(y) \leq p(y)$ for all $y \in Y$.*

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

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

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 $\operatorname{Int} A \neq \emptyset$ and $B \cap \operatorname{Int} A = \emptyset$ 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 $a \in \operatorname{Int} A$ (such an f is necessarily non-zero).*

[2]Banach space is a complete normed space.

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

[4]A function $p \colon X \to \mathbb{R}$ 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 $b\colon X \times Y \to \mathbb{K}$ is a bilinear map. For all $x \in X$, let $b(x, \cdot)\colon Y \to \mathbb{K}$ 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 $b(\cdot, y)\colon X \to \mathbb{K}$ 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, $p_y\colon X \to \mathbb{R}$, defined by $p_y(x) := \left|b(x, y)\right|$ 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**.

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

[6] Topological 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 Banach[7] 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 $u_n(x) \to u(x)$ for any $x \in X$.

Theorem 3.9 (Riesz-Markov)**.** *If K is compact Hausdorff[9] topological space, the dual of $\mathbb{C}(K)$ is isomorphic to the space $\mathscr{M}(K)$ 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.*

[7]Banach space is a complete normed space.
[8]Recall that not everything is a dual space.
[9]Hausdorff 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.