STOCHASTIC MASS TRANSFER

Abstract

The theory of optimal transport (OT) has seen a tremendous development in the last 25 years with fascinating applications ranging from geometric and functional inequalities over PDEs and geometry to image analysis and statistics. In recent years, variants of the optimal transport problem with additional stochastic constraints have received increasing attention, e.g. weak optimal transport (WOT), entropic optimal transport (EOT), martingale optimal transport (MOT) and causal/adapted optimal transport (COT).

The aim of this lecture[1] is to serve as an introduction into the stochastic variants of the transport problem. After a quick recall of the classical OT problem we will start investigating the above mentioned probabilistic versions.

Frequently used notation

- X, Y denote Polish spaces
- For a Polish space X we denote the probability measures over X by $\mathcal{P}(\mathsf{X})$, the set of Borel measures by $\mathcal{M}(\mathsf{X})$, and the Borel sets by $\mathcal{B}(\mathsf{X})$.
- For a map $T : \mathsf{X} \to \mathsf{Y}$ and $\mu \in \mathcal{P}(\mathsf{X})$ we denote the image measure of μ under T by $T(\mu) = T_{\#}\mu = \mu \circ T^{-1}$
- The set of all all couplings between two probability measures μ, ν will be denoted by $\mathsf{Cpl}(\mu, \nu)$.
- $C_b(\mathsf{X})$ denotes the continuous and bounded functions $f : \mathsf{X} \to \mathbb{R}$.
- For integrable $f : \mathsf{X} \to \mathbb{R}$ and $\mu \in \mathcal{M}(\mathsf{X})$ we often write $\mu(f) := \int f d\mu$.

1. The optimal transport problem

In this section we will give a short introduction into the theory of optimal transport. This will serve as a benchmark or guidance for what to expect for the different stochastic variations of the transport problem we will consider in the next sections.

For reference and further reading we refer to the books [San15, AG13, Vil03].

1.1. On how mass is transported.

Definition 1.1. *A topological space (X, τ) is called Polish, iff it is separable and there exists a metric d metrizing τ s.t. (X, d) is a complete metric space.*

Let X, Y be Polish spaces and denote the set of probability measures by $\mathcal{P}(\mathsf{X}), \mathcal{P}(\mathsf{Y})$. Given two distributions $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ we are interested in ways of *transporting mass* distributed according to μ into mass distributed according to ν. In mathematical terms:

Definition 1.2. *For a Borel function $T : \mathsf{X} \to \mathsf{Y}$ we define the push-forward of μ by T or the image measure of μ under T by*

$$T(\mu) := T_{\#}\mu = \mu \circ T^{-1},$$

i.e. $T(\mu)(A) = \mu(T^{-1}(A))$ for all $A \in \mathcal{B}(\mathsf{Y})$. If $T(\mu) = \nu$ we call T a transport map (or Monge transport) from μ to ν.

This problem was first formulated by Gaspard Monge in 1781 in the article *"Sur la theorie des déblais et des remblais"* [Mon81] where he was interested in minimizing the transport cost of moving a pile of sand.

[1] These notes are based on earlier lectures / lecture notes of Julio Backhoff-Veraguas, Martin Huesmann and Gudmund Pammer.

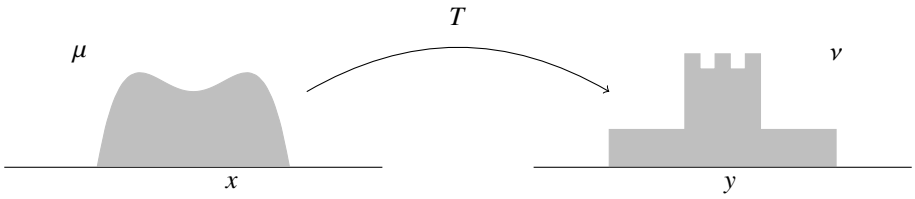


FIGURE 1. A possible transport from a distribution μ to a distribution ν via a map T.

Remark 1.3. In $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$, if μ, ν have densities and T is regular enough, then T is a transport map between μ and ν iff

$$\det(DT)\frac{d\nu}{dx} \circ T = \frac{d\mu}{dx},$$

as follows from the change of variables formula. This is a complicated PDE in the unknown T, called the Monge-Ampère Equation. Finding an optimal map then boils down to finding a solution with further structural properties.

In general, transport maps from μ to ν might not exist:

Example 1.4. Assume $\mu = \delta_0 \in \mathcal{P}(\mathbb{R})$ and $\nu \neq \delta_a$ for all $a \in \mathbb{R}$. Since, $T(\mu) = \delta_{T(0)}$ for any transport map T there cannot be a map T s.t. $T(\mu) = \nu$.

Another problem with the notion of transport maps is that the constraint $T(\mu) = \nu$ is not closed w.r.t. a reasonable topology.

Definition 1.5. *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$. A coupling of μ and ν is a measure $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ with marginals μ and ν, i.e.*

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

The set of all couplings of μ and ν will be denoted by $\mathsf{Cpl}(\mu, \nu)$.

Stochastically, a coupling π of μ and ν is a joint law of two random variables (X, Y) such that $\text{law}_\pi(X) = \mu$ and $\text{law}_\pi(Y) = \nu$. In particular, conditioning on $X = x$ we can interpret the regular conditional probability $\pi(\cdot|X = x)$ as a plan on how to transport the mass at x. Therefore, we will often call coupling transport plans. Analytically, this corresponds to disintegrating π w.r.t. its first marginal μ to obtain a family of probability measures $(\pi_x(dy))_{x\in\mathsf{X}}$. In terms of the projections

$$\text{proj}_\mathsf{X} : \mathsf{X} \times \mathsf{Y} \to \mathsf{X}, (x, y) \mapsto x, \text{proj}_\mathsf{Y} : \mathsf{X} \times \mathsf{Y} \to \mathsf{Y}, (x, y) \mapsto y$$

a measure $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ is an element of $\mathsf{Cpl}(\mu, \nu)$ iff $\text{proj}_\mathsf{X}(\pi) = \mu$ and $\text{proj}_\mathsf{Y}(\pi) = \nu$. A further equivalent characterization of $\mathsf{Cpl}(\mu, \nu)$ is

$$\{\pi \in \mathbb{P}(\mathsf{X} \times \mathsf{Y}) : \int f(x)\, d\pi(x, y) = \int f\, d\mu, \int g(y)\, d\pi(x, y) = \int g\, d\mu \text{ for } f \in C_b(\mathsf{X}), g \in C_b(\mathsf{Y})\}. \tag{1.1}$$

The set $\mathsf{Cpl}(\mu, \nu)$ is always non-empty. Indeed the product coupling (stochastically, the independent coupling) satisfies $\mu \otimes \nu \in \mathsf{Cpl}(\mu, \nu)$.

Remark 1.6. Observe, that any transport map $T : \mathsf{X} \to \mathsf{Y}$ from μ to ν induces a transport plan $\pi_T := (\text{Id}, T)(\mu) \in \mathsf{Cpl}(\mu, \nu)$. We call π_T a Monge coupling or the coupling induced by the map T.

We give some further examples of transport maps / couplings:

Example 1.7. Let ν be a probability measure on $\mathbb{R}$ and write F^ν for its distribution function. The corresponding quantile function is given by the generalized inverse $q^\nu : (0, 1) \to \mathbb{R}$ defined by

$$q^\nu(u) := \inf\{x : F(x) > 0\}. \tag{1.2}$$

Writing λ for the Lebesgue measure on the unit interval, we have $q^\nu_\#(\lambda) = \nu$, that is q^ν if a Monge-map which takes λ to ν.

If ν has atoms, then so does $F^\nu(\nu)$ and in particular F^ν is *not* a Monge-map from ν to λ.

On the other hand, if μ is a continuous probability on $\mathbb{R}$ (i.e. has no atoms), than $F^\mu_\#(\mu) = \lambda$. In this situation the map $T := q^\nu \circ F^\mu$ is a Monge-transport from μ to ν, the so called *monotone transport* mapping.

Example 1.8. If μ, ν are (non necessarily continuous) measures on the real line, $\pi := (q^\mu, q^\nu)_\#\lambda$ is a coupling of μ, ν, the so called *co-monotone coupling*.

1.2. **The Monge and Kantorovich optimal transport problem.** Fix a Borel measurable function $c : \mathsf{X} \times \mathsf{Y} \to [0, \infty]$. We will interpret c as the cost of transporting a unit of mass from $x \in \mathsf{X}$ to $y \in \mathsf{Y}$. Therefore, we will call such a function a cost function.

Slightly more generally we will also consider $c : \mathsf{X} \times \mathsf{Y} \to (-\infty, \infty]$ which is lower bounded in the sense that there exists $a \in L^1(\mu), b \in L^1(\nu)$, such that $a(x) + b(y) \le c(x, y)$ for $x \in \mathsf{X}, y \in \mathsf{Y}$. Problems for such cost functions can be reduced to the case of non-negative costs functions by considering the mapping $(x, y) \mapsto c(x, y) + a_-(x) + b_-(y)$ which leads to an equivalent optimization problem.

In virtually all applications c is continuous or at least lower semi-continuous and we will freely make this assumption if it simplifies arguments.

Definition 1.9. *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ and c a cost function. The Monge problem is to solve*

$$P_c^M := P_c^M(\mu, \nu) := \inf \int c(x, T(x))\, \mu(dx), \tag{MP}$$

where the infimum runs over all transport maps $T : \mathsf{X} \to \mathsf{Y}$ such that $T(\mu) = \nu$. Any map T attaining the infimum in (MP) *is called optimal transport map.*

Definition 1.10. *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ and c a cost function. The Kantorovich problem is to solve*

$$P_c^K := P_c^K(\mu, \nu) := \inf \int c(x, y)\, \pi(dx, dy), \tag{KP}$$

where the infimum runs over all couplings $\pi \in \mathsf{Cpl}(\mu, \nu)$. Any coupling π attaining the infimum in (KP) *is called optimal coupling or optimal transport plan.*

Remark 1.11 (Kantorovich problem in probabilistic terms). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a (non-atomic) probability space and write X, Y, Z for random variables on Ω. Then $\mathsf{Cpl}(\mu, \nu) = \{ \text{law}(X, Y) : X \sim \mu, Y \sim \nu\}$ and we can formulate the transport problem as

$$P_c^K := \inf\{\mathbb{E}[c(X, Y)] : X \sim \mu, Y \sim \nu\}.$$

As we will see, the Kantorovich problem is much better behaved than the Monge problem. For instance, the following properties are immediate.

Remark 1.12.
- The set $\mathsf{Cpl}(\mu, \nu)$ is convex.
- The map $\pi \mapsto \int c\, d\pi$ is linear.

Moreover, $\mathsf{Cpl}(\mu, \nu)$ is compact in a natural topology which will allow us to show existence of optimal couplings under some assumption on the cost function c: Recall that a sequence of measures $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}(\mathsf{X})$ converges weakly to $\mu \in \mathcal{P}(\mathsf{X})$ iff

$$\int f d\mu_n \to \int f d\mu, \quad \text{for all } f \in C_b(\mathsf{X}),$$

where $C_b(\mathsf{X})$ denote the continuous and bounded functions on X. We call the induced topology on $\mathcal{P}(\mathsf{X})$ the weak topology.

Theorem 1.13 (Prokhorov)**.** *Let X be a Polish space. A family $A \subseteq \mathcal{P}(\mathsf{X})$ of probability measures on X is relatively compact w.r.t. the weak topology iff it is tight, i.e. for every $\varepsilon > 0$ there exists $K_\varepsilon \subseteq \mathsf{X}$ compact such that*

$$\sup_{\mu\in A} \mu(\mathsf{X} \setminus K_\varepsilon) \le \varepsilon.$$

For a proof we refer to [Bil99].

Lemma 1.14. *If $A_1 \subseteq \mathcal{P}(\mathsf{X}), A_2 \subseteq \mathcal{P}(\mathsf{Y})$ are tight so is $A_3 := \{\pi \in \mathcal{P}(\mathsf{X}\times\mathsf{Y}) : \mathrm{proj}_{\mathsf{X}}(\pi) \in A_1$ and $\mathrm{proj}_{\mathsf{Y}}(\pi) \in A_2\}$.*

Proof. Let $\pi \in A_3$ and $\varepsilon > 0$ be given. Pick $K_1 \subseteq \mathsf{X}, K_2 \subseteq \mathsf{Y}$ such that $\mu(\mathsf{X}\setminus K_1) \leq \varepsilon, \nu(\mathsf{Y}\setminus K_2) \leq \varepsilon$ for all $\mu \in A_1, \nu \in A_2$. Since, $K_1 \times K_2 \subseteq \mathsf{X}\times\mathsf{Y}$ is compact the claim follows from

$$\pi(\mathsf{X}\times\mathsf{Y}\setminus K_1\times K_2) \leq \pi((\mathsf{X}\setminus K_1)\times\mathsf{Y}) + \pi(\mathsf{X}\times(\mathsf{Y}\setminus K_2)) = \mu(\mathsf{X}\setminus K_1) + \nu(\mathsf{Y}\setminus K_2) \leq 2\varepsilon.$$

□

Corollary 1.15. *The set $\mathsf{Cpl}(\mu,\nu)$ is compact.*

Proof. Since $\{\mu\} \subseteq \mathcal{P}(\mathsf{X}), \{\nu\} \subseteq \mathcal{P}(\mathsf{Y})$ are tight, $\mathsf{Cpl}(\mu,\nu)$ is tight by Lemma 1.14. It remains to show that it is closed. Pick $(\pi_n)_{n\in\mathbb{N}} \subseteq \mathsf{Cpl}(\mu,\nu)$ with limit π. We have to show that π has marginals μ and ν. Pick $\varphi \in C_b(\mathsf{X})$ and define $\bar{\varphi}(x,y) := \varphi(x)$ so that $\bar{\varphi} \in C_b(\mathsf{X}\times\mathsf{Y})$. Then, we know that

$$\int \varphi\, d\pi = \int \bar{\varphi}\, d\pi = \lim_n \int \bar{\varphi}\, d\pi_n = \lim_n \int \varphi\, d\pi_n = \int \varphi\, d\mu$$

so that $\mathrm{proj}_{\mathsf{X}}(\pi) = \mu$. Similarly, it follows that $\mathrm{proj}_{\mathsf{Y}}(\pi) = \nu$. □

A function $f : Z \to (-\infty,\infty]$ is lower semi-continuous (l.s.c.) if for all sequence $z, z_1, z_2, \ldots \in Z$, $\lim_{n\to\infty} z_n = z$ we have $\liminf f(z_n) \geq f(z)$. Equivalently, f is l.s.c. if there is a sequence of continuous functions $f_1, f_2, \ldots : Z \to [0,\infty)$ such that $f = \sup_n f_n$. (Clearly $f_1, f_2, \ldots$ can be chosen to be upper bounded and if f is lower bounded, they can be chosen to be upper and lower bounded.) Alternative a function f is l.s.c. iff its epigraph is closed or iff its level sets $\{x : f(x) \leq \alpha\}$ are closed.

If $c : \mathsf{X}\times\mathsf{Y} \to [0,\infty)$ is lower semi continuous, also the mapping

$$\pi \to \int c\, d\pi$$

is lower semicontinuous (on $\mathcal{P}(Z)$) since it is a supremum of the continuous bounded functions $\pi \to \int c \wedge n\, d\pi, n = 1, 2, \ldots$.

Note also that a lower semicontinuous function attains its infimum on every compact set. From these observations we obtain:

Theorem 1.16. *Assume that $c : \mathsf{X}\times\mathsf{Y} \to [0,\infty]$ is lower semi-continous and bounded from below. Then there exists a minimizer π^* to* (KP)*, i.e. $\pi^* \in \arg\min_{\pi\in\mathsf{Cpl}(\mu,\nu)} \int c d\pi$.*

As a consequence of the relaxation of the Monge problem to the Kantorovich problem we can guarantee the existence of optimal couplings in some generality. However, there are several natural questions. For instance,

- When is the optimal coupling unique?
- What is the relationship between (KP) and (MP)?
- Can we characterize the structure of optimal couplings? Are there necessary/sufficient conditions for optimality?

A powerful tool to answer these questions lies in the notion of cyclical monotonicity (more precisely, c-cyclical monotonicity) and the dual problem.

1.3. **The dual problem and characterization of optimal couplings.** A fundamental insight of Kantorovich was the transport problem admits a *dual formulation*:

Definition 1.17 (Dual problem)**.** *For $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ the dual problem is to maximize*

$$\int \varphi(x) d\mu(x) + \int \psi(y) d\nu(y)$$

over $\varphi \in C_b(\mathsf{X}), \psi \in C_b(\mathsf{Y})$ such that $\varphi(x) + \psi(y) \leq c(x,y)$ for all $(x,y) \in \mathsf{X}\times\mathsf{Y}$. We denote the maximal value by $D_c^K := D_c^K(\mu,\nu)$.

We will often write $\varphi \oplus \psi$ for the function given by

$$\varphi \oplus \psi(x, y) = \varphi(x) + \psi(y), \quad (x, y) \in \mathsf{X} \times \mathsf{Y}.$$

It is often useful allow for the larger set of dual candidates

$$\mathcal{D}(c) = \{(\varphi, \psi) : \varphi : X \to [-\infty, \infty), \psi : Y \to [-\infty, \infty), \varphi \in L^1(\mu), \psi \in L^1(\nu), \varphi \oplus \psi \leq c\}$$

which has the advantage that it will be easier to find maximizers for the dual problem.

It is immediate that $D_c^M \leq P_c^M$ since for any candidates φ, ψ, π it follows from the marginal constraint on π that

$$\int \varphi d\mu + \int \psi d\mu = \int (\varphi(x) + \psi(y))\pi(dx, dy) \leq \int c \, d\pi.$$

We will say that *duality* holds if $D_c^M = P_c^M$.

Remark 1.18. Let φ, ψ be integrable with $\varphi \oplus \psi \leq c$. Then $\mu(\varphi) + \nu(\psi) = \pi(c)$ if and only if duality holds and φ, ψ dual maximizers. In this case $\pi(\{c = \varphi \oplus \psi\}) = 1$, in fact it is straightforward (please check for yourself) that $\bar{\pi} \in \mathsf{Cpl}(\mu, \nu)$ is optimal if and only if $\bar{\pi}(\{c = \varphi \oplus \psi\}) = 1$. Theorem 1.30 below provides relatively general conditions which guarantee that we are in this situation.

For the understanding of the transport problem it is extremely useful that duality holds:

Theorem 1.19 (Duality)**.** *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ and $c : \mathsf{X} \times \mathsf{Y} \to [0, \infty]$ be l.s.c. Then,*

$$P_c^K = \inf_{q \in \mathsf{Cpl}(\mu,\nu)} \int c \, dq = \sup\{\mu(\varphi)+\nu(\psi) : \varphi \in C_b(\mathsf{X}), \psi \in C_b(\mathsf{Y}), \varphi(x)+\psi(y) \leq c(x, y)\} = D_c^k.$$

A relatively simple approach is based on the following result from convex analysis:

Theorem 1.20 (see e.g. [Str85, Thm. 45.8] or [AH96, Thm. 2.4.1])**.** *Let K, L be convex subsets of vector spaces H_1 resp. H_2, where H_1 is locally convex and let $F : K \times L \to \mathbb{R}$ be given. If*

(1) K is compact,
(2) $F(\cdot, y)$ is continuous and convex on K for every $y \in L$,
(3) $F(x, \cdot)$ is concave on L for every $x \in K$

then

$$\sup_{y \in L} \inf_{x \in K} F(x, y) = \inf_{x \in K} \sup_{y \in L} F(x, y).$$

Proof in the compact case. We will give a proof under the additional assumption that X, Y are compact spaces and that c is continuous bounded. Write

$$\chi(\pi) = \begin{cases} 0 & \text{if } \pi \in \mathsf{Cpl}(\mu, \nu) \\ \infty & \text{else.} \end{cases} = \sup_{(\varphi,\psi) \in C_b(\mathsf{X}) \times C_b(\mathsf{Y})} \mu(\varphi) + \nu(\psi) - \int (\varphi(x) + \psi(y)) d\pi(x, y).$$

As $\mathcal{P}(X \times Y)$ is compact for compact spaces X, Y we can apply Theorem 1.20 to interchange inf and sup to obtain

$$\begin{aligned}
&\inf_{\pi \in \mathsf{Cpl}(\mu,\nu)} \int c \, d\pi = \inf_{\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})} \int c \, d\pi + \chi(\pi) \\
&= \inf_{\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})} \sup_{(\varphi,\psi) \in C_b(\mathsf{X}) \times C_b(\mathsf{Y})} \int c(x, y) - \varphi(x) - \psi(y) \, d\pi + \mu(\varphi) + \nu(\psi) \\
&= \sup_{(\varphi,\psi) \in C_b(\mathsf{X}) \times C_b(\mathsf{Y})} \inf_{\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})} \int c(x, y) - \varphi(x) - \psi(y) \, d\pi + \mu(\varphi) + \nu(\psi) \\
&= \sup_{(\varphi,\psi) \in C_b(\mathsf{X}) \times C_b(\mathsf{Y})} \inf_{(x,y) \in \mathsf{X} \times \mathsf{Y}} c(x, y) - \varphi(x) - \psi(y) + \mu(\varphi) + \nu(\psi) \\
&= \sup_{(\varphi,\psi) \in C_b(\mathsf{X}) \times C_b(\mathsf{Y}), \varphi \oplus \psi \leq c} \mu(\varphi) + \nu(\psi).
\end{aligned}$$

□

From the above proof one can first obtain duality for general Polish spaces X, Y by approximating μ, ν with compactly supported measures. It is then easy to extend from the case of a continuous bounded cost function to a l.s.c. cost function $c : \mathsf{X} \times \mathsf{Y} \to [0, \infty]$ by approximating c from below with continuous bounded functions.

Other proofs of transport duality often use further tools from convex analysis such as the Fenchel-Moreau theorem.

Below we give another proof of duality which uses the notion of c-cyclical monotonicity which is of interest in its own right. It allows to characterize optimality of transport plans through a "combinatoric optimality property" of its support set:

Definition 1.21. *Let $\Gamma \subseteq \mathsf{X} \times \mathsf{Y}$ and $c \colon \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$. We call Γ* c-cyclically monotone *if for all $n \in \mathbb{N}$ and sequences $(x_1, y_1), \dots, (x_n, y_n) \in \Gamma$, we have with the convention $y_{n+1} := y_1$*

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{i+1}). \tag{1.3}$$

Let $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ and $c \colon \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$. We call π c-cyclically monotone *if there is a c-cyclically monotone set Γ with $\pi(\Gamma) = 1$.*

This is a generalization of the notion of *cyclical monotonicity* which arose in convex analysis and corresponds to c-cyclical monotonicity in the special case where $c(x, y) = -xy$, $x, y \in \mathbb{R}^d$.

Note that if c is continuous than the closure of every c-cyclically monotone set is again c-cyclically monotone. In this case a transport plan is c-cyclically monotone precisely if $\operatorname{supp} \pi$ is c-cyclically monotone.

Lemma 1.22 (c-cyclical monotonicity: necessary)**.** *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$, and $c \in C(\mathsf{X} \times \mathsf{Y})$. If $\pi \in Cpl(\mu, \nu)$ is an optimal coupling (w.r.t. the cost function c) with finite value, then* $\operatorname{supp}(\pi)$ *is c-cyclically monotone (so in particular π is c-cyclically monotone).*

In fact c-cyclical monotonicity is also sufficient for optimality. However the proof is more involved and will be given later.

Proof of Lemma 1.22. Let $\pi^* \in \mathrm{Cpl}(\mu, \nu)$ be an optimizer for $V_c(\mu, \nu)$ and assume that there exist $n \in \mathbb{N}$ and $(x_1, y_1), \dots, (x_n, y_n) \in \operatorname{supp}(\pi)$ such that

$$\sum_{i=1}^{n} c(x_i, y_i) > \sum_{i=1}^{n} c(x_i, y_{i+1}).$$

By continuity of c there are neighbourhoods U_i of x_i and V_j of y_j such that for all $u_i \in U_i, v_j \in V_j, 1 \leq i, j \leq n$

$$\sum_{i=1}^{n} c(u_i, v_i) > \sum_{i=1}^{n} c(u_i, v_{i+1}) \tag{1.4}$$

In the next step, we will use this property to construct a competitor $\bar{\pi}$ of π with strictly lower transport cost. To this end, consider $m_i := \pi(U_i \times V_i)$ (which is positive as $\operatorname{supp}(\pi) \ni (x_i, y_i) \in U_i \times V_i$) and $\frac{1}{m_i}\pi|_{U_i \times V_i} \in \mathrm{Cpl}(\mu_i, \nu_i)$ (where (μ_i, ν_i) simply denote the renormalized marginals of $\pi^*|_{U_i \times V_i}$). We define

$$\bar{\pi} := \pi + \frac{\min_j m_j}{n} \sum_{i=1}^{n} \mu_i \otimes \nu_{i+1} - \frac{1}{m_i}\pi^*|_{U_i \times V_i}.$$

Since $\pi|_{U_i \times V_i} - \frac{\min_j m_j}{m_i}\pi|_{U_i \times V_i} \geq 0$, we have that

$$\pi - \sum_{i=1}^{n} \frac{\min_j m_j}{n\, m_i}\pi|_{U_i \times V_i} \in \mathrm{Cpl}\left(\mu - \frac{\min_j m_j}{n} \sum_i \mu_i, \nu - \frac{\min_j m_j}{n} \sum_i \nu_i\right),$$

is a positive measure and observe $\bar{\pi} \in \mathrm{Cpl}(\mu, \nu)$. By (1.4) we find $\int c\, d\bar{\pi} < \int c\, d\pi$ which contradicts optimality of π. □

Given a candidate pair $(\varphi, \psi) \in \mathcal{D}(c)$ we can always improve it by replacing φ (which satisfies $\varphi(x) \leq c(x,y) - \psi(y)$) by

$$\tilde{\varphi}(x) := \inf_y c(x,y) - \psi(y).$$

Then, $(\tilde{\varphi}, \psi) \in \mathcal{D}(c)$ and since $\varphi \leq \tilde{\varphi}$ it follows that $\mu(\varphi) \leq \mu(\tilde{\varphi})$ so that the pair $(\tilde{\varphi}, \psi)$ yields a higher value in the dual problem. Note that $\tilde{\varphi}$ is the biggest function f such that $f(x) + \psi(y) \leq c(x,y)$. Similarly, we can replace ψ by $\tilde{\psi}$ defined by

$$\tilde{\psi}(y) := \inf_x c(x,y) - \tilde{\varphi}(x)$$

producing an even better candidate for the dual problem. This motivates the following definition

Definition 1.23 (c-transform)**.** *Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ be a Borel measurable cost function. For a function $\varphi : \mathsf{X} \to \mathbb{R}$ we define its c-transform (also called c-conjugate function) $\varphi^c : \mathsf{Y} \to \mathbb{R}$ by*

$$\varphi^c(y) := \inf_{x \in \mathsf{X}} c(x,y) - \varphi(x). \tag{1.5}$$

Analogously, we define the c-transform of $\psi : \mathsf{Y} \to \mathbb{R}$ by

$$\psi^c(x) := \inf_{y \in \mathsf{Y}} c(x,y) - \psi(y).$$

We say that a function $\psi : \mathsf{Y} \to \mathbb{R}$ is c-concave if $\psi = \varphi^c$ for some φ (and analogously for $\psi : \mathsf{Y} \to \mathbb{R}$).

Formally the definition of the c-transform of a function χ depends on whether the domain of χ is X or Y. In most applications we will have $X = Y$ and c will be symmetric.

The c-transform of a function χ 'regularizes' χ: if c is continuous, then (1.5) is an infimum of continuous functions and thus upper semicontinuous (i.e. its negative is l.s.c.). In particular all c-concave functions are then u.s.c. and in particular measurable.

From the previous considerations it is clear that in the dual problem we can restrict to pairs of c-concave functions. One could go on trying to "improve" these functions, however, we have the following result:

Lemma 1.24. *Suppose that c is real valued. For any $\varphi : \mathsf{X} \to \mathbb{R} \cup \{-\infty\}$ it holds that $\varphi^{cc} := (\varphi^c)^c \geq \varphi$. We have $\varphi^{cc} = \varphi$ iff φ is c-concave (i.e. for any $\varphi : \mathsf{X} \to \mathbb{R}$ it holds that $\varphi^{ccc} := ((\varphi^c)^c)^c = \varphi^c$.); in general, φ^{cc} is the smallest c-concave function larger than φ.*

Proof. Exercise. □

Importantly, c-concave functions share many properties of concave (convex) functions. Indeed, the terminology is inspired by the Legendre transform

$$\varphi^*(y) := \sup_{y \in \mathsf{X}} xy - \varphi(x)$$

from convex analysis. Specifically, up to a sign switch, the Legendre transform corresponds to c transform for the for cost function $c(x,y) = -xy$, see Example 1.27 below.

Definition 1.25. *Let $\varphi : \mathsf{X} \to [-\infty, \infty)$ be c-concave. Its c-superdifferential is defined as*

$$\partial^c \varphi = \{(x,y) \in \mathsf{X} \times \mathsf{Y} : \varphi(x) + \varphi^c(y) = c(x,y)\}.$$

We also write[2]

$$\partial^c \varphi(x) = \{y : (x,y) \in \partial^c \varphi\} = \{y : \varphi(x) + \varphi^c(y) = c(x,y)\} = \arg\min_y c(x,y) - \varphi^c(y). \tag{1.6}$$

Similarly, we define the c-superdifferential of a c-concave function $\psi : \mathsf{Y} \to [-\infty, \infty)$.

Remark 1.26. Note that if φ is a c-concave function and φ, c are continuous bounded (to avoid integrability issues), then any transport plan concentrated on $\partial^c \varphi$ is optimal between its marginals.

[2]It is worthwhile to reflect a bit about the equalities in (1.6).

Example 1.27. a) Let $\mathsf{X} = \mathsf{Y}$, $c(x,y) = d(x,y)$ be a distance. Then φ is c-concave ("d-concave") if and only if φ is 1-Lipschitz and $\varphi^c = -\varphi$.

Proof. $\varphi^d(\cdot) = \inf_x d(x,\cdot) - \varphi(x)$ is an infimum over 1-Lipschitz functions and thus 1-Lipschitz, hence any d-concave function is 1-Lipschitz. If φ is 1-Lipschitz, then $\varphi(y) - \varphi(x) \le d(x,y)$ implies $-\varphi(x) \le \inf_y d(x,y) - \varphi(y) = \varphi^c(x)$ and here equality is attained for $x = y$. Conversely, it is immediate that any 1-Lipschitz function with $\varphi^d(x) = -\varphi(x)$ is d-concave. □

b) $\mathsf{X} = \mathsf{Y} = \mathbb{R}^n$, $c(x,y) = -x \cdot y$, the standard Euclidean inner product. In this case $\varphi^c = -((-\varphi)^*)$, where

$$\chi^*(y) := \sup_y xy - \chi(x)$$

denotes the Legendre-transform. In particular, φ is c-concave iff φ is concave and u.s.c. In convex analysis the subdifferential $\partial\chi(x)$ of a convex function $\chi : \mathbb{R}^n \to (-\infty, \infty]$ in a point x equals

$$\partial\chi(x) = \{y : y(x' - x) \le \chi(x') - \chi(x), \forall x \in \mathbb{R}^n\} = \arg\max_y xy - \chi(x),$$

with the superdifferential (or supergradient) being defined analogously. In particular, the c-superdifferential of φ is precisely the classical superdifferential $\partial\varphi$ of φ from convex analysis.

The link between optimal transport and convex analysis becomes even more transparent if we switch to *maximization* in the primal transport problem and to *minimization* in the dual problem, i.e. consider the optimization problems

$$P_c^K = \sup_{\pi \in \mathsf{Cpl}(\mu,\nu)} \int xy \, d\pi(x,y), \; D^k = \inf\{\mu(\varphi) + \nu(\psi) : \varphi(x) + \psi(y) \ge xy\}. \tag{1.7}$$

In this formulation, the c-transform becomes precisely the convex conjugate and c-convexity is just ordinary convexity.

c) Put $c(x,y) = \frac{1}{2}|x-y|^2$. Then, φ is c-concave iff $\bar{\varphi}(x) := \frac{|x|^2}{2} - \varphi(x)$ is convex and l.s.c.

Rockafeller's theorem (see e.g. [Vil03, Theorem 2.27]) characterizes when a set $\Gamma \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is the subgradient of a convex function through cyclical monotonicity.

We spell it out for c-concave functions:

Theorem 1.28. *Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$. A set $\Gamma \subseteq \mathsf{X} \times \mathsf{Y}$ is c-cyclically monotone iff $\Gamma \subseteq \partial^c\varphi$ for some c-concave function φ.*

Proof. First observe that $\partial^c\varphi$ is c-cyclical monotone. Indeed, for any $n \in \mathbb{N}$ and tuples $(x_1,y_1),\dots,(x_n,y_n) \in \partial^c\varphi$ with $y_{n+1} := y_1$ by definition of c-concavity that

$$\sum_{i=1}^n c(x_i,y_i) = \sum_{i=1}^n \varphi(x_i) + \varphi^c(y_i) = \sum_{i=1}^n \varphi(x_i) + \varphi^c(y_{i+1}) \le \sum_{i=1}^n c(x_i,y_{i+1}).$$

The converse direction is a bit technical and we first try to give some motivation for the proof. For this we suppose that we have found a pair of functions a, b such that $a \oplus b \le c$ with equality on the set Γ. Let further $(x_0,y_0) \in \Gamma$ and assume that $a(x_0) = 0$. Consider $(x_1,y_1),\dots,(x_n,y_n) \in \Gamma$ and let $x \in \mathsf{X}$ be arbitrary. Then we have

$$\underbrace{c(x,y_n) - c(x_n,y_n)}_{\ge a(x)+b(y_n)-a(x_n)-b(y_n)} + \underbrace{c(x_n,y_{n-1}) - c(x_{n-1},y_{n-1})}_{\ge a(x_n)+b(y_{n-1})-a(x_{n-1})-b(y_{n-1})} + \dots + \underbrace{c(x_1,y_0) - c(x_0,y_0)}_{\ge a(x_1)+b(y_0)-a(x_0)-b(y_0)} \ge a(x). \tag{1.8}$$

Hence fixing $x \in \mathsf{X}$ and $(x_0,y_0) \in \Gamma$, (1.8) gives an upper bound of $a(x)$ for any choice of $(x_1,y_1),\dots,(x_n,y_n) \in \Gamma$. Taking the infimum over all such choices will us thus give a natural ("largest possible") candidate for a dual function.

That is, to rigorously start our proof we set $\varphi(x_0) = 0$ for $x \in \mathsf{X}$

$$\begin{aligned}\varphi(x) := \inf\{&c(x,y_n) - c(x_n,y_n) + c(x_n,y_{n-1}) - c(x_{n-1},y_{n-1}) + \dots + c(x_1,y_0) - c(x_0,y_0) : \\ &n \in \mathbb{N}; (x_1,y_1),\dots,(x_n,y_n) \in \Gamma\}\end{aligned} \tag{1.9}$$

Since c is real valued and $\Gamma \neq \emptyset$ we have $\varphi < \infty$ on X. Note also that c-monotonicity implies that $\varphi(x_0) \geq 0$, and that equality is attained by choosing $n = 1$ and $(x_1, y_1) = (x_0, y_0)$.

Next, writing

$$-\psi(y) := \inf\{ \quad - c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + \ldots + c(x_1, y_0) - c(x_0, y_0) : \\ n \in \mathbb{N}; (x_1, y_1), \ldots, (x_n, y_n) \in \Gamma, y_n = y\}$$

we see that

$$\varphi(x) = \inf_{y \in \mathsf{Y}} c(x, y) - \psi(y) = \psi^c(x).$$

(Observe, that $\psi(y) > -\infty$ iff $y \in \mathrm{proj}_{\mathsf{Y}}(\Gamma)$, i.e. there is $x \in \mathsf{X}$ such that $(x, y) \in \Gamma$.)

To show that $\Gamma \subseteq \partial^c \varphi$ it is sufficient to show $\varphi(x) + \varphi^c(y) \geq c(x, y)$ on Γ since the other inequality follows from c-concavity. Since $\varphi^c = \psi^{cc} \geq \psi$ (see Lemma 1.24) it is enough to show $\varphi(x) + \psi(y) \geq c(x, y)$ on Γ. So pick $\varepsilon > 0$ and $(x, y) \in \Gamma$. Since $\varphi = \psi^c$ there is some $\tilde{y} \in \mathrm{proj}_{\mathsf{Y}}(\Gamma)$ such that $c(x, \tilde{y}) - \psi(\tilde{y}) < \varphi(x) + \varepsilon$. From the definition of ψ it follows that $-\psi(y) \leq -c(x, y) + c(x, \tilde{y}) - \psi(\tilde{y})$ (estimating the inf with a particular choice of tuples approximating $-\psi(\tilde{y})$). Together this gives $-\psi(y) \leq -c(x, y) + c(x, \tilde{y}) - \psi(\tilde{y}) < -c(x, y) + \varphi(x) + \varepsilon$. Since $\varepsilon > 0$ is arbitrary this proves the claim. □

Remark 1.29. If c is not continuous, the functions φ and φ^c constructed in Theorem 1.28 may not be Borel measurable. When c is Borel measurable and Γ a Borel subset of $\mathsf{X} \times \mathsf{Y}$, then one can show that φ and φ^c are universally measurable. This means, for any marginals μ and ν there exist Borel measurable functions $\tilde{\varphi} \colon \mathsf{X} \to [-\infty, \infty)$ and $\tilde{\psi} \colon \mathsf{Y} \to [-\infty, \infty)$ such that $\tilde{\varphi} = \varphi$ μ-a.s. and $\tilde{\psi} = \varphi^c$ ν-a.s.

Theorem 1.28 allows us to prove the following result, sometimes referred to as fundamental theorem of optimal transport, or characterization of optimizers, or monotonicity principle of OT.

Theorem 1.30 (Fundamental theorem of OT)**.** *Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ be continuous and assume that that $|c(x, y)| \leq a(x) + b(y)$ for some $a \in L^1(\mu), b \in L^1(\nu)$. Let $\pi \in \mathsf{Cpl}(\mu, \nu)$. Then the following are equivalent:*

i) π is an optimal coupling;
ii) the support $\mathrm{supp}(\pi)$ of π is c-cyclical monotone;
iii) there exists a c-concave function φ with s.t. $\mathrm{supp}(\pi) \subseteq \partial^c \varphi$ (i.e. $\pi(\{(x, y) : \varphi(x) + \varphi^c(y) = c(x, y)\} = 1$).

In this case $\varphi \in L^1(\mu), \varphi^c \in L^1(\nu)$, duality holds and φ, φ^c are maximizers of the dual problem. Furthermore every primal optimizer is supported by $\partial^c \varphi$.

Proof. We want to replace c by the mapping $(x, y) \mapsto c(x, y) + a(x) + b(y)$ to reduce the argument to the case where $c \geq 0$. Clearly this works provided a, b are continuous but if a, b are merely measurable, the new cost function might no longer be continuous. This problem can be avoided with the help of a trick from descriptive set theory. [Kec95, Theorem 08.15] asserts that one can refine the topologies of X, Y to (still) Polish topologies with respect to which the functions a, b are continuous. Doing this, we can proceed with the proof under the assumption that c is continuous and non-negative. (Note that is in general *not* possible to equip X, Y with finer Polish topologies such that a given Borel function defined on $\mathsf{X} \times \mathsf{Y}$ becomes continuous)

We now proceed with the proof.

i) ⇒ ii): This follows by Lemma 1.22.
ii) ⇒ iii): This follows by Theorem 1.28.
iii) ⇒ i): As

$$\varphi \oplus \varphi^c \leq c \leq a \oplus b$$

it follows that φ_+ and ψ_+ (where $\psi := \varphi^c$) are integrable. Given a function χ we write $\chi^{(n)} := (\chi \vee n) \wedge n$. Then $\varphi^{(n)} \to \varphi$, $\int \varphi^{(n)} \, d\mu \to \int \varphi \, d\mu$ and analogously for ψ. Furthermore

$\varphi^{(n)} \oplus \psi^{(n)} \leq c$ with equality holding π-a.s. Thus

$$\int c\, d\pi = \int \lim_n \varphi^{(n)} \oplus \psi^{(n)}\, d\pi = \lim_n \int \varphi^{(n)} \oplus \psi^{(n)}\, d\pi =$$
$$\lim_n \int \varphi^{(n)}\, d\mu + \lim_n \int \psi^{(n)}\, d\nu = \int \varphi\, d\mu + \int \psi\, d\nu.$$

Thus φ, ψ are (integrable) dual maximizers, π is a primal maximizer and all primal maximizers are concentrated on $\partial\varphi = \{\varphi \oplus \psi = c\}$.

We conclude by Remark 1.18. □

Remark 1.31. As mentioned above, duality for l.s.c. $c : \mathsf{X}\times\mathsf{Y} \to [0,\infty]$ claimed in Theorem 1.19 follows directly from Theorem 1.30 by approximating c from below with continuous bounded functions. Duality for upper bounded measurable functions was established in [Kel84] and duality for lower bounded, finitely valued cost functions was established in [BS11]. Optimal transport plans are c-cyclically monotone for all measurable $c : \mathsf{X}\times\mathsf{Y} \to [0,\infty]$ while c-cyclically monotone transport plans are optimal for measurable $c : \mathsf{X}\times\mathsf{Y} \to [0,\infty]$ provided that $\{c = \infty\}$ is contained in the union of a closed set and a $\mu\otimes\nu$-null set [BGMS09a, Bei15] but not for arbitrary l.s.c. $c : \mathsf{X}\times\mathsf{Y} \to [0,\infty]$, see [AP03].

Example 1.32. A standard example where duality does not hold (i.e. there is a 'duality gap') is the following: Let $X, Y = [0,1], \mu = \nu = \mathsf{Leb}_{|[0,1]}$ and

$$c(x,y) = \begin{cases} 0 & x < y \\ 1 & x = y \\ \infty & \text{else} \end{cases}.$$

Then the primal problem has the value 1 while the dual problem has the value 0.

Remark 1.33. An important consequence of Theorem 1.30 is that the optimality of a given coupling depends solely on geometric properties of its support (and it does not matter how the mass is distributed over the support). In particular, if π is optimal and $\tilde\pi$ another probability with $\mathrm{supp}(\tilde\pi) \subseteq \mathrm{supp}(\pi)$, then it is an optimal coupling between its marginals. For instance, a restriction of an optimal coupling is optimal (between its marginals).

One can argue similarly for transport maps T. If there exists a c-concave functions φ such that for all $x \in \mathsf{X}$ it holds that $T(x) \in \partial^c\varphi(x)$, then for any $\mu \in \mathcal{P}(\mathsf{X})$ the map T is optimal between μ and $T(\mu)$ (up to integrability isssues of the cost c w.r.t. μ and ν). Hence, it makes sense to say that T is an optimal transport map without specifying any measure.

Definition 1.34. *A c-concave function φ such that the pair (φ, φ^c) is a maximizing pair for the dual problem is called a c-concave Kantorovich potential, or Kantorovich potential, of the measures μ, ν.*

1.4. **Consequences of Theorem of 1.30 – Brenier's Theorem.** As first consequence of Theorem of 1.30 (and Example 1.27) we obtain

Corollary 1.35 ((Kantorovich-Rubinstein formula))**.** *Let $\mathsf{X} = \mathsf{Y}$, $c(x,y) = d(x,y)$ a distance and assume that μ, ν have finite first moment.*[3] *Then*

$$\inf_{\pi\in\mathsf{Cpl}(\mu,\nu)} \int c\, d\pi = \sup_{\varphi \text{ is 1-Lipschitz}} \int \varphi(d\mu - d\nu)$$

and inf, sup *are attained. If φ is a maximizer of the dual problem, then π is optimal if and only if*

$$\pi(\{(x,y) : \varphi(x) - \varphi(y) = d(x,y)\}) = 1.$$

Note that the right hand side immediately implies that $\inf_{\pi\in\mathsf{Cpl}(\mu,\nu)} \int cdq =: W_1(\mu,\nu)$ is a distance. Moreover, we can use this formula and extend this to non-probability measures as well. In this case for any finite non-negative measure η it holds that $W_1(\mu+\eta, \nu+\eta) = W_1(\mu,\nu)$.

[3] I.e. $\int d(x_0,x)\, d\mu(x) < \infty$ for some and then any $x_0 \in \mathsf{X}$.

By Theorem 1.30 we know that an optimal coupling is concentrated on the superdifferential $\partial^c\varphi$ of a c-concave function φ. In particular, if we can show that for μ-a.e. $x \in \mathsf{X}$ the set $\partial^c\varphi(x)$ is single-valued any optimal coupling π needs to be induced by a transport map.

We will first use this observation in the context of the maximization version of the transport problem for the cost function $c(x, y) = xy, x, y \in \mathbb{R}^d$. We assume that μ, ν have finite second moments such that xy is dominated by $a \oplus b$ where $a(x) = b(x) = x^2$ is μ- and ν-integrable so that Theorem 1.30 is applicable. Thus $\pi \in \mathsf{Cpl}(\mu, \nu)$ is optimal for

$$P_c^K = \sup_{\pi \in \mathsf{Cpl}(\mu,\nu)} \int xy \, d\pi(x, y). \tag{1.10}$$

if and only if there exists a l.s.c. convex $\varphi : \mathbb{R}^n \to [-\infty, \infty)$ such that π is concentrated on the subgradient

$$\partial\varphi = \{(x, y) : \varphi(x) + \varphi^*(y) = xy\}.$$

By Rademacher's theorem, every convex function is almost surely differentiable (where it is finite). From this we obtain a particularly satisfying characterization of the dual maximizer in the case of absolutely continuous marginal μ:

Corollary 1.36. *Assume that $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ have finite second moment and that $\mu \ll \mathsf{Leb}$. Then there is a unique optimal $\pi^* \in \mathsf{Cpl}(\mu, \nu)$ for the problem* (1.10)*. Given a Kantorovich potential φ, the optimizer π^* is of the form $(Id, \nabla\varphi)_\#\mu$. In particular there exists a unique optimal transport mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ and T is the gradient of a convex function. Moreover T is the only gradient of a convex function which pushes μ to ν.*

Proof. From Theorem 1.30 we know that there exists a convex potential φ and that $\pi(\partial\varphi) = 1$ for any primal optimizer π.

Since a convex function is locally Lipschitz (on the set where it is finite) it is differentiable Leb-a.e. by Rademacher's theorem so that $\nabla\varphi(x)$ exists μ-a.e. Furthermore, on the set of differentiability of φ it holds that $\nabla\varphi(x) = y$ iff $y \in \partial\varphi(x)$. Hence, any optimal coupling π is concentrated on the graph of $\nabla\varphi$.

This immediately implies that the optimal coupling is unique. Indeed, assume there are two optimal couplings π_1, π_2 both concentrated on the graph of some maps T_1, T_2. Then, $\pi_3 = \frac{1}{2}(\pi_1 + \pi_2)$ is optimal again by linearity of the Kantorovich problem. By the first part of the theorem, π_3 has to be concentrated on the graph of some function. By construction it is concentrated on the union of the graphs of T_1 and T_2. This is only possible if $T_1 = T_2$ μ-a.s.

For the last statement, note that any $\nabla\bar{\varphi}$ with $\nabla\bar{\varphi}(\mu) = \nu$ is optimal since the graph defines a c-cyclically monotone set. By uniqueness, we can conclude. □

Remark 1.37. In probabilistic terms, (1.10) reads

$$\sup\{\mathbb{E}[XY] : X \sim \mu, Y \sim \nu\}. \tag{1.11}$$

The covariance of random variables X, Y is given by $\mathrm{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, so problem (1.12) is further equivalent to finding the coupling of X and Y which maximizes the mutual covariance, i.e.

$$\sup\{\mathrm{cov}(X, Y) : X \sim \mu, Y \sim \nu\}. \tag{1.12}$$

Remark 1.38 (Uniqueness). Observe, that we proved uniqueness by showing that any optimal coupling has to satisfy a property which is not stable under convex combinations (here being concentrated on the graph of a function). This is essentially the only way we can prove uniqueness of optimal couplings.

Being the gradient of a convex function $\varphi : \mathbb{R}^n \to [-\infty, \infty)$ can be seen as a natural n-dimensional analogue of be an increasing (non-decreasing) function.

Definition 1.39. *Let $\mu \in \mathcal{P}(\mathbb{R}^n)$. A function $T : \mathbb{R}^n \to \mathbb{R}^n$ is called* monotone *if there exists a convex function $\varphi : \mathbb{R}^n \to [-\infty, \infty)$ which is μ-a.s. differentiable and satisfies $T = \nabla\varphi$. μ-a.s.*

Rather then maximizing w.r.t. the cost function $(x, y) \mapsto xy$, we can consider the minimization problem for the quadratic distance cost function $(x, y) \mapsto \frac{1}{2}|x - y|^2$. In this terms, Corollary 1.36 amounts

Theorem 1.40 (Brenier's Theorem)**.** *Let* $\mathsf{X} = \mathsf{Y} = \mathbb{R}^n$, $c(x, y) = \frac{1}{2}|x - y|^2$. *Assume that* $\mu, \nu \in \mathcal{P}(\mathsf{X})$ *have finite second moment and that* $\mu \ll \mathsf{Leb}$. *Then, there is a unique optimal coupling* $\pi^* \in \mathsf{Cpl}(\mu, \nu)$. *This optimizer is of the form* $\pi* = (\mathrm{Id}, \nabla\bar{\varphi})(\mu)$ *for some convex function* $\bar{\varphi} : \mathsf{X} \to \mathbb{R}$.[4]

Moreover, there exists a μ*-a.e. unique map* T *of the form* $T = \nabla\bar{\varphi}$ *such that* $T(\mu) = \nu$ *and for any convex* $\tilde{\varphi}$, *the map* $\tilde{T} = \nabla\tilde{\varphi}$ *is optimal between* μ *and* $\tilde{T}(\mu)$.

Remark 1.41. One can relax the condition $\mu \ll \mathsf{Leb}$ a little bit. From the proof it is clear that it is sufficient to assume that μ does not charge the set of non-differentiability points of convex functions. For instance it would be sufficient to assume that μ does not charge any set of Hausdorff dimension less or equal than $n - 1$. We will refer to this property by saying μ does not charge *small sets*.

Remark 1.42. The question of regularity of the optimal transport maps is an interesting story in itself which goes way beyond the scope of this course. For instance if μ and ν have α-Hölder continuous density and convex support then the optimal map is $C^{1,\alpha}$. If the densities are only bounded from above and below the optimal maps are only C^α for some $\alpha < 1$.

In Brenier's theorem we could show that $\partial^c\varphi(x)$ is single-valued μ a.s. by playing everything back to convex functions. However, the two properties that we really needed are the following: If φ is c-concave and $(x, y) \in \partial^c\varphi$ then

- $\varphi, c(\cdot, y)$ are differentiable at x (μ-a.e.) with $\nabla\varphi(x) = \nabla_x c(x, y)$
- $\nabla_x c(x, \cdot)$ is invertible.

Assuming differentiability of φ and c, the second part of the first item can be argued via c-concavity ($\varphi(x) = c(x, y) - \varphi^c(y), \varphi(z) \le c(z, y) - \varphi^c(y)$ all $z \in \mathsf{X}$). If the second item holds, then $y = (\nabla_x c(x, y))^{-1}(\nabla\varphi(x))$ so that we can write down a map $x \mapsto y$ with $(x, y) \in \partial^c\varphi$ implying uniqueness as for Brenier's result.

Let us consider two cases:

i) $c(x, y) = h(x - y)$, with h superlinear and strictly convex (e.g. $h(x) = |x|^2/2$)
ii) $c(x, y) = h(|x - y|)$, with h strictly concave.

Let us start with *i*). Then, $\nabla_x c(x, y) = \nabla h(x - y)$ and ∇h is defined (a.e.) and invertible with $(\nabla h)^{-1} = \nabla h^*$ where $h^*(y) = \sup_x x \cdot y - h(x)$ is the Legendre transform of h. This means that $\nabla_x c(x, y) = u \Leftrightarrow y = x - \nabla h^*(u)$. Thus, if φ is c-concave and differentiable at x, then

$$\partial^c\varphi(x) = \{x - \nabla h^*(\nabla\varphi(x))\}.$$

In this situation one can show that a c-concave function is locally Lipschitz on the interior of the set where it is finite (short $\mathsf{int}(\mathsf{Dom}(\varphi))$). Then, Rademacher's theorem implies that φ is differentiable Leb-a.e. on $\mathsf{int}(\mathsf{Dom}(\varphi))$. Summarizing we obtain

Theorem 1.43 (Gangbo-McCann, [GM96])**.** *Let* $\mathsf{X} = \mathsf{Y} = \mathbb{R}^n$, $c(x, y) = h(x - y)$ *where* h *is superlinear, strictly convex and bounded from below. Let* $\mu, \nu \in \mathcal{P}(\mathsf{X})$ *with* $\mu \ll \mathsf{Leb}$. *Assume that* $P_c^K < \infty$. *Then, there exists a unique optimal coupling* q^*. *It is a Monge coupling induced by a transport map of the form* $T(x) = x - \nabla h^*(\nabla\varphi(x))$ *for some c-concave function* φ.

Furthermore, any map T *of this form is optimal between* μ *and* $T(\mu)$.

Let us turn to item ii) so $c(x, y) = h(|x - y|)$ with $h : \mathbb{R}_+ \to \mathbb{R}$ strictly concave and $h \ge 0$.

Theorem 1.44 (Gangbo-McCann, [GM96])**.** *Assume* $P_c^K < \infty$. *Put* $\mu_0 = (\mu - \nu)_+, \nu_0 = (\mu - \nu)_-, \mu \wedge \nu = \mu - \mu_0 = \nu - \nu_0$. *Then, there is a unique optimal coupling* q^*. *Write*

[4] Explicitly we can take $\bar{\varphi} = \frac{|x|^2}{2} - \varphi(x)$, where φ is a Kantorovich potential.

$q^* = q^*_d + q^*_o$ with $q^*_d = q^*_{|\{(x,x):x\in\mathbb{R}^d\}}$. Then, $q^* = (\mathrm{Id},\mathrm{Id})(\mu\wedge\nu)$ and q_o is a Monge coupling induced by a map T of the form $T(x) = x - \nabla h^*(\nabla\varphi(x))$ μ-a.e. for some c-concave φ.

The crucial idea to prove this result relies on the following observation. Wlog we can assume $c(x,x) = h(0) = 0$. Then the strict concavity of h implies that c is a metric with strict triangular inequality (Exercise!). Then, the common mass has to stay put. Indeed, we have the following result:

Lemma 1.45. *Let $\mu, \nu \in \mathcal{P}(\mathsf{X})$, c a metric on X. Let $q \in \mathsf{Cpl}(\mu,\nu), \mu\wedge\nu = \mu - (\mu-\nu)_+ = \nu - (\nu-\mu)_+$. Then, $q_d \le (\mathrm{Id},\mathrm{Id})(\mu\wedge\nu)$. If c satisfies the strict triangular inequality and q is optimal for c, then there is equality.*

Proof. Exercise. □

Arguing as in the convex case yields the result.

Example 1.46. *(The one-dimensional case)* Let $\mathsf{X} = \mathsf{Y} = \mathbb{R}$, $c(x,y) = h(y-x)$ for some strictly convex h, e.g. $h(r) = |r|^p$, $p > 1$. Pick a c-c.m. set Γ and $(x_i, y_i) \in \Gamma$ for $i = 1, 2$. Wlog we can assume that $y_1 < y_2$. We want to understand whether c-c.m. forces $x_1 \le x_2$ or $x_2 \le x_1$? Note that these are *geometric constraints* on Γ.

Put $a = y_2 - y_1 > 0$. Setting

$$b = y_1 - x_1, \quad d = y_1 - x_2$$

we have

$$b + a = y_2 - x_1, \quad d + a = y_2 - x_2.$$

Since Γ is c-c.m., (1.3) with $n = 2$ implies

$$\begin{aligned} & h(b) + h(d+a) \le h(b+a) + h(d) \\ \Leftrightarrow \quad & h(d+a) - h(d) \le h(b+a) - h(b). \end{aligned}$$

Since h is strictly convex and $a > 0$ the map $x \mapsto h(x+a) - h(x)$ is (strictly) increasing implying $b > d$ and hence $x_1 < x_2$.

Note that this property uniquely determines any coupling π living on Γ to be the quantile coupling/monotone rearrangement between its marginals. More precisely, if π has marginals μ and ν with cumulative distribution functions F_μ and F_ν, then

$$\pi = (F_\mu^{-1}, F_\nu^{-1})(\mathsf{Leb}_{|[0,1]}).$$

Observe, that in this situation we only considered cyclical monotonicity using 2-cycles. A set $\Gamma \subseteq \mathsf{X}\times\mathsf{Y}$ satisfying (1.3) for $n = 2$ is called monotone set. In dimension 1 monotonicity is equivalent to c-cyclical monotonicity for $c(x,y) = h(x-y)$ and h strictly convex. In higher dimensions this is not true any more.

1.5. **Kantorovich-Wasserstein distance and interpolation of probability measures.**

For various applications of optimal transport a key object are the Kantorovich-Wasserstein distances $\mathcal{W}_p$. They inherit various geometric properties of the base space and induce an useful interpolation of probability measures.

We denote the set of probability measures with finite p-th moment by

$$\mathcal{P}_p(\mathsf{X}) = \left\{\mu \in \mathcal{P}(\mathsf{X}) : \int d^p(x,x_0)\,\mu(dx) < \infty, \text{ for some, hence any } x_0 \in \mathsf{X}\right\}.$$

Definition 1.47. *The p-Wasserstein distance $\mathcal{W}_p$ is defined for $\mu,\nu \in \mathcal{P}_p(\mathsf{X})$ as*

$$\mathcal{W}_p(\mu,\nu) = \left(\inf_{\pi\in\mathsf{Cpl}(\mu,\nu)} \int d^p(x,y)\,\pi(dx,dy)\right)^{\frac{1}{p}}.$$

Note that by Jensen's inequality $\mathcal{W}_p(\mu,\nu) \le \mathcal{W}_q(\mu,\nu)$ whenever $1 \le p \le q < \infty$, in particular $\mathcal{W}_1$ is the weakest of all Wasserstein distances.

Observe that $\mathcal{W}_p(\delta_x,\delta_y) = d(x,y)$, hence, $\mathcal{W}_p$ can be seen as a natural extension of d from X to $\mathcal{P}_p(\mathsf{X})$.

Let us show that $\mathcal{W}_p$ is in fact a distance.

Theorem 1.48. *$\mathcal{W}_p$ defines a distance on $\mathcal{P}_p(\mathsf{X})$.*

Proof. Since the $d(x,y) = d(y,x)$ it follows that $\mathcal{W}_p(\mu,\nu) = \mathcal{W}_p(\nu,\mu)$ and $\mathcal{W}_p(\mu,\mu) = 0$. If $\mathcal{W}_p(\mu,\nu) = 0$ there is a coupling π (note that d^p is continuous and bounded from below so that we have existence of optimal coupligns) which is concentrated on the diagonal $\{(x,x) : x \in \mathsf{X}\}$ since $d(x,y) = 0$ iff $x = y$. Hence, $\mu = \nu$.

It remains to show the triangle inequality. To this end, pick $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(\mathsf{X})$. and let q_1 be optimal between μ_1 and μ_2 and q_2 be optimal between μ_2 and μ_3. Then there exists Markov-process X_1, X_2, X_3 on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_i \sim \mu_i, i = 1,2,3$ and $(X_1, X_2) \sim q_1, (X_2, X_3) \sim q_2$. Using the triangle inequality on $L^p(\mathbb{P})$ it follows that

$$\begin{aligned}\mathcal{W}_p(\mu_1,\mu_3) \le (\mathbb{E}|X_1 - X_3|^p)^{\frac{1}{p}} &\le (\mathbb{E}|X_1 - X_2|^p)^{\frac{1}{p}} + (\mathbb{E}|X_2 - X_3|^p)^{\frac{1}{p}} \\ &= \mathcal{W}_p(\mu_1,\mu_2) + \mathcal{W}_p(\mu_2,\mu_3)\end{aligned}$$

Finally, we need to show that $\mathcal{W}_p$ is real valued. From the triangle inequality we obtain

$$\mathcal{W}_p(\mu,\nu) \le \mathcal{W}_p(\mu,\delta_{x_0}) + \mathcal{W}_p(\nu,\delta_{x_0}) = \left(\int d^p(x,x_0)(d\mu(x) + d\nu(x))\right)^{\frac{1}{p}} < \infty$$

by definition of $\mathcal{P}_p(\mathsf{X})$. □

Theorem 1.49. *If the metric on X is bounded, then $\mathcal{W}_p$ metrizes the weak topology.*

In the proof we use the following probabilistic characterization of the weak topology:

Theorem 1.50 (Skorohod representation)**.** *Let X be a Polish space and $\mu_n, \mu \in \mathcal{P}(\mathsf{X})$ with $\mu_n \to \mu$ weakly. Then there exists a probability space supporting random variables X_n, X with $X_n \sim \mu_n$ and $X \sim \mu$ such that $X_n \to X$ a.s.*

Proof of Theorem 1.49. Let $\mu_n \to \mu$ weakly. Then there exists by the Skorohod representation theorem a probability space supporting X_n, X with $X_n \sim \mu_n$ and $X \sim \mu$ such that $X_n \to X$ a.s. Write $\pi^n := \mathrm{law}(X^n, X) \in \mathcal{P}(\mu_n, \mu)$. We have

$$\limsup_n \mathcal{W}_p^p(\mu_n,\mu) \le \limsup_n \int d^p(x_n,y)\, d\pi^n(x_n,x) = \limsup_n \mathbb{E}[d^p(X^n,X)] = 0,$$

by dominated convergence.

For the converse direction assume that $\mathcal{W}_1(\mu_n,\mu) \to 0$. Let $f \in C_b(\mathsf{X})$ and $\varepsilon > 0$. On a bounded Polish space, Lipschitz functions are uniformly dense in the continuous bounded functions, so pick a Lipschitz function g which is uniformly $\varepsilon/2$-close to f. Then we have by (the trivial part of) Kantorovich-Rubinstein

$$\mu_n(f) - \mu(f) \le \mu_n(g) - \mu(g) + \varepsilon \le \varepsilon + \mathrm{Lip}(g)\mathcal{W}_1(\mu_n,\mu)$$

and, by symmetry also $\mu(f) - \mu_n(f) \le \varepsilon + \mathrm{Lip}(g)\mathcal{W}_1(\mu_n,\mu)$. □

Definition 1.51 (Weak convergence in $\mathcal{P}_p(\mathsf{X})$)**.** *Let $(\mu_n)_n$ be a sequence in $\mathcal{P}_p(\mathsf{X})$ and $\mu \in \mathcal{P}_p(\mathsf{X})$. We say that $(\mu_n)_n$ converges weakly in $\mathcal{P}_p(\mathsf{X})$ to μ if for some (and therefore any) $x_0 \in \mathsf{X}$*

$$\mu_n \to \mu \ \text{ and } \ \int d^p_{\mathsf{X}}(x,x_0)\, d\mu_n(x) \to \int d^p_{\mathsf{X}}(x,x_0)\, d\mu(x).$$

We stress that in the above definition the condition "$\int d^p(x,x_0)\,\mu_n(dx) \to \int d^p(x,x_0)\,\mu(dx)$ for some $x_0 \in \mathsf{X}$" can be replaced by "$\int f(x)\,\mu_n(dx) \to \int f(x)\,\mu(dx)$ for all f continuous of at most p-th order growth".

Theorem 1.52. *The p-Wasserstein distance metrizes weak convergence in $\mathcal{P}_p(\mathsf{X})$.*

Proof. Let $\mu_n, \mu \in \mathcal{P}_p(\mathsf{X})$, $n \in \mathbb{N}$ and $x_0 \in \mathsf{X}$.

If $\mathcal{W}_p(\mu_n, \mu) \to 0$ then we know by Theorem 1.49 that $\mu_n \to \mu$ weakly. Moreover, by the triangle inequality

$$\left| \int d_{\mathsf{X}}^p(x, x_0) \, (\mu_n(dx) - \mu(dx)) \right| \leq \left| \mathcal{W}_p^p(\mu_n, \delta_{x_0}) - \mathcal{W}_p^p(\mu, \delta_{x_0}) \right| \leq 2^{p-1} \mathcal{W}_p(\mu_n, \mu) \to 0,$$

which proves that $\mu_n \to \mu$ in $\mathcal{P}_p(\mathsf{X})$.

On the other hand, if $\mu_n \to \mu$ in $\mathcal{P}_p(\mathsf{X})$ then there exists by the Skorohod representation theorem a probability space with $X_n \sim \mu_n$, $X \sim \mu$ and $X_n \to X$ a.s. As $d_{\mathsf{X}}^p(x, y) \leq 2^{p-1}(d_{\mathsf{X}}^p(x, x_0) + d_{\mathsf{X}}^p(y, x_0))$, we have by Fatou's lemma that

$$\limsup_n \mathbb{E} d_{\mathsf{X}}^p(X^n, X) - 2^{p-1}(d_{\mathsf{X}}^p(X^n, x_0) + d_{\mathsf{X}}^p(X, x_0)) \leq -2^p \mathbb{E} d_{\mathsf{X}}^p(X, x_0),$$

and therefore $\lim_n \mathbb{E} d^p(X^n, X) = 0$. As in Theorem 1.49 we conclude that $\mathcal{W}_p(\mu_n, \mu) \to 0$. □

Lemma 1.53. *A family $\mathscr{E}$ is relatively compact in $\mathcal{P}_p(\mathsf{X})$ if and only if $\mathscr{E}$ is tight and*

$$\lim_{R \to \infty} \sup_{\mu \in \mathscr{E}} \int \mathbb{1}_{\{x \in \mathsf{X} : d_{\mathsf{X}}(x, x_0) \geq R\}} d_{\mathsf{X}}^p(x, x_0) \, \mu(dx) = 0. \tag{1.13}$$

Proof. In the proof of Theorem 1.52 we have seen that $\mu_n \to \mu$ in $\mathcal{P}_p(\mathsf{X})$ if and only if there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $X_n \sim \mu_n$, $X \sim \mu$ such that $d_{\mathsf{X}}^p(X, X_n) \to 0$ in $L^1(\mathbb{P})$. By the Vitali convergence theorem we have that $(d_{\mathsf{X}}^p(X, X_n))_n$ is uniformly integrable, from where it easily follows that also $(d_{\mathsf{X}}(X_n, x_0))_n$ is uniformly integrable. Hence,

$$\lim_{R \to \infty} \sup_n \int \mathbb{1}_{\{x \in \mathsf{X} : d_{\mathsf{X}}(x, x_0) \geq R\}} d_{\mathsf{X}}^p(x, x_0) \, \mu_n(dx) = 0.$$

If $\mathscr{E}$ is relatively compact, then any sequence $(\mu_n)_n$ in $\mathscr{E}$ admits a convergent subsequence with limit in $\mathcal{P}_p(\mathsf{X})$. Then $\mathscr{E}$ is tight and by the first observation we also get (1.13).

If $\mathscr{E}$ is tight and satisfies (1.13), then any sequence admits a convergent subsequence with limit in $\mathcal{P}(\mathsf{X})$. Since $\sup_{\mu \in \mathscr{E}} \int d_{\mathsf{X}}^p(x, x_0) \, \mu(dx) < \infty$ and the map $\mu \mapsto \int d_{\mathsf{X}}^p(x, x_0) \, \mu(dx)$ is l.s.c. on $\mathcal{P}(\mathsf{X})$, the limit has finite p-th moment. When $(\mu_n)_n$ is a sequence in $\mathscr{E}$ such that $\mu_n \to \mu$ in $\mathcal{P}(\mathsf{X})$, then there exists by the Skorohod representation theorem a probability space with $X_n \sim \mu_n$, $X \sim \mu$ and $X_n \to X$ a.s. By (1.13), $(d_{\mathsf{X}}^p(X_n, x_0))_n$ is uniformly integrable and $d_{\mathsf{X}}^p(X_n, x_0) \to d_{\mathsf{X}}^p(X, x_0)$ in L^1, which implies that $\mu_n \to \mu$ in $\mathcal{P}_p(\mathsf{X})$. As $(\mu_n)_n$ was arbitrary, $\mathscr{E}$ is relatively compact in $\mathcal{P}_p(\mathsf{X})$. □

Finally we show:

Theorem 1.54. *$(\mathcal{P}_p(\mathsf{X}), \mathcal{W}(\mathsf{X}))$ is Polish.*

Proof. The weak topology is separable, so we only have to show that $(\mathcal{P}_p(\mathsf{X}), \mathcal{W}_p(\mathsf{X}))$ is complete. That is, given a Cauchy-sequence $(\mu_n)_n$, we have to find an accumulation point. For simplicity we assume that d is bounded and $p = 1$, but both assumptions are not necessary. Passing to a subsequence if necessary, we may assume that $\mathcal{W}(\mu_n, \mu_{n+1}) < 2^{n+1}$. Next pick $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables $X_1, X_2, \ldots$ such that $X_n \sim \mu_n, \mathbb{E} d(X_n, X_n + 1) = \mathcal{W}(\mu_n, \mu_{n+1})$. Then $(X_n)_n$ is an $L_1(\mathbb{P})$-Cauchy sequence. As X is complete, it converges along a subsequence to a random variable X and $\mu := \mathrm{law} X$ is the desired accumulation point of $(\mu_n)_n$. □

1.6. **Displacement interpolation.** Let $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$ endowed with the Euclidean norm and let (X_0, X_1) be distributed according to an $\mathcal{W}_p$-optimal coupling $\pi \in \mathrm{Cpl}(\mu_0, \mu_1)$. For two points $x_0, x_1 \in \mathbb{R}^d$ a curve connecting x_0 and x_1 is given by $x_t := (1 - t)x_0 + t x_1$ which is the unique constant speed geodesic, i.e., $|x_t - x_s| = |t - s||x_0 - x_1|$. The coupling π induces via the *displacement interpolation* a curve of measures $(\mu_t)_{t \in [0,1]}$

$$\mu_t := \mathrm{law}(X_t) \quad \text{where } X_t := (1 - t)X_0 + tX_1.$$

Define $\pi^{s,t} := \mathrm{law}(X_s, X_t) \in \mathrm{Cpl}(\mu_s, \mu_t)$ which yields

$$\mathcal{W}_p(\mu_s, \mu_t) \le \mathbb{E}\left[|X_s - X_t|^p\right]^{\frac{1}{p}} = \mathbb{E}\left[|t-s|^p |X_0 - X_1|^p\right]^{\frac{1}{p}} = |t-s|\mathcal{W}_p(\mu_0, \mu_1).$$

Combining this (when $0 \le s \le t \le 1$) with the triangle inequality we get

$$\begin{aligned}\mathcal{W}_p(\mu_0, \mu_1) &\le \mathcal{W}_p(\mu_0, \mu_s) + \mathcal{W}_p(\mu_s, \mu_t) + \mathcal{W}_p(\mu_t, \mu_1) \\ &\le (s + (t-s) + (1-t))\mathcal{W}_p(\mu_0, \mu_1) = \mathcal{W}_p(\mu_0, \mu_1),\end{aligned}$$

which leads to

$$|t-s|\mathcal{W}_p(\mu_0, \mu_1) = \mathcal{W}_p(\mu_s, \mu_t) \quad \forall s, t \in [0,1]. \tag{1.14}$$

Curves $(\mu_t)_{t\in[0,1]}$ in $\mathcal{P}_p(\mathsf{X})$ that satisfy (1.14) are called *constant speed geodesics*. We conclude that optimal couplings induce via displacement interpolation constant speed geodesics on $\mathcal{P}_p(\mathbb{R}^d)$. The converse is also true (c.f. [Lis09, Theorem 6]): if $(\mu_t)_{t\in[0,1]}$ is a constant speed geodesic in $\mathcal{P}_p(\mathbb{R}^d)$ then it is the displacement interpolation w.r.t. an optimal coupling in $\mathrm{Cpl}(\mu_0, \mu_1)$.

Example 1.55 (convex interpolation vs displacement interpolation). When $p > 1$ the (unique) p-Wasserstein constant speed geodesic between two dirac measures δ_x and δ_y is given by $(\delta_{x_t})_t$ where $x_t := (1-t)x + ty$. Contrary to that, the convex interpolation $\nu_t := (1-t)\delta_x + t\delta_y$ is not of constant speed, since

$$\mathcal{W}_p(\nu_t, \nu_s) = |t-s|^{\frac{1}{p}}|x-y|,$$

which is strictly greater than $|t-s||x-y| = |t-s|\mathcal{W}_p(\delta_x, \delta_y)$ as $p > 1$.

Let $\nabla\varphi\colon \mathbb{R}^d \to \mathbb{R}^d$ be the gradient of a convex function with $\nu = \nabla\varphi_{\#}\mu$ and $X_0 \sim \mu$. Define $X_t := (1-t)X_0 + t\nabla\varphi(X_0)$ and observe that a.s. $X = (X_t)_{t\in[0,1]} \in \mathrm{AC}$ (where AC denotes here the set of absolutely continuous curves on $\mathbb{R}^d$) with derivative $\dot{X}_t = \nabla\varphi(X_0) - X_0$. We know by Brenier's theorem that the law of $(X_0, \nabla\varphi(X_0))$ is an $\mathcal{W}_2$-optimal coupling, hence,

$$\mathcal{W}_2^2(\mu, \nu) = \mathbb{E}\left[|X_0 - X_1|^2\right] = \int_0^1 \mathbb{E}\left[|\dot{X}_t|^2\right] dt.$$

On the other hand, when $X = (X_t)_{t\in[0,1]}$ is a.s. absolutely continuous such that $X_0 \sim \mu$ and $X_1 \sim \nu$, we have

$$\mathcal{W}_2^2(\mu, \nu) \le \mathbb{E}\left[|X_0 - X_1|^2\right] = \mathbb{E}\left[\left|\int_0^1 \dot{X}_t\, dt\right|^2\right] \le \int_0^1 \mathbb{E}\left[|\dot{X}_t|^2\right].$$

We have derived the following dynamic optimal transport formulation: Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, it holds that

$$\mathcal{W}_2^2(\mu, \nu) = \inf\left\{\int_0^1 \mathbb{E}[|\dot{X}_t|^2]\, dt : X \in \mathrm{AC}, X_0 \sim \mu, X_1 \sim \nu\right\}.$$

Benamou and Brenier [BB00] introduced a dynamic formulation of the Wasserstein-2 distance where they connect the Wasserstein-2 distance to curves of measures $(\mu_t)_t$ satisfying weak formulation of the continuity equation $\partial_t\mu_t + \nabla\cdot(\mu_t v_t) = 0$ for some velocity field $v(t,x)$. Otto [Ott01] observed that the Benamou-Brenier formula can formally be interpreted as the length distance induced by a Riemannian metric on $\mathcal{P}_2(\mathbb{R}^d)$. This observation allowed to interpret solutions of certain PDEs as gradient flows w.r.t. the Wasserstein geometry. Even though this is a very exciting direction of research, proceeding here would go beyond the scope of these lecture notes For the interested reader we refer to [FG21] for an introduction to this topic and [AGS08] for a thorough study of gradient flows in $\mathcal{P}_p(\mathsf{X})$.

1.7. **Multi-marginal optimal transport.** Given $\mu_1, \dots, \mu_N \in \mathcal{P}_p(\mathsf{X})$ and a cost function $c\colon \mathsf{X}^N \to \mathbb{R} \cup \{\infty\}$, the multi-marginal optimal transport problem is given by

$$V_c(\mu_1, \dots, \mu_N) := \inf_{\pi \in \mathrm{Cpl}(\mu_1, \dots, \mu_N)} \int c(x_1, \dots, x_N)\, d\pi(x_1, \dots, x_N), \tag{1.15}$$

where $\mathrm{Cpl}(\mu_1, \dots, \mu_N) := \{\pi \in \mathcal{P}(\mathsf{X}^N) : \mathrm{proj}^n_\# \pi = \mu_n,\ n = 1, \dots, N\}$. With minor modifications of the proofs in the previous section, we find when c is lower semicontinuous and bounded from below that

(1) the optimal value $V_c(\mu_1, \dots, \mu_N)$ is attained,
(2) the optimal value map $(\mu_1, \dots, \mu_N) \mapsto V_c(\mu_1, \dots, \mu_N)$ is lower semicontinuous,
(3) there holds duality

$$V_c(\mu_1, \dots, \mu_N) = \sup_{\substack{(\varphi_1, \dots, \varphi_N) \in C_b(\mathsf{X})^N \\ \sum_{n=1}^N \varphi(x_n) \le c(x_1, \dots, x_N)}} \sum_{n=1}^N \mu_n(\varphi_n).$$

A very natural multi-marginal transport problem is given by the so called Wasserstein-barycenter introduced in [AC11].

The Wasserstein distance induces a metric on the set of probability measures that lifts the metric structure of the base space X onto the set of probability measures on X. Given data from multiple sources in form of distributions $\mu_1, \dots, \mu_N \in \mathcal{P}_p(\mathsf{X})$, the Wasserstein barycenter is a way of summarizing them which contrary to a convex combination also takes into account the geometry of the underlying space. A p-Wasserstein barycenter of $(\mu_1, \dots, \mu_N)$ is a minimizer of

$$\inf_{\mu \in \mathcal{P}_p(\mathsf{X})} \frac{1}{N} \sum_{n=1}^N \mathcal{W}_p^p(\mu_n, \mu). \tag{1.16}$$

Theorem 1.56. *Let $\mu_1, \dots, \mu_N \in \mathcal{P}_2(\mathbb{R}^d)$ and $c(x_1, \dots, x_N) := \frac{1}{N} \sum_{n=1}^N |x_n - T(x_1, \dots, x_N)|^2$ where $T(x_1, \dots, x_N) := \bar{x}_N := \frac{1}{N} \sum_{n=1}^N x_n$. Then* (1.16) *admits a minimizer and we have*

$$V_c(\mu_1, \dots, \mu_N) = \inf_{\mu \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{N} \sum_{n=1}^N \mathcal{W}_p^2(\mu_n, \mu). \tag{1.17}$$

Moreover, if $\pi^* \in Cpl(\mu_1, \dots, \mu_N)$ is optimal then $T_\# \pi^* \in \mathcal{P}_p(\mathbb{R}^d)$ is a p-Wasserstein barycenter of $(\mu_1, \dots, \mu_N)$.

Proof. Plainly we have

$$\inf_{\mu \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{N} \sum_{n=1}^N \mathcal{W}_p^2(\mu_n, \mu) = \inf_{X, X_1 \sim \mu_1, \dots, X_N \sim \mu_N} \sum_{i=1}^N \mathbb{E}[|X_i - X|^p] \tag{1.18}$$

$$= \inf_{X_1 \sim \mu_1, \dots, X_N \sim \mu_N} \sum_{i=1}^N \mathbb{E}[|X_i - \bar{X}_N|^p] \tag{1.19}$$

$$= \inf_{\pi \in \mathrm{Cpl}(\mu_1, \dots, \mu_N)} \int c(x_1, \dots, x_N)\, d\pi(x_1, \dots, x_N). \tag{1.20}$$

□

2. Weak optimal transport

Let π be a probability measure on $\mathsf{X}\times\mathsf{Y}$ with first marginal μ. The disintegration theorem tells us that for any measurable functions $c\colon \mathsf{X}\times\mathsf{Y}\to\mathbb{R}\cup\{\infty\}$ that is bounded from below, we have the identity

$$\pi(c) = \int\int c(x,y)\,\pi_x(dy)\,\mu(dx).$$

In this case, we can define the measurable map $C(x,\rho) := \int c(x,y)\,\rho(dy)$ with domain $\mathsf{X}\times\mathcal{P}(\mathsf{Y})$, which allows us to reformulate the optimal transport problem as

$$V_c(\mu,\nu) = \inf_{\pi\in\mathrm{Cpl}(\mu,\nu)}\int\int c(x,y)\,\pi_x(dy)\,\mu(dx) = \inf_{\pi\in\mathrm{Cpl}(\mu,\nu)}\int C(x,\pi_x)\,\mu(dx).$$

2.1. Problem formulation.

Weak optimal transport is a generalization of optimal transport that is concerned with cost functions $C\colon \mathsf{X}\times\mathcal{P}_p(\mathsf{Y})\to\mathbb{R}\cup\{\infty\}$. Cost functions of this type are called *weak transport costs*. For $\mu\in\mathcal{P}_p(\mathsf{X})$ and $\nu\in\mathcal{P}_p(\mathsf{Y})$, the weak optimal transport problem is defined as

$$V_C^{\mathrm{WT}}(\mu,\nu) := \inf_{\pi\in\mathrm{Cpl}(\mu,\nu)}\int_{\mathsf{X}} C(x,\pi_x)\,\mu(dx). \tag{2.1}$$

Example 2.1 (Entropic optimal transport). Let $\mu\in\mathcal{P}(\mathsf{X})$, $\nu\in\mathcal{P}(\mathsf{Y})$, and $c\in M_b(\mathsf{X}\times\mathsf{Y})$ a cost function. The entropic transport problem is given by

$$V_c^{\mathrm{EOT}}(\mu,\nu) := \inf_{\pi\in\mathrm{Cpl}(\mu,\nu)}\pi(c) + H(\pi|\mu\otimes\nu),$$

where H denotes the relative entropy

$$H(\eta|\rho) = \begin{cases}\int\log\left(\frac{d\eta}{d\rho}(x)\right)\,d\eta(x) & \eta\ll\rho,\\ \infty & \text{else,}\end{cases}$$

for $\eta,\rho\in\mathcal{P}(\mathsf{X})$. This problem can be reformulated as a weak transport problem: When $\pi\ll\mu\otimes\nu$ the density satisfies $\frac{d\pi}{d\mu\otimes\nu}(x,y) = \frac{d\pi_x}{d\nu}(y)$ which we use to compute

$$H(\pi|\mu\otimes\nu) = \int\log\left(\frac{d\pi}{d\mu\otimes\nu}\right)d\pi = \int\int\log\left(\frac{d\pi_x}{d\nu}\right)d\pi_x\,d\mu = \int H(\pi_x|\nu)\,\mu(dx),$$

hence,

$$V_c^{\mathrm{EOT}}(\mu,\nu) = \inf_{\pi\in\mathrm{Cpl}(\mu,\nu)}\int\left(\int c(x,y)\,\pi_x(dy) + H(\pi_x|\nu)\right)\mu(dx).$$

We will discuss that $(\mu,\nu)\mapsto H(\mu|\nu)$ is bounded from below, jointly lower semicontinuous and jointly convex in a later chapter.

Definition 2.2. *Let $\mu,\nu\in\mathcal{P}_1(\mathbb{R}^d)$, and define* mean$\colon\mathcal{P}_1(\mathbb{R}^d)\to\mathbb{R}:\rho\mapsto\int y\,\rho(dy)$.

A coupling $\pi\in Cpl(\mu,\nu)$ is called a martingale coupling *if*

$$\mathrm{mean}(\pi_x) = x \quad \textit{for } \mu\textit{-a.e. } x.$$

The subset of all martingale couplings is denoted by $Cpl^M(\mu,\nu)$.

Example 2.3 (Martingale optimal transport). Let $c\colon\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$ and $\mu,\nu\in\mathcal{P}_1(\mathbb{R}^d)$. The martingale optimal transport problem consists of minimizing

$$V_c^{\mathrm{MOT}}(\mu,\nu) := \inf_{\pi\in\mathrm{Cpl}^{\mathrm{M}}(\mu,\nu)}\pi(c). \tag{2.2}$$

The martingale constraint on the disintegration kernel $(\pi_x)_x$ can be incorporated into a weak transport cost. Indeed, we can define

$$C(x,\rho) := \begin{cases}\int c(x,y)\,\rho(dy) & \mathrm{mean}(\rho) = x,\\ \infty & \text{else.}\end{cases}$$

Since the mapping mean$\colon\mathcal{P}_1(\mathbb{R}^d)\to\mathbb{R}$ is continuous, we have that the set $\{(x,\rho) : \mathrm{mean}(\rho) = x\}\subseteq\mathbb{R}^d\times\mathcal{P}_1(\mathbb{R}^d)$ is closed. We observe that, when c is bounded from below and lower semicontinuous, then C is bounded from below, lower semicontinuous and

convex in its second argument. Since $\pi(c) = \int C(x, \pi_x)\,\mu(dx) < \infty$ implies that π is a martingale coupling, we conclude

$$V_c^{\mathrm{MOT}}(\mu,\nu) = \inf_{\pi\in \mathrm{Cpl}(\mu,\nu)} \int C(x,\pi_x)\,\mu(dx).$$

2.2. **Basic properties.** The fact that disintegration of measures is a non-continuous operation, complicates the study of the weak transport problem. For example, given a weak transport cost $C \in C_b(\mathsf{X} \times \mathcal{P}(Y))$ the map

$$\pi \mapsto \int C(x,\pi_x)\,\mu(dx).$$

is in general just measurable.

We note that in Examples 2.1 and 2.3 the function $C\colon \mathsf{X} \times \mathcal{P}_p(\mathsf{Y}) \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous, bounded from below and convex in its second argument.

Indeed, this is satisfied for most natural appearances of weak transport problems and we will mostly work under this assumption.

What makes the weak transport problem slightly challenging, is that lower semi-continuity of $C : \mathcal{P}(\mathsf{X} \times \mathcal{P}(\mathsf{Y})) \to (-\infty, \infty]$ is not as directly applicable as in the case of the classical transport. We first need to pass from the transport plan π to its disintegration with respect to the first marginal μ. This operation is in general not continuous with respect to the weak topology and requires some attention.

To deal with the disintegration, we introduce the injection $J\colon \mathcal{P}(\mathsf{X}\times\mathsf{Y}) \to \mathcal{P}(\mathsf{X}\times\mathcal{P}(\mathsf{Y}))$ which is defined by

$$J(\pi) := ((x,y) \mapsto (x,\pi_x))_{\#}\,\pi.$$

(1) The injection J is well-defined, since the disintegration kernel is unique up to null sets.

(2) Measurability of J is trickier. A possible way is to first show that

$$\mathscr{F} := \{P \in \mathcal{P}(\mathsf{X}\times\mathcal{P}(\mathsf{Y})) :\ P \text{ is supported by the graph of a function }\}$$

is a G_δ-subset of $\mathcal{P}(\mathsf{X} \times \mathcal{P}(\mathsf{Y}))$. Thus, $\mathscr{F}$ equipped with the trace topology is Polish. Restricting the image yields that $J\colon \mathcal{P}(\mathsf{X}\times\mathsf{Y}) \to \mathscr{F}$ is a bijection and its right-inverse is continuous. (Explicitly, the right inverse is given by the mapping $\hat{I}$ defined in (2.5) below.)

Therefore, by the Lusin-Souslin theorem [Kec95, Theorem 15.1] (which states that the injective image of a Borel set under a continuous function between two Polish spaces is again Borel) we have that $J^{-1}(B) = J^{-1}(B \cap \mathscr{F}) \in \mathcal{B}(\mathcal{P}(\mathsf{X}\times\mathsf{Y})$. We conclude that J is measurable.

2.3. **The intensity operator.** To better exploit the (lower semi-) continuity of the cost function C, we will introduce a counterpart to the set of transport plans in $\mathbb{P}(\mathsf{X}\times\mathbb{P}(\mathsf{Y}))$. The formal definition of this set $\Lambda(\mu,\nu)$ will be given subsequently in Definition 2.9.

Definition 2.4 (Intensity operator)**.** *The intensity operator $I\colon \mathcal{P}(\mathcal{P}(\mathsf{X})) \to \mathcal{P}(\mathsf{X})$ maps a measure $P \in \mathcal{P}(\mathcal{P}(\mathsf{X}))$ to the element $I(P) \in \mathcal{P}(\mathsf{X})$ satisfying, for any $f \in C_b(\mathsf{X})$,*

$$\int_{\mathsf{X}} f(x)\,I(P) = \int_{\mathcal{P}(\mathsf{X})} p(f)\,P(dp). \tag{2.3}$$

Remark 2.5. Some immediate properties of the intensity operator:

(1) The intensity operator is well-defined as for all $B \in \mathcal{B}(\mathsf{X})$ the map $\mathcal{P}(\mathsf{X}) \ni p \mapsto p(B)$ is measurable and therefore it is straightforward to verify that the right-hand side in (2.3) defines a probability measure on X. Instead of (2.3), we could define the intensity by asserting that for all measurable $B \subseteq \mathsf{X}$

$$I(P)(B) = \int_{\mathcal{P}(\mathsf{X})} p(B)\,P(dp). \tag{2.4}$$

(2) It is also possible (and sometimes useful) to related the intensity operator to a form of expectation. Specifically, let Z be a $\mathbb{P}(\mathsf{X})$-valued random variable. Its expectation is the probability $\mathsf{E}(Z) \in \mathbb{P}(\mathsf{X})$ satisfying

$$\mathsf{E}[Z](B) = \mathbb{E}[Z(B)]$$

for all measurable $B \subseteq \mathsf{X}$. We thus have

$$\mathsf{E}[Z] = I(\mathrm{law}(Z)).$$

The expectation of a measure-valued random variables Z relates to the intensity of $\mathrm{law}(Z)$ in precisely the same way as the usual expectation of a random variable X relates to the barycenter of the law of X.

Of course one can also define the conditional expectation of a measure valued random variable (which corresponds to the intensity of the conditional law of this random variable). E inherits the basic properties of the usual expectation operator such as the tower law.

(3) From the definition of the weak topology, we have that, for $f \in C_b(\mathsf{X})$, the map $F_f \colon \mathcal{P}(\mathsf{X}) \to \mathbb{R} \colon p \mapsto p(f)$ is continuous, which has continuity of I w.r.t. the weak topology as consequence.

(4) Restricted to $\mathcal{P}_p(\mathcal{P}_p(\mathsf{X}))$, the intensity operator maps continuously to $\mathcal{P}_p(\mathsf{X})$, indeed it is 1-Lipschitz for the respective Wasserstein metrics.

Proof. Indeed, fix $P, Q \in \mathcal{P}_p(\mathcal{P}_p(\mathsf{X}))$ and observe that, for $f, g \in C_b(\mathsf{X})$ with $f(x) + g(y) \le d^p_{\mathsf{X}}(x, y)$, we have $\eta(f) + \rho(g) \le \inf_{\pi \in \mathrm{Cpl}(\eta,\rho)} \int d^p_{\mathsf{X}}(x, y)\, d\pi = \mathcal{W}^p_p(\eta, \rho)$ for all $\eta, \rho \in \mathcal{P}_p(\mathsf{X})$. Therefore, we have by optimal transport duality

$$\begin{aligned}\mathcal{W}^p_p(I(P), I(Q)) &= \sup_{\substack{f(x)+g(y)\le d^p_{\mathsf{X}}(x,y),\\ f,g\in C_b(\mathsf{X})}} P(F_f) + Q(F_g)\\ &\le \sup_{\substack{F(\eta)+G(\rho)\le \mathcal{W}^p_p(\eta,\rho)\\ F,G\in C_b(\mathcal{P}_p(\mathsf{X}))}} P(F) + Q(G) = \mathcal{W}^p_p(P, Q).\end{aligned}$$ □

(5) We have the following Jensen-type inequality: let $f \colon \mathcal{P}_p(\mathcal{P}_p(\mathsf{X})) \to (-\infty, \infty]$ be lower semicontinuous, convex and bounded from below, and $P \in \mathcal{P}_p(\mathcal{P}_p(\mathsf{X}))$. Then

$$f(I(P)) \le \int f(\rho)\, dP(\rho).$$

Indeed if Z is a random variable with law P, then this amounts to

$$f(\mathsf{E}(Z)) \le \mathbb{E}(f(Z)).$$

Proof. With the law of large numbers, for example, one can show that there is a sequence of discrete measures with $P^n \to P$ and $P^n(f) \to P(f)$. We obtain Jensen's inequality

$$f(I(P)) \le \liminf_n f(I(P^n)) \le \liminf_n P^n(f) = P(f) = \int f(\rho)\, dP(\rho),$$

where we used continuity of I and lower semicontinuity for the first and convexity for the second inequality. □

Lemma 2.6. *A set $\mathscr{E} \subseteq \mathcal{P}_p(\mathcal{P}_p(\mathsf{X}))$ is relatively compact if and only if $I(\mathscr{E}) \subseteq \mathcal{P}_p(\mathsf{X})$ is relatively compact.*

Proof idea. To sketch the main idea we prove the assertion for d_{X} bounded, which means that all spaces are equipped with the weak topology, c.f. Theorem 1.49. The general case is not much different but slightly more technical as it requires, besides showing tightness, to adequately treat moments. A complete proof can be found in [BVBP19].

Assume that $\mathscr{E}$ is relatively compact. Since the image of a relatively compact set under a continuous map is again relatively compact, the first implication follows by continuity of I.

Next, assume that $I[\mathscr{E}]$ is relatively compact in $\mathcal{P}_p(\mathsf{X})$. Fix $\varepsilon > 0$ and choose for every $n \in \mathbb{N}$ a compact $K_n \subseteq \mathsf{X}$ such that

$$\sup_{\mu \in I[\mathscr{E}]} \mu(K_n^{\mathrm{c}}) < \frac{\varepsilon^2}{2^{2n}}.$$

The following set

$$K := \left(\left\{ \mu \in \mathcal{P}_p(\mathsf{X}) : \forall n \in \mathbb{N},\, \mu(K_n^{\mathrm{c}}) \leq \frac{\varepsilon}{2^n} \right\} \right)$$

is by Portmanteau's theorem closed and by construction tight, thus, compact by Prohorov's theorem. Let $P \in \mathscr{E}$. Using Markov's inequality, the next computation shows tightness

$$P(K^{\mathrm{c}}) \leq \sum_n P(\{\mu : \mu(K_n^{\mathrm{c}}) > \frac{\varepsilon}{2^n}) \leq \sum_n \frac{2^n}{\varepsilon} \int \rho(K_n^{\mathrm{c}})\, P(d\rho) < \sum_n \frac{\varepsilon}{2^n} = \varepsilon. \qquad \square$$

It is convenient to introduce another intensity operator $\hat{I} \colon \mathcal{P}_p(\mathsf{X} \times \mathcal{P}_p(\mathsf{Y})) \to \mathcal{P}_p(\mathsf{X} \times \mathsf{Y})$ where $\hat{I}(P)$ is given by

$$\hat{I}(P)(f) = \int \int f(x,y)\, p(dy)\, P(dx, dp) \quad \forall f \in M_b(\mathsf{X} \times \mathsf{Y}). \tag{2.5}$$

Similarly to the intensity operator I, we have that $\hat{I}$ is well-defined and continuous.

Corollary 2.7. *A subset $\mathscr{E} \subseteq \mathcal{P}_p(\mathsf{X} \times \mathcal{P}_p(\mathsf{Y}))$ is relatively compact if and only if $\hat{I}[\mathscr{E}] \subseteq \mathcal{P}_p(\mathsf{X} \times \mathsf{Y})$ is relatively compact.*

Proof. The set $\hat{I}[\mathscr{E}]$ is relatively compact if and only if its set of X- and Y-marginals are relatively compact in $\mathcal{P}_p(\mathsf{X})$ and $\mathcal{P}_p(\mathsf{Y})$ respectively. By Lemma 2.6 this is precisely the case when the set of X- and $\mathcal{P}_p(\mathsf{Y})$-marginals are relatively compact. But, this is equivalent to $\mathscr{E}$ being relatively compact. $\square$

Lemma 2.8. *Let $P = \mu \otimes \kappa \in \mathcal{P}(\mathsf{X} \times \mathcal{P}(\mathsf{Y}))$ for a kernel $\kappa \colon \mathsf{X} \to \mathcal{P}(\mathcal{P}(\mathsf{Y}))$. Then we have the identity $\hat{I}(P) = \mu \otimes I(\kappa)$. In particular, $I(\kappa) \colon \mathsf{X} \to \mathcal{P}(\mathsf{Y})$ is a disintegration of $\hat{I}(P)$.*

Proof. Fix $f \in M_b(\mathsf{X} \times \mathsf{Y})$. By definition of I, $\hat{I}$ and as I is measurable we find

$$\hat{I}(P)(f) = \int \int \int f(x,y)\, p(dy)\, \kappa_x(dp)\, \mu(dx) = \int \int f(x,y)\, I(\kappa_x)(dy)\, \mu(dx).$$

Hence, $I(\kappa)$ is a disintegration. $\square$

Definition 2.9. *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$. Then we set*

$$\Lambda(\mu, \nu) := \hat{I}^{-1}(Cpl(\mu, \nu)) \subseteq \mathcal{P}(\mathsf{X} \times \mathcal{P}(\mathsf{Y})).$$

Note that

$$\Lambda(\mu,\nu) = \{P : \mathrm{proj}_{\mathsf{X}}(P) = \mu, I \circ \mathrm{proj}_{\mathcal{P}(\mathsf{Y})}(P) = \nu\} = \{P : \int \varphi(x)\, dP(x,\rho) = \mu(\varphi), \int \psi(\rho)\, dP(x,\rho) = \nu(\psi), \varphi \in C_b(\mathsf{X}), \psi \in C$$

We then have

Corollary 2.10. *Assume that $\mu \in \mathcal{P}_p(\mathsf{X}), \nu \in \mathcal{P}_p(\nu)$. Then $\Lambda(\mu, \nu)$ is compact with respect to p-weak convergence.*

2.4. **Existence.** From Corollary 2.10 we know that

$$\inf_{P \in \Lambda(\mu,\nu)} \int C(x,p)\, dP(x,p)$$

is attained for lower semicontinuous c. If C is convex in the second argument and we denote $\pi := \hat{I}(P), \bar{P} := J(\pi)$, we have by Jensen's inequality that

$$P(C) = \int\int C(x,p)\, P_x(dp)\, \mu(dx) \geq \int C(x, I(P_x))\, \mu(dx) = \int C(x, \pi_x)\, \mu(dx) = \bar{P}(C). \tag{2.6}$$

Putting these facts together we obtain that for l.s.c. C which is convex in the second argument

$$\inf_{\pi\in \mathrm{Cpl}(\mu,\nu)} \int C(x,\pi_x)\, d\mu(x) = \inf_{P\in\Lambda(\mu,\nu)} \int C(x,p)\, dP(x,P) \tag{2.7}$$

and both sides are attained.

Indeed we will show a bit more below.

Proposition 2.11. *Let $C\colon \mathsf{X}\times\mathcal{P}_p(\mathsf{Y}) \to \mathbb{R}\cup\{\infty\}$ be jointly lower semicontinuous, bounded from below and convex in the $\mathcal{P}_p(\mathsf{Y})$-coordinate, then the mapping*

$$\mathcal{P}_p(\mathsf{X}\times\mathsf{Y}) \ni \pi \mapsto \int C(x,\pi_x)\,\mu(dx)$$

is lower semicontinuous.

Proof. Let $(\pi^n)_n$ with $\pi^n = \mu_n\otimes\pi^n_x$ be a sequence that converges in $\mathcal{P}_p(\mathsf{X}\times\mathsf{Y})$ to $\pi = \mu\otimes\pi_x$. Then we have by Lemma 2.7 that also the set $(P^n := J(\pi^n))_n$ is relatively compact. Therefore, we can find a convergent subsequence $(P^{n_k})_k$ with limit P such that $\liminf_n P^n(C) = \liminf_k P^{n_k}(C)$. The intensity $\hat{I}$ is continuous, which yields $\pi = \lim_n \pi^n = \lim_n \hat{I}(J(\pi^n)) = \hat{I}(P)$ and $P = \mu\otimes P_x$. Due to Jensen's inequality (which holds here true since $p\mapsto C(x,p)$ is lower semicontinuous and convex) we find

$$P(C) = \int\int C(x,p)\, P_x(dp)\,\mu(dx) \geq \int C(x, I(P_x))\,\mu(dx) = \int C(x,\pi_x)\,\mu(dx), \tag{2.8}$$

where we used that by Lemma 2.8 $\pi_x = I(P_x)$ μ-a.s. Hence, we conclude

$$\liminf_n \int C(x,\pi^n_x)\,\mu_n(dx) = \liminf_n P^n(C) \geq P(C) = \int C(x,\pi_x)\,\mu(dx).$$

□

Theorem 2.12. *The optimal value map $V_C^{\mathrm{WOT}}\colon \mathcal{P}_p(\mathsf{X})\times\mathcal{P}_p(\mathsf{Y}) \to \mathbb{R}\cup\{\infty\}$ is jointly lower semicontinuous, convex in its second argument and the infimum is attained.*

Proof. To see convexity, note that when $\pi^1,\pi^2 \in \mathrm{Cpl}(\mu,\nu)$ so is $\pi := \frac{1}{2}(\pi^1+\pi^2)$ and $\pi_x = \frac{1}{2}(\pi^1_x+\pi^2_x)$, thus, $\frac{1}{2}\int C(x,\pi^1_x)+C(x,\pi^2_x)\,\mu(dx) \geq \int C(x,\pi_x)\,\mu(dx)$.

Attainment and lower semicontinuity of the optimal value is a consequence of compactness (see Lemma 1.53) and lower semicontinuity of the value function $\pi\mapsto\int C(x,\pi_x)\,\mu(dx)$ (see Proposition 2.11). □

2.5. **Duality.**

Definition 2.13. *Given the cost C, the set of admissible dual variables is given by*

$$\mathcal{D}(C) := \left\{(\varphi,\psi) : \varphi\in M(\mathsf{X};\bar{\mathbb{R}}), \psi\in M_p(\mathsf{Y}),\ \varphi(x)+\rho(\psi)\leq C(x,\rho)\right\}.$$

Definition 2.14 (C-conjugate)**.** *Let $C\colon \mathsf{X}\times\mathcal{P}_p(\mathsf{Y}) \to \mathbb{R}\cup\{\infty\}$ be a cost function. For a function $\psi\colon \mathsf{Y}\to\mathbb{R}$ we define its C-transform $\psi^C\colon \mathsf{X}\to\mathbb{R}\cup\{\infty\}$ by*

$$\psi^C(x) := \inf_{\rho\in\mathcal{P}_p(\mathsf{Y})} C(x,\rho) - \rho(\psi).$$

Remark 2.15. Similarly to the c-transform from optimal transport, one can argue by abstract arguments from descriptive set theory that ψ^C is analytically measurable if C and ψ are Borel. This suffices for our purposes. In the examples below, either C will be continuous (thus, ψ^C is upper semicontinuous) or one can directly argue measurability.

Theorem 2.16 (Duality)**.** *Let $C\colon \mathsf{X}\times\mathcal{P}_p(\mathsf{Y}) \to \mathbb{R}\cup\{\infty\}$ be lower semicontinuous. Then*

$$\inf_{\pi\in Cpl(\mu,\nu)} \int C(x,\pi_x)\,\mu(dx) = \sup_{\substack{(\varphi,\psi)\in\mathcal{D}(C)\\ \varphi\in C_b(\mathsf{X}),\psi\in C_p(\mathsf{Y})}} \mu(\varphi)+\nu(\psi) = \sup_{\psi\in C_{b,p}(\mathsf{Y})} \mu(\psi^C)+\nu(\psi).$$

Remark 2.17. For general $\psi \in C_p(\mathsf{Y})$, their C-transform might not be μ-integrable. To avoid this, we restrict to ψ that are additionally bounded from above, that is $\psi \in C_{b,p}(\mathsf{Y})$. We have for $(\varphi, \psi) \in \mathcal{D}(C) \cap (C_b(\mathsf{X}) \times C_p(\mathsf{Y}))$ by monotone convergence that

$$\varphi(x) \le \psi^C(x) = \inf_{k\in\mathbb{N}} (\psi \vee k)^C(x),$$

thus, $\mu(\varphi) + \nu(\psi) \le \lim_k \mu((\psi \vee k)^C) + \nu(\psi \vee k)$.

Proof. We prove the theorem under the additional assumption that X, Y are compact. In this case we can use the min-max argument we have seen in the classical case almost verbatim. As before we write

$$\chi(P) = \begin{cases} 0 & \text{if } P \in \Lambda(\mu,\nu) \\ \infty & \text{else.} \end{cases} = \sup_{(\varphi,\psi)\in C_b(\mathsf{X})\times C_b(\mathsf{Y})} \mu(\varphi) + \nu(\psi) - \int \varphi(x) + p(\psi)\, dP(x,p).$$

As $\mathcal{P}(X \times \mathcal{P}(Y))$ is compact for compact spaces X, Y we can apply Theorem 1.20 to interchange inf and sup to obtain

$$\begin{aligned}
&\inf_{P\in\Lambda(\mu,\nu)} \int C\, dP = \inf_{P\in\mathcal{P}(\mathsf{X}\times\mathcal{P}(\mathsf{Y}))} \int c\, d\pi + \chi(\pi) \\
&= \inf_{P\in\mathcal{P}(\mathsf{X}\times\mathcal{P}(\mathsf{Y}))} \sup_{(\varphi,\psi)\in C_b(\mathsf{X})\times C_b(\mathsf{Y})} \int C(x,p) - \varphi(x) - p(\psi)\, dP(x,p) + \mu(\varphi) + \nu(\psi) \\
&= \sup_{(\varphi,\psi)\in C_b(\mathsf{X})\times C_b(\mathsf{Y})} \inf_{P\in\mathcal{P}(\mathsf{X}\times\mathcal{P}(\mathsf{Y}))} \int C(x,p) - \varphi(x) - p(\psi)\, dP(x,p) + \mu(\varphi) + \nu(\psi) \\
&= \sup_{(\varphi,\psi)\in C_b(\mathsf{X})\times C_b(\mathsf{Y})} \inf_{(x,p)\in\mathsf{X}\times\mathcal{P}(\mathsf{Y})} C(x,p) - \varphi(x) - p(\psi) + \mu(\varphi) + \nu(\psi) \\
&= \sup_{(\varphi,\psi)\in C_b(\mathsf{X})\times C_b(\mathsf{Y}),\varphi(x)+p(\psi)\le C(x,p)} \mu(\varphi) + \nu(\psi).
\end{aligned}$$

□

To prove the result in the general case, one can either use approximation arguments or apply the Fenchel-Moreau theorem, see [BVBP20].

2.6. **Strassen's theorem.** Given two marginals $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$ Strassen [Str65] characterized the existence of a martingale coupling with said marginals. By Jensen's inequality a necessary condition is that μ and ν are in the convex order:

Definition 2.18 (convex order)**.** *We say μ and ν are in convex order and write $\mu \preceq_c \nu$ if, for all convex $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, $\mu(f) \le \nu(f)$.*

To show Strassen's theorem, we need the following lemma:

Lemma 2.19. *Let $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be lsc and bounded from below. Then $\varphi^{**}(x) = \inf\{\rho(\varphi) : \rho \in \mathcal{P}_1(\mathbb{R}^d), \mathrm{mean}(\rho) = x\}$.*

Proof. Observe that φ^{**} is the largest lower semicontinuous and convex function below φ. For example, this can be seen from the Fenchel-Moreau theorem and by monotonicity of the biconjugate, i.e., $\varphi_1 \le \varphi_2$ then $\varphi_1^{**} \le \varphi_2^{**}$.

On the one hand, one can verify that $\psi(x) := \inf\{\rho(\varphi) : \rho \in \mathcal{P}_1(\mathbb{R}^d), \mathrm{mean}(\rho) = x\}$ is lower semicontinuous, convex and dominated by φ, thus, $\psi \le \varphi^{**}$. On the other hand, we have by Jensen's inequality that

$$\varphi^{**}(\mathrm{mean}(\rho)) \le \rho(\varphi^{**}) \le \rho(\varphi),$$

thus, $\varphi^{**} \le \psi$ which completes the proof. □

Theorem 2.20 (Strassen)**.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Then $\mu \preceq_c \nu$ if and only if $Cpl^M(\mu,\nu) \neq \emptyset$.*

Proof. Let $C(x,\rho)\colon \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d) \to [0,\infty]$ be given by

$$C(x,\rho) := \begin{cases} 0 & \mathrm{mean}(\rho) = x, \\ \infty & \text{else.} \end{cases}$$

Clearly, $\mathrm{Cpl}^{\mathrm{M}}(\mu,\nu)$ is non empty iff the weak transport problem associated to the cost C is 0. Since C is bounded from below, lower semicontinuous and convex, we can apply the duality theorem and get

$$\inf_{\pi\in\mathrm{Cpl}(\mu,\nu)}\int C(x,\pi_x)\,\mu(dx)=\sup_{\psi\in C_{b,1}(\mathbb{R}^d)}\nu(\psi)+\mu(\psi^C).$$

We find for any $\psi\in C_{b,1}(\mathbb{R}^d)$ that

$$\psi^C(x)=\inf_{\rho\in\mathcal{P}_1(\mathbb{R}^d)}C(x,\rho)-\rho(\psi)=\inf_{\substack{\rho\in\mathcal{P}_1(\mathbb{R}^d)\\ \mathrm{mean}(\rho)=x}}\rho(-\psi)=(-\psi)^{**}(x), \tag{2.9}$$

where we applied Lemma 2.19 for the last equality. As $(-\psi)^{**}\le-\psi$ we have

$$\sup_{\psi\in C_{b,1}(\mathbb{R}^d)}\nu(\psi)+\mu(\psi^C)=\sup_{\psi\in C_{b,1}(\mathbb{R}^d)}-\nu((-\psi)^{**})+\mu((-\psi)^{**})=\sup_{\substack{\psi\in C_{b,1}(\mathbb{R}^d),\\ \text{concave}}}\nu(-\psi)-\mu(-\psi).$$

We conclude that $\mathrm{Cpl}^{\mathrm{M}}(\mu,\nu)\neq\emptyset$ if and only if $\mu\preceq_c\nu$. □

2.7. **Convex Kantorovich-Rubinstein duality.** Now we let $C(x,\rho):=|x-\mathrm{mean}(\rho)|$. In this case we obtain the *barycentric Kantorovich-Rubinstein formula*.

$$\inf_{\pi\in\mathrm{Cpl}(\mu,\nu)}\int|x-\mathrm{mean}(\pi_x)|\,\mu(dx)=\sup_{\substack{\psi\colon\mathbb{R}^d\to\mathbb{R}\\ \text{1-Lipschitz, concave}}}\nu(\psi)-\mu(\psi). \tag{2.10}$$

Proof. Towards applying duality (Theorem 2.16) we compute

$$\psi^C(x)=\inf_{\rho\in\mathcal{P}_1(\mathbb{R}^d)}|x-\mathrm{mean}(\rho)|-\rho(\psi)=\inf_{z\in\mathbb{R}^d}\inf_{\substack{\rho\in\mathcal{P}_1(\mathbb{R}^d)\\ \mathrm{mean}(\rho)=z}}|x-z|-\rho(\psi)=\inf_{z\in\mathbb{R}^d}|x-z|+(-\psi)^{**}(z),$$

where we used Lemma 2.19 for the last equality. As $(x,z)\mapsto|x-z|+(-\psi)^{**}(z)$ is jointly convex, we deduce convexity of ψ^C. Moreover, ψ^C is also 1-Lipschitz with $\psi^C(x)\ge-\psi(x)$. By Example 1.27 we conclude that $\psi^C=-\psi$ if ψ is concave and 1-Lipschitz. We conclude that

$$V_C^{\mathrm{WOT}}(\mu,\nu)=\sup_{\substack{\psi\colon\mathbb{R}^d\to\mathbb{R}\\ \text{1-Lipschitz, concave}}}\nu(\psi)-\mu(\psi).$$

□

No that the barycentric Kantorovich-Rubinstein formula implies Strassen's theorem, indeed it can be seen as a quantitative version of Strassen's theorem.

2.8. **Brenier-Strassen theorem.**

Lemma 2.21. *Let C be continuous. If $\pi^*\in Cpl(\mu,\nu)$ be optimizer of $V_C^{\mathrm{WOT}}(\mu,\nu)<\infty$. Then* $\mathrm{supp}(J(\pi^*))=\mathrm{cl}(\{(x,\pi_x^*):x\in\mathsf{X}\})\subseteq\mathsf{X}\times\mathcal{P}_p(\mathsf{Y})$ *satisfies:*

when $(x_1,p_1),(x_2,p_2)\in\mathrm{supp}(J(\pi^*))$ and $q_1,q_2\in\mathcal{P}_p(\mathsf{Y})$ with $p_1+p_2=q_1+q_2$ then $C(x_1,p_1)+C(x_2,p_2)\le C(x_1,q_1)+C(x_2,q_2)$.

Proof. This can be shown similarly to Lemma 1.22. □

Theorem 2.22. *Let $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ and $\nu\in\mathcal{P}_1(\mathbb{R}^d)$. There exists a unique $\mu^*\preceq_c\nu$ s.t.*

$$\mathcal{W}_2^2(\mu^*,\mu)=\inf_{\eta\preceq_c\nu}\mathcal{W}_2^2(\eta,\mu)=\inf_{\pi\in Cpl(\mu,\nu)}\int|x-\mathrm{mean}(\pi_x)|^2\,\mu(dx). \tag{2.11}$$

Moreover, there exists a continuously differentiable convex function $\varphi:\mathbb{R}^d\to\mathbb{R}$ with $\nabla\varphi$ being 1-Lipschitz, such that $\nabla\varphi(\mu)=\mu^*$. Finally, an optimal coupling $\pi^*\in\Pi(\mu,\nu)$ for $V(\mu,\nu)$ exists, and a coupling $\pi\in\Pi(\mu,\nu)$ is optimal for $V(\mu,\nu)$ if and only if $\mathrm{mean}(\pi_x)=\nabla\varphi(x)$ *μ-a.s.*

Proof. First, we prove (2.11). If $\pi \in \mathrm{Cpl}(\mu,\nu)$, set $T(x) := \mathrm{mean}(\pi_x)$ and observe that $(\mathrm{Id}, T)_\#\mu \in \mathrm{Cpl}(\mu, T_\#\mu)$ with $T_\#\mu \preceq_c \nu$ by Jensen's inequality. We have $\mathcal{W}_2^2(\mu, T_\#\mu) \le \int |x - \mathrm{mean}(\pi_x)|^2 \,\mu(dx)$.

If $\pi^1 = \mu \otimes K^1 \in \mathrm{Cpl}(\mu,\eta)$ with $\eta \preceq_c \nu$ and $\pi^2 = \eta \otimes K^2 \in \mathrm{Cpl}^{\mathrm{M}}(\eta,\nu)$ (which exists by Strassen's theorem), then the gluing $\pi := \mu \otimes (K^1 K^2) \in \mathrm{Cpl}(\mu,\nu)$ with

$$\begin{aligned}\int |x-y|^2 \,\pi^1(dx,dy) &= \int\int |x-y|^2 K_x^1(dy)\,\mu(dx)\\ &= \int\int |x - \mathrm{mean}(K_y^2)|^2 \,K_x^1(dy)\,\mu(dx) \ge \int |x - \mathrm{mean}(K^1K_x^2)|^2\,\mu(dx),\end{aligned}$$

where we used $\mathrm{mean}(K_y^2) = y$ π^1-a.s. for the second equality and Jensen's inequality for the first inequality. We have shown (2.11).

Since $\mathcal{W}_2^2(\mu, \delta_{\mathrm{mean}(\nu)})$ is finite, so is (2.11).

If $\pi^1, \pi^2 \in \mathrm{Cpl}(\mu,\nu)$ are both optimal, so is $\frac{1}{2}(\pi^1 + \pi^2)$. By strict convexity of $|\cdot|^2$ this shows that $\mathrm{mean}(\pi_x^1) = \mathrm{mean}(\pi_x^2)$, from which we derive that there is a unique $\mu^* \preceq_c \nu$ that minimizes $\inf_{\eta \preceq_c \nu} \mathcal{W}_2^2(\mu,\eta)$.

Next, we claim that if $\pi^* \in \mathrm{Cpl}(\mu,\nu)$ then $T(x) = \mathrm{mean}(\pi_x^*)$ is μ-a.s. 1-Lipschitz. Indeed, by Lemma 2.21 we have, for $(x,\pi_x^*), (y,\pi_z^*) \in \mathrm{supp}(J(\pi^*))$ that

$$|x - T(x)|^2 + |z - T(z)|^2 \le |x - \mathrm{mean}((1-h)\pi_x^* h\pi_z^*)|^2 + |z - \mathrm{mean}((1-h)\pi_z^* + h\pi_x^*)|^2.$$

Differentiating the right-hand side w.r.t. h and evaluating at 0 yields

$$0 \le (x - T(x))\cdot(T(x) - T(z)) + (z - T(z))\cdot(T(z) - T(x)) = (x-z)\cdot(T(x) - T(z)) - |T(x) - T(z)|^2.$$

By the Cauchy-Schwartz inequality we deduce that $|T(x) - T(z)| \le |x - z|$.

So let μ^* be optimal and T be the optimal map with $T_\#\mu = \mu^*$. If $\mu \sim \mathsf{Leb}^d$, then Theorem 1.40 would tell us the existence of a convex function φ with $\nabla\varphi_\#\mu = \mu^*$, which has to be differentiable as $T = \nabla\varphi$ a.e. As T is 1-Lipschitz, we conclude that φ is differentiable with $T = \nabla\varphi$.

In general, we can approximate μ by a sequence $\mu^n \to \mu$ and $\mu^n \sim \mathsf{Leb}^d$. This can be easily achieved by scaled convolution with a non-degenerate Gaussian kernel. For each n we find convex φ_n with 1-Lipschitz $\nabla\varphi_n$. W.l.o.g. let $\varphi_n(0) = 0$. Since $\{\eta \preceq_c \nu\}$ and $(\mu_n)_n$ are both tight, there is a compact set K with

$$\inf_n \mu_n(K) \wedge \nabla\varphi_\#^n\mu_n(K) > \frac{1}{2}.$$

In particular, for every n there is $x \in K$ with $\nabla\varphi^n(x) \in K$. As K is bounded, this shows that $\sup_n \sup_{x\in K} |\nabla\varphi^n(x)| < \infty$. Therefore, $(\varphi_n)_n$ and $(\nabla\varphi_n)_n$ are both locally equicontinuous and locally bounded, which allows us to apply the Arzelà-Ascoli theorem and find subsequences such that $\varphi_{n_k} \to \varphi$ and $\nabla\varphi_{n_k} \to T$. As locally uniform convergence preserves gradient fields, we conclude that $\nabla\varphi = T$.

Finally, by lower semicontinuity we have that $\liminf_n \mathcal{W}_2^2(\mu_n, \nabla\varphi_\#^n\mu_n) \ge \mathcal{W}_2^2(\mu, \nabla\varphi_\#\mu)$. We claim that $\nabla\varphi_\#\mu$ is optimal. It suffices to show that $\limsup_n V_C^{\mathrm{WOT}}(\mu_n,\nu) \le V_C^{\mathrm{WOT}}(\mu,\nu)$. To this end, let $\pi = \mu \otimes K \in \mathrm{Cpl}(\mu,\nu)$ be optimal for V_C^{WOT} and $\mu^n \otimes K^n \in \mathrm{Cpl}(\mu^n,\mu)$ be $\mathcal{W}_2$-optimal. Define $\pi^n := \mu \otimes (K^n K) \in \mathrm{Cpl}(\mu^n,\nu)$. We have

$$\begin{aligned}\int |x - \mathrm{mean}(K^nK_x)|^2, \mu^n(dx) &= \int \left| x - \mathrm{mean}(K_x^n) + \int y - \mathrm{mean}(K_y)\,K_x^n(dy)\right|^2 \mu^n(dx)\\
&\le \int |x-y|^2 - 2(x - \mathrm{mean}(K_x^n))\cdot\Big(\int y - \mathrm{mean}(K_y)\,K_x^n(dy)\Big) + \left|\int y - \mathrm{mean}(K_y)\,K_x^n(dy)\right|^2 \mu^n(dx)\\
&\le \mathcal{W}_2^2(\mu^n,\mu) + 2\left(\int |x - \mathrm{mean}(K_x^n)|^2\,\mu^n(dx)\right)^{\frac12}\left(\int |y - \mathrm{mean}(K_y)|^2\,\mu(dy)\right)^{\frac12} + V_C^{\mathrm{WOT}}(\mu,\nu)\\
&\le \mathcal{W}_2^2(\mu^n,\mu) + 2\mathcal{W}_2(\mu^n,\mu)V_C^{\mathrm{WOT}}(\mu,\nu)^{\frac12} + V_C^{\mathrm{WOT}}(\mu,\nu).\end{aligned}$$

For $n \to \infty$ the final display converges to $V_C^{\mathrm{WOT}}(\mu,\nu)$, which shows $\limsup_n V_C^{\mathrm{WOT}}(\mu_n,\nu) \le V_C^{\mathrm{WOT}}(\mu,\nu)$. □

3. Martingale optimal transport & robust finance

3.1. **Motivation: A very quick primer in (robust) finance.** In this section we introduce the martingale optimal transport problem, that is motivated by the pricing problem in robust finance. We start with a very quick review of some of the main concepts in finance. For a more detailed introduction in mathematical finance (without robustness) we refer to the manuscripts [FS16] and [DS06] (the first chapters give a very nice and quick introduction without the technical subtleties from the general theory).

Let us consider the following simplified setup:

- we are given one risk-free asset and one risky asset, that can be traded at times $t = 0, 1, \dots, T$;
- there is no interest rate (so w.l.g. the risk-free asset can be assumed to have constant value 1), and the market is friction-less (no trading costs, both long and short positions are allowed, holding fractions of assets is allowed,...);
- the value of the risky asset at time t is denoted by S_t, and its evolution till maturity T by $S = (S_0, \dots, S_T)$;
- a derivative or option written on the underlying S is a financial security whose value/payoff depends on the evolution of S, i.e. it is a function of S. We write $f(S) = f(S_0, \dots, S_T)$.

The fundamental problem in mathematical finance is to find a fair price for $f = f(S)$. We start by noticing that, if $f \leq g$ then we expect $\mathsf{price}(f) \leq \mathsf{price}(g)$, and for any quantity $a \geq 0$, we expect $\mathsf{price}(af) = a \cdot \mathsf{price}(f)$. That is, we are looking for a functional "price" which is a linear operator on the space of all derivatives. We will see below that this will take the form of an expectation w.r.t. a probability measure in an appropriate space.

We will only consider self-financing strategies, in the sense that, given an initial amount $x \in \mathbb{R}$ invested in the two assets at time 0, at every future time $t = 1, \dots, T$ the holdings can be rebalanced between the two assets, without injecting money in the portfolio nor withdrawing from it. Then, given an initial amount x, a trading strategy will be completely defined by the holdings in the risky asset S, that we denote by $H = (H_0, \dots, H_{T-1})$ (so that $H_t \in \mathbb{R}$ are the units of S held in our portfolio between time t and $t + 1$). We will use the notation $H \cdot S$ for the gain or loss process from trading in S using the strategy H. From the self-financing property, one can see that $H \cdot S$ is a stochastic integral, so that in discrete time we have $(H \cdot S)_0 = 0$ and

$$(H \cdot S)_t := \sum_{s=0}^{t-1} H_s(S_{s+1} - S_s), \quad t = 1, \dots, T.$$

In particular, as H runs over the set of all admissible strategies, the set of random variables obtained as terminal values of such integrals, $(H \cdot S)_T$, represents all payoffs reachable starting with zero initial endowment.

Classical setting. In the classical (model-dependent) approach, to model uncertainty related to future evolution of S, one fixes a filtered probability space $(\Omega, \mathcal{F} = (\mathcal{F}_t)_{t=0}^T, \mathbb{P})$, where the filtration $\mathcal{F}$ represent the evolution of the available information, and one assumes S to be $\mathcal{F}$-adapted (i.e., the evolution $S_0, \dots, S_t$ is known at time t). Similarly, the strategy H is assumed to be $\mathcal{F}$-adapted, so that holdings at time t are decided according to the information available at that time. We denote by $\mathcal{H}$ the set of self-financing adapted strategies.

A crucial assumption in order to be able to talk of fair price in a market, is that of absence of opportunities which are "too good", in the sense of (self-financing) trading strategies that produce gains without incurring in any risk. In this spirit, we say that the market satisfies the *no arbitrage* condition, short NA, if, for any $H \in \mathcal{H}$,

$$(H \cdot S)_T \geq 0 \ \mathbb{P} - \text{a.s.} \quad \Rightarrow \quad (H \cdot S)_T = 0 \ \mathbb{P} - \text{a.s.}. \tag{NA}$$

This means that, by starting with no initial capital, and investing in the market, if we obtain an a.s. non-negative payoff, then it must be a.s. zero.

If two measures $\mathbb{Q}$ and $\mathbb{P}$ are equivalent, meaning that they have the same null-sets, we write $\mathbb{Q} \sim \mathbb{P}$. We say S is a $\mathbb{Q}$-martingale if it is a martingale under $\mathbb{Q}$. This simply means that S is $\mathbb{Q}$-integrable and that $\mathbb{E}_{\mathbb{Q}}[S_{t+1}|\mathcal{F}_t] = S_t$ for all $t = 0, \dots, T-1$. We recall one of the pillars of mathematical finance, the so-called fundamental theorem of asset pricing (FTAP), connecting (NA) to the theory of martingales.

Theorem 3.1 ([DMW90])**.** *$\mathbb{P}$ satisfies* NA *$\Leftrightarrow$ there exists $\mathbb{Q} \sim \mathbb{P}$ such that S is a $\mathbb{Q}$-martingale.*

Note that the direction $\Leftarrow$ of the proof is straightforward since any nonnegative martingale with mean zero is constant. The other direction is far less trivial.

Any measure $\mathbb{Q}$ as in Theorem 3.1 is called an equivalent martingale measure, or pricing measure. The reason for the latter is justified by the fact that fair pricing rules are exactly given by $\mathbb{E}_{\mathbb{Q}}$, for any fixed equivalent martingale measure $\mathbb{Q}$. Here we only provide the argument that shows that any such expectation provides indeed a fair pricing. We are under the assumption of NA, and we consider p being a fair price for a derivative f if by introducing the product f at price p in the market we do not introduce arbitrage. Note that, in this extended market, the notion of NA reads as

$$a(f-p) + (H \cdot S)_T \geq 0 \; \mathbb{P}-\text{a.s.} \quad \Rightarrow \quad a(f-p) + (H \cdot S)_T = 0 \; \mathbb{P}-\text{a.s.}, \tag{3.1}$$

for all $a \in \mathbb{R}$ and $H \in \mathcal{H}$. Now, if there is an equivalent martingale measure $\mathbb{Q}$ s.t. $\mathbb{E}_{\mathbb{Q}}[f] = p$, the r.h.s. of (3.1) has expectation zero under $\mathbb{Q}$, thus the implication in (3.1) is clearly satisfied. Conversely, that NA in the extended market implies that there is an equivalent martingale measure $\mathbb{Q}$ s.t. $\mathbb{E}_{\mathbb{Q}}[f] = p$, can be seen from the next theorem, that relates the two fundamental problems of pricing and hedging (i.e. replication) of financial derivatives. We recall that a derivative f is said to be replicable (resp. super-replicable) if there exists an endowment $x \in \mathbb{R}$ and a trading strategy $H \in \mathcal{H}$ s.t.

$$x + (H \cdot S)_T = f \text{ (resp. } \geq f) \;\; \mathbb{P}\text{-a.s..} \tag{3.2}$$

Theorem 3.2. *Let $\mathbb{P}$ satisfy* (NA) *and let $f \in L^1(\mathbb{P})$. Then*

$$\sup\{\mathbb{E}_{\mathbb{Q}}[f] : \mathbb{Q} \textit{ equivalent martingale measure}\} = \inf\{a \in \mathbb{R} : \exists H \in \mathcal{H}, a + (H \cdot S)_T \geq f \; \mathbb{P}\textit{-a.s.}\}.$$

The above theorems states that the "maximal fair price" for a derivative is given by the smallest initial amount needed in order to super-replicate the derivative. Similarly, the "minimal fair price" equals the biggest initial amount needed in order to sub-replicate the derivative:

$$\inf\{\mathbb{E}_{\mathbb{Q}}[f] : \mathbb{Q} \text{ equivalent martingale measure}\} = \sup\{a \in \mathbb{R} : \exists H \in \mathcal{H}, a + (H \cdot S)_T \leq f \; \mathbb{P}\text{-a.s.}\}. \tag{3.3}$$

We are interested in determining these extreme values, as they define the interval of all fair prices for the derivative $f(S)$.

Robust setting. In contrast to the classical approach, the model-independent approach lifts any assumption about an underlying probability space, and instead aims to analyse the problems of pricing and hedging only based on available market prices. Here $S = (S_0, \dots, S_T)$ can be regarded as the canonical process on $\mathbb{R}_{\geq 0}^{T+1}$.

Denote by $C_{t,k} = (S_t - k)_+$ the payoff of a European Call option with strike k and maturity t. This is the typical example of a particular, frequently traded financial derivative. Hence we assume that these options are "liquidly" traded (at time 0) and that the market gives us the price for these options, namely the function $(t,k) \mapsto p_t(k)$ is given:

$$\mathsf{price}(C_{t,k}) = p_t(k).$$

Let us fix t and check which properties the function $k \mapsto p_t(k)$ should reasonably have:

(1) As the payoff is nonnegative, we should have $p_t \geq 0$.

(2) Let $k_1 < k_2$, $\lambda \in (0,1)$, and set $k := (1-\lambda)k_1 + \lambda k_2$. Then, it holds that

$$(1-\lambda)C_{t,k_1} + \lambda C_{t,k_2} \geq C_{t,k}.$$

Consequently, for a linear pricing rule,

$$(1-\lambda)p_t(k_1) + \lambda p_t(k_2) \geq p_t(k) = p_t((1-\lambda)k_1 + \lambda k_2),$$

so that p_t is convex in k.

(3) As the stock price is non-negative, we have $C_{t,0} = S_t$, so that $p_t(0) = S_0$.

(4) For every value of S_t it holds that $\lim_{k\to\infty} C_{t,k} = 0$. Hence, $\lim_{k\to\infty} p_t(k) = 0$.

(5) For $k_1 < k_2$ we have

$$0 \leq C_{t,k_1} - C_{t,k_2} \leq k_2 - k_1,$$

and therefore, for a monotone pricing rule,

$$0 \leq p_t(k_1) - p_t(k_2) \leq k_2 - k_1,$$

so that p_t is decreasing and convex with slope at least -1 (close to 0) and at most 0 (close to ∞).

Interestingly, any function p_t satisfying these five properties is induced by a measure in the following sense:

Lemma 3.3 (Breeden-Litzenberger [BL78])**.** *Assume that $k \mapsto p(k)$ satisfies properties (1)-(5) above. Then, there exists a unique probability μ on $\mathbb{R}_+$ s.t.*

$$p(k) = \int (x-k)_+ \mu(dx).$$

Moreover,

$$p(0) = \int x\mu(dx) \quad \text{and} \quad \mu((k,\infty]) = -p'(k+),$$

where $p'(k+)$ denotes the right derivative of p at k.

Proof. By convexity of p, the right derivative exists. From (5) it is clear that $|p'(0+)| \leq 1$ and allowing an atom at 0 the function p therefore defines a unique probability measure on $\mathbb{R}_+$. Since conditions (3) and (4) above fix the boundary data the rest is straightforward, e.g. by using calculus for Riemann-Stieltjes integrals. □

As a consequence of Lemma 3.3, if the prices at a given maturity t for Call options with strike k for all $k \geq 0$ are known, then there exists a unique measure μ_t such that, for every derivative with payoff $f(S_t)$, we have

$$\mathsf{price}(f) = \int f d\mu_t.$$

The reason is that we can approximate any such f via the functions $C_{t,k}$, i.e.

$$f \sim \sum_{i=1}^{n} C_{t,k_i}.$$

Differently said, Lemma 3.3 implies that the knowledge of the prices for all Call options with maturity t uniquely defines the distribution of S_t under the measure used for pricing.

As such options are liquidly traded in $t = 0$, they can be used for hedging in a static sense, i.e., these can be bought or sold in 0 and such position kept till maturity. We will see how this explicitly appears in the super-replication theorem presented in the next section.

In the current setting, the extreme pricing values of a derivative f are given by

$$\inf / \sup\{\mathbb{E}_{\mathbb{Q}}[f] : \mathbb{Q} \text{ martingale measure}, S_i \sim_{\mathbb{Q}} \mu_i\}. \tag{3.4}$$

In the following section we will analyse this problem in some detail looking again at the "basic" questions of existence, duality (which has the interpretation of robust sub/superhedging) and characterization of optimizers.

3.2. **Existence, duality, and geometry of optimizers: discrete time.** We want to analyze (3.4) and focus for (notational) simplicity on the case of one period, so $t \in \{1, 2\}$. All the results that we present have multi-period versions, some of them are only notational more complex, while some (e.g. geometry of optimizers) are on a technical as well as on an intuitive level much more complex.

For comparison with the first chapter of these notes, we focus on the minimization problem only, and consider

$$\mathsf{X} = \mathsf{Y} = \mathbb{R},\ \mu := \mu_1 \in \mathcal{P}_1(\mathbb{R}),\ \nu := \mu_2 \in \mathcal{P}_1(\mathbb{R}),\ c(\cdot) := f(\cdot),$$

so that the object of study becomes:

$$V_c^M(\mu, \nu) := \inf_{\pi \in \mathrm{Cpl}^{\mathrm{M}}(\mu,\nu)} \int c\, d\pi, \tag{MOT}$$

where $\mathrm{Cpl}^{\mathrm{M}}(\mu, \nu) = \{\pi \in \mathrm{Cpl}(\mu, \nu) : \mathbb{E}_\pi[S_2|S_1] = S_1\}$, c.f. Definition 2.2. In analogy to (KP) we call this the Martingale Optimal Transport problem.

Remark 3.4. In the literature on MOT, one often considers the maximization problem instead of the minimization problem due to the relation to the superhedging result. However, mathematically speaking, the maximization and minimization problems are equivalent. For consistency with the rest of our manuscript, we have opted to work with the minimization problem.

Example 3.5. To familiarize with the martingale constraint, we note:

(1) Due to the martingale constraint, there cannot be Monge-type martingale couplings $T(\mu) = \nu$ other than for $T = \mathrm{Id}$, since then, for $\pi \in \mathrm{Cpl}^{\mathrm{M}}(\mu, \nu)$ with $(\mathrm{Id}, T)_{\#}\mu = \pi$,
$$T(x) = \mathrm{mean}(\pi_x) = x \quad \mu\text{-a.s.}$$
Hence, $\mathrm{Cpl}^{\mathrm{M}}(\mu, \mu) = \{(\mathrm{Id}, \mathrm{Id})_{\#}\mu\}$ consists of a single coupling whereas in general $\mathrm{Cpl}^{\mathrm{M}}(\mu, \mu)$ consists of many couplings. In general, the minimal mass splitting one can hope for is that the optimal martingale coupling is concentrated on the graph of two functions.
(2) When $\mu = \delta_x$ and $\nu \in \mathcal{P}_1(\mathbb{R})$ with $\mathrm{mean}(\nu) = x$, then $\mathrm{Cpl}(\mu, \nu) = \{\mu \otimes \nu\} = \mathrm{Cpl}^{\mathrm{M}}(\mu, \nu)$.
(3) For general martingales, there can be many martingale couplings. Let $\mu = \frac{1}{2}(\delta_{-1} + \delta_1)$ and $\nu = \frac{1}{3}(\delta_{-4} + \delta_0 + \delta_4)$ and define, for $\frac{1}{4} \le \alpha \le \frac{2}{3}$,
$$\pi^\alpha(dx, dy) := \mu \times \pi_x^\alpha \quad \text{where}$$
$$\pi_x^\alpha = \begin{cases} (\frac{2}{3} + \frac{1}{4} - \alpha)\delta_{-4} + (\frac{2}{3} - \frac{5}{4} + 2\alpha)\delta_0 + (\frac{2}{3} - \alpha)\delta_4 & x = 1, \\ (\alpha - \frac{1}{4})\delta_{-4} + (\frac{5}{4} - 2\alpha)\delta_0 + \alpha\delta_4 & x = 1. \end{cases}$$
It is straightforward to check that $\{\pi^\alpha : \frac{1}{4} \le \alpha \le \frac{2}{3}\} = \mathrm{Cpl}^{\mathrm{M}}(\mu, \nu)$, so that there are uncountably many martingale couplings with marginals μ and ν.

We recall that, by Strassen theorem (Theorem 2.20), the set $\mathrm{Cpl}^{\mathrm{M}}(\mu, \nu)$ is non-empty if and only if μ is in convex order with ν.

Lemma 3.6. *Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, then $Cpl^M(\mu, \nu)$ is convex and compact in $\mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$.*

Proof. We start the proof by characterizing the martingale constraint. Let $\pi \in \mathcal{P}_1(\mathbb{R}^d \times \mathbb{R}^d)$ with $\mathrm{proj}_1 \pi = \mu$. We have

$$\begin{aligned} \mathrm{mean}(\pi_x) = x \quad \mu\text{-a.s.} &\iff \int |x - \mathrm{mean}(\pi_x)|\ \mu(dx) = 0 \\ &\iff \forall i = 1, \dots, d,\ \int |x_i - \mathrm{mean}(\pi_x)_i|\ \mu(dx) = 0, \end{aligned}$$

where we write $y = (y_i)_{i=1}^d \in \mathbb{R}^d$. Since $|x_i| = \operatorname{sign}(x_i)x_i$, $\operatorname{sign}\colon \mathbb{R} \to \{-1, 0, 1\}$ is measurable, and $C_b(\mathbb{R}^d)$ is dense in $L^1(\mu)$, we obtain

$$\operatorname{mean}(\pi_x) = x \quad \mu\text{-a.s.} \iff \forall g \in C_b(\mathbb{R}^d; \mathbb{R}^d),\ \int g(x)(x-y)\,\pi(dx, dy) = 0.$$

The map $(x, y) \mapsto g(x)(x-y) =: h(x, y)$ is continuous with at most linear growth. Therefore, $\pi \mapsto \pi(h)$ is convex and continuous on $\mathcal{P}_1(\mathbb{R} \times \mathbb{R})$, c.f. Definition 1.51 and subsequent discussion. Since

$$\left\{\pi \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d) : \pi \text{ is martingale measure}\right\} = \bigcap_{g \in C_b(\mathbb{R}^d;\mathbb{R}^d)} \left\{\pi \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d) : \int g(x)(x-y)\,\pi(dx, dy) = 0\right\},$$

it immediately follows that the set of all martingale probabilities is convex and closed in $\mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$ (for any $p \geq 1$). Hence, as $\operatorname{Cpl}(\mu, \nu)$ is convex and by Lemma 1.53 a compact subset of $\mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$, we conclude that $\operatorname{Cpl}^{\mathrm{M}}(\mu, \nu)$ (as intersection of a convex, closed and a convex, compact set) is also convex and compact. □

As in the case of classical transport we obtain:

Corollary 3.7. *Assume that c is l.s.c. and bounded from below. Then there exists an optimal martingale coupling.*

This settles the problem of existence. Note that for these arguments neither the restriction to a single period nor to dimension one are necessary. Next we turn to the question of duality.

Theorem 3.8. *Let c be l.s.c. and bounded from below, then*

$$V_c^{MOT}(\mu, \nu) = \sup\{\mu(f) + \nu(g) : f, g, H \in C_b(\mathbb{R}) \text{ s.t. } f(x) + g(y) + H(x)(y-x) \leq c(x, y)\ \forall x, y \in \mathbb{R}\}. \quad (3.5)$$

Remark 3.9. The dual problem can be interpreted as a robust subhedging problem of the option with payoff c. In this sense, Theorem 3.8 is a robust version (i.e. independent of the reference model $\mathbb{P}$) of (3.3). The strategies used for subreplicating c in (3.5) involve trading in the asset S and in the European options f, g with maturity $t = 1$ and $t = 2$ respectively. Since $\mu = \mu_1$ and $\nu = \mu_2$ are known, the payoffs of these options can be approximated by the liquidly traded European call options with maturity $t = 1$ and $t = 2$, and their fair prices are given by $\mu(f)$ and $\nu(g)$. So, (3.5) tells us that the smallest robust fair price for the derivative c is given by the largest initial amount needed to buy the options f and g in order to sub-replicate c.

In a multi-period setting, the trading in S would be done dynamically (that is, possibly changing the portfolio in a self-financing way at every intermediate time), while the trading in the European options would still be done in a static way (i.e. with buy-and-hold strategies established at time 0, since this is the time when these are liquidly traded).

Mathematically, the term $H(x)(y-x)$ can be understood as a Lagrange multiplier accounting for the martingale constraint in the primal problem.

Proof of Theorem 3.8. Justified by Lemma 3.6 and its proof, we get by applying Theorem 1.20 with $K = \operatorname{Cpl}(\mu, \nu)$, $Y = C_b(\mathbb{R})$ and $h(\pi, H) = \int c(x, y) - H(x)(y-x)\,\pi(dx, dy)$ and Theorem 1.19, that

$$\begin{aligned} V_c^{MOT}(\mu, \nu) &= \inf_{\pi \in \operatorname{Cpl}(\mu,\nu)} \sup_{H \in C_b(\mathbb{R})} h(\pi, H) = \sup_{H \in C_b(\mathbb{R})} V_{c_H}(\mu, \nu) \\ &= \sup_{\substack{(f,g,H) \in C_b(\mathbb{R}) \times C_b(\mathbb{R}) \\ \forall (x,y) \in \mathbb{R}^2,\ f(x)+g(y) \leq c(x,y) - H(x)(y-x)}} \mu(f) + \nu(g), \end{aligned}$$

where $c_H(x, y) := c(x, y) - H(x)(y-x)$. □

The next examples illustrate that duality and dual attainment is a delicate question.

Example 3.10. (1) Let $\mu = \nu = \mathsf{Leb}_{|[0,1]}$. Then $\mathrm{Cpl}^{\mathrm{M}}(\mu,\nu) = \{\hat{Q}\}$ with $\hat{Q} = (\mathrm{Id}, \mathrm{Id})_{\#}\mu$. Let

$$c(x,y) = \mathbb{1}_{\{x=y\}} = \begin{cases} 1 & x = y \\ 0 & \text{else} \end{cases}.$$

Then $V_c^M = 1$ and we claim that $D_c^M = 0$. Indeed, let φ, ψ and h be Borel bounded s.t. $\varphi(x) + \psi(y) + H(x)(y-x) \le c(x,y)$ for all $x, y \in [0,1]$. Then $\varphi(x) + \psi(y) + H(x)(y-x) \le 0$ for all $x \neq y$. Fixing $\varepsilon > 0$, by Lusin's theorem, there is a Borel set $A \subseteq [0,1]$ with $\mu(A) > 1-\varepsilon$ s.t. $\psi_{|A}$ is continuous. Moreover, A can be chosen to be perfect (i.e., every point of A is a limit point of A). Let $x \in A$ and $(x_n)_n \subseteq A$ with $x_n \to x$ and $x_n \neq x$. Then

$$\varphi(x) + \psi(x_n) + H(x)(x_n - x) \le 0$$

for all $n \in \mathbb{N}$ and thus $\varphi(x) + \psi(x) \le 0$. As $\varepsilon > 0$ is arbitrary, $\mu(\{x : \varphi(x) + \psi(x) \le 0\}) = 1$ and hence

$$\int \varphi d\mu + \int \psi d\nu = \int \varphi(x)dx + \int \psi(x)dx \le 0.$$

A different argument: Define $T^\varepsilon(x) = \varepsilon + x \,(\mathrm{mod}\, 1)$, so that for any $\varepsilon \in (0,1)$ we have $T^\varepsilon(x) \neq x$, $T^\varepsilon(\mathsf{Leb}_{|[0,1]}) = \mathsf{Leb}_{|[0,1]}$, and $T^\varepsilon(x) - x$ is equal to ε on $[0, 1-\varepsilon]$ and equal to $\varepsilon - 1$ on $(1-\varepsilon, 1]$. Hence $\varphi(x) + \psi(T^\varepsilon(x)) \le -H(x)[T^\varepsilon(x) - x]$ and integrating w.r.t. Lebesgue we get

$$\int \varphi d\mu + \int \psi d\nu = \int \varphi(x)dx + \int \psi(x)dx \le \int_{1-\varepsilon}^1 H(x)dx - \varepsilon \int_0^1 H(x)dx \to 0.$$

In any case, a maximizing triplet is given by $\varphi = \psi = H = 0$.

(2) Consider the following setting:

$$\mu = \frac{1}{2}(\delta_{-1} + \delta_1), \quad \nu = \frac{1}{4}\mathsf{Leb}_{(-2,2)}, \text{ and } c(x,y) := \begin{cases} 0 & xy < 0, \\ -\sqrt{|xy|} & xy \ge 0. \end{cases}$$

We claim that then the right-hand side in (3.5) is not attained. Indeed, assume that $(f,g,H) \in L^1(\mu) \times L^1(\nu) \times L^1(\mu)$ with $f(x) + g(y) - H(x)(y-x) \le c(x,y)$ for all $(x,y) \in \mathbb{R}^2$. As $\mathrm{Cpl}^{\mathrm{M}}(\mu,\nu) = \{\frac{1}{2}(\delta_{-1} \otimes (\frac{1}{2}\mathsf{Leb}_{(-2,0)}) + \delta_1 \otimes (\frac{1}{2}\mathsf{Leb}_{(0,2)}) =: \pi\}$, we have that π is the optimizer and find

$$\pi(f \oplus g) = \pi(c) = 0,$$

from where we deduce (modulo a ν-null set) that

$$f(x) + g(y) + H(x)(y-x) = 0 \quad \text{on } \{(x,y) \in \{-1,1\} \times (-2,2) : xy \ge 0\}.$$

We derive that g is piece-wise affine and is given by

$$g(y) = \begin{cases} -f(1) - H(1)(y-1) & y > 0, \\ -f(-1) - H(-1)(y+1) & y < 0. \end{cases} \tag{3.6}$$

Thus, for ν-a.e. $y \in (0,2)$ we have

$$\begin{aligned} -\sqrt{y} &\ge f(1) + g(-y) + H(1)(-y-1), \\ -\sqrt{y} &\ge f(-1) + g(y) + H(-1)(y+1). \end{aligned}$$

Adding these two inequalities and using (3.6), we obtain

$$\begin{aligned} -2\sqrt{y} &\ge f(1) + f(-1) + g(y) + g(-y) + y(H(-1) - H(1)) + H(-1) - H(1) \\ &= 2y(H(-1) - H(1)), \end{aligned}$$

which is impossible to hold for all ν-a.e. $y \in (0,2)$ because

$$\lim_{y \searrow 0} -\frac{\sqrt{y}}{y} = -\infty < H(-1) - H(1).$$

We conclude that (3.5) is not attained.

The reason that duality and attainment fail in the Examples 3.10 is that the dual problem, as defined in Theorem 3.8, is too restrictive. Financially speaking: in the dual problem we hedge against scenarios that are not possible given the market observations! To clarify this, we need the concept of irreducibility.

Definition 3.11. *Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$ be in the convex order. We call the pair (μ, ν) irreducible if there exists $\pi \in Cpl^M(\mu, \nu)$ such that*

$$\forall A, B \in \mathcal{B}(\mathbb{R}^d)\colon\ \mu(A), \nu(B) > 0 \implies \pi(A \times B) > 0.$$

In one dimension, the convex order can be described via potential functions.

Definition 3.12. *For a finite measure μ on $\mathbb{R}$ with $\int |x|\,\mu(dx) < \infty$, we define its potential function $u_\mu\colon \mathbb{R} \to \mathbb{R}$ by*

$$u_\mu(y) := \int |x - y|\,\mu(dx).$$

It is possible to read off various properties of μ, as total mass, mean, atoms etc., from the behaviour of u_μ.

Lemma 3.13. *Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$. Then*

(1) $\mu \leq_{cx} \nu$ if and only if $u_\mu \leq u_\nu$.

Further, if ν is not a Dirac measure and $\mu \leq_{cx} \nu$, then

(2) (μ, ν) is irreducible if and only if $\{u_\mu < u_\nu\}$ is an interval and $\mu(\{u_\mu < u_\nu\}) = 1$.

Proof. The first claim follows from the following two observations: Firstly, the function $y \mapsto |x - y|$ is convex and secondly, and convex function can be approximated from below by convex combinations of elements in $\{f(y) = a|x - y| + by + c : a \in \mathbb{R}^+, x, b, c \in \mathbb{R}\}$.

The second claim is slightly harder. The missing parts will be shown in the next subsection.

Assume that there exists $z \in \mathbb{R}$ s.t. $u_\mu(z) = u_\nu(z)$. Then any $\pi \in \mathrm{Cpl}^{\mathrm{M}}(\mu, \nu)$ satisfies

$$\begin{aligned} u_\nu(z) \quad &= \quad \int |y - z|\,\pi(dx, dy) = \int\int |y - z|\,\pi_x(dy)\,\mu(dx) \\ &\overset{\text{Jensen}}{\geq} \quad \int |\operatorname{mean}(\pi_x) - z|\,\mu(dx) = \int |x - z|\,\mu(dx) = u_\mu(z) = u_\nu(z). \end{aligned}$$

In particular, we get equality in Jensen's inequality, and hence any $\pi \in \mathrm{Cpl}^{\mathrm{M}}(\mu, \nu)$ satisfies

$$\begin{aligned} &\pi(y \geq z \mid x > z) = 1, \\ &\pi(y \leq z \mid x < z) = 1, \\ &\pi(y = z \mid x = z) = 1 \qquad (\text{if } \mu(\{z\}) > 0\ ). \end{aligned}$$

In other words, z is a barrier for any martingale-transport plan between μ and ν, i.e. the level z cannot be strictly crossed by any such martingale. □

Theorem 3.14. *Let (μ, ν) be irreducible, compactly supported, and $c \in M_b(\mathbb{R}^2)$. Then there exists $(f, g, H) \in L^1(\mu) \times L^1(\nu) \times M(\mathbb{R})$ with $f(x) + g(y) + H(x)(y - x) \leq c(x, y)$ for μ-a.e. x and ν-a.e. y such that*

$$V_c^{\mathrm{MOT}}(\mu, \nu) = \mu(f) + \nu(g).$$

Proof. Very technical. □

Remark 3.15. Here, we presented only the case when (μ, ν) is irreducible. In one dimension and μ and ν in convex order, there is a unique decomposition of them into subprobability measures $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ such that $\mu_1 = \nu_1$, (for $k \geq 2$) (μ_k, ν_k) is irreducible, and any martingale coupling $\pi \in \mathrm{Cpl}^{\mathrm{M}}(\mu, \nu)$ can be written as

$$\pi = \sum_k \pi^k \quad \text{where } \pi^k \in \mathrm{Cpl}^{\mathrm{M}}(\mu_k, \nu_k).$$

Building on this decomposition, it is possible to also split the primal and dual problems along these decompositions.

In order to get a better understanding of irreducibility, we study a particular martingale in continuous time which induces (in case of irreducibility) a martingale coupling with the required property. In contrast to potential functions, this approach is not limited to $d = 1$.

3.3. **Stretched Brownian motion.** We have seen in Section 1 that the optimal transport problem associated with the cost function $c(x, y) = |x - y|^2$ is structurally rich. Contrary to that, this cost is not particularly interesting in the martingale world: if $\pi \in \mathrm{Cpl}_M(\mu, \nu)$ then we have

$$\begin{aligned}\int |x-y|^2\,\pi(dx,dy) &= \int |x|^2 - 2x\cdot y + |y|^2\,\pi(dx,dy)\\ &= \int |x|^2\,\mu(dx) - \int 2x\cdot\int y\,\pi_x(dy)\,\mu(dx) + \int |y|^2\,\nu(dy)\\ &= \int |y|^2\,\nu(dy) - \int |x|^2\,\mu(dx).\end{aligned}$$

So, the value of $\int |x-y|^2\,\pi(dx,dy)$ only depends on the marginals as long as π is a martingale coupling. Similarly in continuous time, when B is a d-dimensional Brownian motion and $X = \int \sigma_t\,dB_t$ with $X_0 \sim \mu$ and $X_1 \sim \nu$ for some progressively measurable square-integrable process σ, we have by Itô's isometry that

$$\mathbb{E}\left[\int_0^1 \mathrm{tr}\left(\sigma_t^2\right)\,dt\right] = \mathbb{E}\left[|X_1 - X_0|^2\right] = \int |y|^2\,\nu(dy) - \int |x|^2\,\mu(dx).$$

Another perspective on optimal transport is provided by the Benamou-Brenier formula. We recall that

$$\mathcal{W}_2^2(\mu,\nu) = \inf\left\{\int_0^1 \mathbb{E}\left[|\dot X_t|^2\right]\,dt : X_0\sim\mu, X_1\sim\nu, X\in\mathrm{AC}\right\}. \tag{3.7}$$

In the Euclidean geometry, one could interpret the right-hand side as follows: Choose as a reference motion straight lines (in direction $v \in \mathbb{R}^d$) and find a process X that connects the marginals μ and ν resembling this reference motion. We compute

$$\int_0^1 |\dot X_t - v|^2\,dt = \int_0^1 |\dot X_t|^2 - 2\dot X_t\cdot v + |v|^2\,dt = \int_0^1 |\dot X_t|^2\,dt - 2(X_1 - X_0)\cdot v + |v|^2.$$

and obtain by taking expectations that (3.7) is (up to constants) equivalent to the minimization of

$$\inf\left\{\int_0^1 \mathbb{E}\left[|\dot X_t - v|^2\right]\,dt : X_0\sim\mu, X_1\sim\nu, X\in\mathrm{AC}\right\}.$$

Arguable, in the martingale world, when speed is measured w.r.t. the quadratic variation, Brownian motion is the analogon to the straight lines of the Euclidean world. Consider a filtered probability spaces $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_t, \mathbb{P})$ with a Brownian motion B. Let X be a real-valued $\mathbb{F}$-martingale with $X_0 \sim \mu$ and $X_1 \sim \nu$ and compute

$$\mathbb{E}\left[[X-B]_1\right] = \mathbb{E}\left[[X]_1 - 2[X,B]_1 + [B]_1\right] = \mathbb{E}\left[|X_1 - X_0|^2\right] - 2\mathbb{E}\left[[X,B]_1\right] + 1 \tag{3.8}$$

Observe that, when X is of the form $X_t = \int \sigma_t\,dB_t$ for some progressively measurable σ, then, as $[X-B]_1 = \int_0^1 (\sigma_t - 1)^2\,dt$, $\mathbb{E}\left[[X,B]_1\right] = \mathbb{E}\left[\int_0^1 \sigma_t\,dt\right]$. In multiple dimensions the minimization of (3.8) is equivalent to the maximization of the trace of the covariance

$$V_{\mathrm{MBB}}(\mu,\nu) := \sup\left\{\int_0^1 \mathbb{E}\left[\mathrm{tr}(\sigma_t)\right]\,dt : X_0\sim\mu, X_1\sim\nu, X = X_0 + \int \sigma_t\,dB_t\right\}, \tag{MBB}$$

where the infimum runs over all filtered probability spaces that support a Brownian motion B and progressively measurable σ with values in the symmetric and positive semidefinite matrices. Note that as consequence of the product rule for Itô processes one could also maximize $\mathbb{E}[(X_1 - X_0)\cdot B_1]$ in (MBB).

Definition 3.16. *An optimizer of* (MBB) *is called stretched Brownian motion.*

The next example explains the wording *stretched*.

Example 3.17 (Bass martingale). Let $\mu = \delta_{\bar{x}}$, $\bar{x} \in \mathbb{R}^d$, and ν be arbitrary with mean $\bar{x}$, so that $\mu \preceq_c \nu$. Recall that γ^d denotes a d-dimensional standard Gaussian. Let $\nabla\varphi$ be the Brenier map with $\nabla\varphi_{\#}\mu = \nu$. Let $B \equiv (B_t)_{t\in[0,1]}$ be a d-dimensional standard Brownian motion with natural filtration $(\mathcal{F}_t)_{t\in[0,1]}$. Denote by P_s the heat semigroup, i.e. $P_s g(x) = \mathbb{E}[g(x + B_s)]$. Define for $t \in [0, 1]$

$$M_t := \mathbb{E}[\nabla\varphi(B_1)|\mathcal{F}_t] = \mathbb{E}[\nabla\varphi(B_1)|B_t] = P_{1-t}\nabla\varphi(B_t),$$

where the second equality follows by the Markov property and the third equality by independence of increments of Brownian motion ($B_1 = B_t + B_1 - B_t$). Observe that $M_1 = \nabla\varphi(B_1) \sim \nu$ and $M_0 = P_1 f(0) = \mathbb{E}[\nabla\varphi(B_1)] = \text{mean}(\nu) = \bar{x}$ so that $M_0 \sim \mu$. In this way, we can construct martingales between a Dirac mass and any terminal measure ν in convex order. Let X be another martingale with $X_0 \sim \mu$ and $X_1 \sim \nu$. We have

$$\mathbb{E}\left[[X, B]_1\right] = \mathbb{E}\left[(X_1 - X_0)B_1\right] = \mathbb{E}[X_1 B_1] \le \mathbb{E}[M_1 B_1] = \mathbb{E}\left[[M, B]_1\right], \tag{3.9}$$

since $X_1 \sim M_1 = \nabla\varphi(B_1)$ and the latter maximizes the covariation to B_1. Hence, the hereby constructed Bass martingale is an optimizer of $V_{\text{MBB}}(\mu, \nu)$. Further, $M = \int \sigma_t^* \, dB_t$ where σ^* can be easily derived due to Itô's formula and, when $\varphi \in C^2(\mathbb{R}^d)$, then $\sigma_t^* = P_t(H\varphi)(B_t)$.

Theorem 3.18. *The problem* (MBB) *admits an unique-in-law optimizer M^* and M^* is a continuous strong Markov martingale.*

The key to prove this result is to link it to a discrete-time optimization problem in the form of a weak optimal transport problem.

$$V_{\text{MWOT}}(\mu, \nu) := \inf_{\pi\in \text{Cpl}_M(\mu,\nu)} \int \mathcal{W}_2^2(\pi_x, \gamma^d)\, \mu(dx), \tag{MWOT}$$

where γ^d denotes the d-dimensional standard normal distribution. Note that

$$\begin{aligned} \tilde{V}_{\text{WMOT}}(\mu, \nu) :=& V_{\text{MWOT}}(\mu, \nu) - 1 - \int |y|^2 \, \nu(dy) \\ =& \frac{1}{2} \sup_{\pi\in \text{Cpl}_M(\mu,\nu)} \int \sup_{\chi\in \text{Cpl}(\pi_x,\gamma^d)} \int y \cdot z \, d\chi(y, z)\, \mu(dx). \end{aligned} \tag{3.10}$$

Theorem 3.19. *Assume $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ are in the convex order. The optimization problem* (MWOT) *and* (MBB) *are equivalent. More precisely,*

(1) $\tilde{V}_{\text{MWOT}}(\mu, \nu) = V_{\text{MBB}}(\mu, \nu) < \infty$;
(2) (MWOT) *has a unique optimizer π^*;*
(3) (MBB) *has a unique-in-law optimizer M^*;*
(4) $\pi^* = law(M_0^*, M_1^*)$ *and* $M_t^* = P_{1-t}(\nabla\varphi_{M_0^*})(B_t)$ *where $\nabla\varphi_x$ is the Brenier map between γ^d and π_x^*.*

Proof. Since $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ are in the convex order, by Strassen's theorem there is $\pi \in \text{Cpl}_M(\mu, \nu)$. Hence,

$$V_{\text{WMOT}}(\mu, \nu) \le 2 \int |y|^2 + 1 \, \pi(dx, dy) < \infty.$$

As the map $\rho \mapsto \mathcal{W}_2^2(\rho, \gamma^d)$ is $\mathcal{W}_2$-continuous, there is by Proposition 2.11 a coupling $\pi^* \in \text{Cpl}_M(\mu, \nu)$ that attains (MWOT). Uniqueness of the optimizer of (MWOT) follows from strict convexity of $\rho \mapsto \mathcal{W}_2^2(\rho, \gamma^d)$. The latter can be seen as follows: let $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By Brenier's theorem there are convex potentials φ^i with $\nabla\varphi^i_{\#}\gamma^d = \rho_i$. We have

$$\mathcal{W}_2^2\left(\frac{1}{2}(\rho_1 + \rho_2), \gamma^d\right) \le \frac{1}{2}\left(\int |z - \nabla\varphi^1(z)|^2 + |z - \nabla\varphi^2(z)| \, d\gamma^d(z)\right),$$

where equality can only hold if $\nabla\varphi^1 = \nabla\varphi^2$ γ^d-a.s. (as consequence of the uniqueness part of Brenier's theorem), which yields the claim.

Write φ^x for the Brenier potential of γ^d and π_x^*. As in Example 3.17 we construct a martingale M^* via

$$M_t^* := \mathbb{E}\left[\nabla\varphi^x(B_1)|\bar{\mathcal{F}}_t\right] = P_{1-t}(\nabla\varphi^{M_0^*})(B_t),$$

where we enlarge the natural filtration $(\mathcal{F}_t)_t$ generated by the Brownian motion B at initial time by adding an independent random variable $M_0^* \sim \mu$, i.e., $\bar{\mathcal{F}}_t = \mathcal{F}_t \vee \sigma(M_0^*)$. This is always possible on a potentially larger probability space. As the law of (M_0^*, M_1^*) maximizes (3.10), we have for any martingale $X = \int \sigma_t \, dB_t$ with $X_0 \sim \mu$ and $X_1 \sim \nu$, where σ is progressively measurable with values in the positive semidefinite matrices, that

$$\mathbb{E}\left[\int_0^1 \operatorname{tr}(\sigma_t)\, dt\right] = \mathbb{E}\left[(X_1 - X_0)B_1\right] \leq \mathbb{E}\left[(M_1^* - M_0^*) \cdot B_1\right] = \mathbb{E}\left[\int_0^1 \operatorname{tr}(\sigma_t^*)\, dt\right], \tag{3.11}$$

where σ^* is given by Itô's formula. We have shown that both problems, (MBB) and (MWOT), admit optimizers whereas the latter one is unique, and that $V_{\mathrm{MBB}}(\mu, \nu) = \tilde{V}_{\mathrm{MWOT}}(\mu, \nu)$. So let X be another optimizer of (MBB), then there is equality in (3.11) and by uniqueness of the optimizer of (MWOT) we have $(X_0, X_1) \sim (M_0^*, M_1^*)$. As in Example 3.17 we necessarily get that $X_1 = \nabla\varphi^{X_0}(B_1)$. Due to the martingale property we have $X_t = \mathbb{E}[X_1|\mathcal{F}_t] = P_{1-t}(\nabla\varphi^{X_0})(B_t)$, which a.s. determines the paths of X. In particular, $X \sim M^*$. □

Remark 3.20. The proof of Theorem 3.19 reveals the procedure how to build the optimizer of (MBB):

(1) Find the unique optimizer π^* of (MWOT).
(2) For $x \in \mathbb{R}^d$, find the Brenier potential φ^x of γ^d and π_x^*.
(3) Define $M_t^x := \mathbb{E}[\nabla\varphi^x(B_1)|B_t] = P_{1-t}\nabla\varphi^x(B_t)$.
(4) Take $X \sim \mu$ independent of B and let $M_t^* := M_t^X$.

To prove that the stretched Brownian motion M^* connecting μ and ν (with $\mu \leq_{cx} \nu$) has the strong Markov property, we establish a *dynamic programming principle*: Fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_t, \mathbb{P})$ that supports a uniform distribution U independent of $\mathcal{F}_1$. For any $(\mathcal{F}_t)_t$-stopping time $0 \leq \tau \leq 1$ we define

$$V(\tau, 1, \mu, \nu) := \sup_{\substack{M = M_0 + \int \sigma_t \, dB_t \\ M_\tau \sim \mu, M_1 \sim \nu}} \mathbb{E}\left[\int_\tau^1 \operatorname{tr}(\sigma_t)\ dt\right], \tag{3.12}$$

so that $V_{\mathrm{SBM}}(\mu, \nu) = V(0, 1, \mu, \nu)$.

Lemma 3.21 (Dynamic programming principle)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be in the convex order. For every stopping time $0 \leq \tau \leq 1$, we have*

$$V(0, 1, \mu, \nu) = \sup_{\substack{M = M_0 + \int \sigma_t \, dB_t \\ M_0 \sim \mu}} \left\{\mathbb{E}\left[\int_0^\tau \operatorname{tr}(\sigma_t)\, dt\right] + V(\tau, 1, law(M_\tau), \nu)\right\}, \tag{3.13}$$

with the convention that $\sup \emptyset = -\infty$. Further, the process M^* is the unique optimizer of $V(\tau, 1, law(M_\tau^*), \nu)$ and $law(M_1^*|M_\tau^*) = law(M_1^*|\mathcal{F}_\tau)$ almost surely. In particular, M^* is strongly Markovian.

Sketch. Obviously the left-hand side of (3.13) is smaller than the right-hand side of (3.13). To see the reverse inequality, pick $M^1 = M_0^1 + \int \sigma_t^1 \, dB_t$ such that $M_0^1 \sim \mu$ and $\operatorname{law}(M_\tau^1) \leq_{cx} \nu$. With similar reasoning as in the proof of Theorem 3.19 and Remark 3.20 (one has to properly rephrase the problem in a weak transport problem), we can construct an optimizer M^2 of $V(\tau, 1, \operatorname{law}(M_\tau^1), \nu)$ where $\operatorname{law}(M_1^2|M_\tau^2) = \operatorname{law}(M_1^2|\mathcal{F}_\tau)$. By adequately concatenating M^1 with M^2 we obtain a martingale M, admissible in $V(0, 1, \mu, \nu)$, which provides the reverse inequality.

Since M^* is the unique maximizer of $V(0, 1, \mu, \nu)$, we get from the first part, by letting $M^1 = M^*$, that M^* is also the unique maximizer of $V(\tau, 1, \operatorname{law}(M_\tau^*), \nu)$, and therefore, $\operatorname{law}(M_1^*|M_\tau^*) = \operatorname{law}(M_1^*|\mathcal{F}_\tau)$ almost surely.

Finally, M^* has the strong Markov property as by the second part and the construction of M^*, there is, for every measurable map $g \in M_b(\mathbb{R})$, a measurable map $f\colon \mathbb{R} \times [0, 1] \to \mathbb{R}$ such that for every stopping time τ,

$$f(X_\tau, \tau) = \mathbb{E}(g(X_{\tau+h})|\mathcal{F}_\tau) \quad \text{a.s.}$$

□

Proposition 3.22. *Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ in the convex order, it holds*

$$V_{\mathrm{WMOT}}(\mu,\nu) = \sup_{\substack{\psi \in C_2(\mathbb{R}^d) \\ |\cdot|^2 - \psi \text{ convex}}} \nu(\psi) + \mu(\psi^C), \tag{3.14}$$

where

$$C(x,\rho) = \begin{cases} \mathcal{W}_2^2(\rho,\gamma^d) & \mathrm{mean}(\rho) = x, \\ +\infty & \text{else,} \end{cases} \quad \text{and } \psi^C(x) = \inf_{\substack{\rho \in \mathcal{P}_2(\mathbb{R}^d) \\ \mathrm{mean}(\rho) = x}} \mathcal{W}_2^2(\rho,\gamma^d) - \rho(\psi).$$

Further, the right-hand side is attained by a dual potential ψ if and only (μ,ν) is irreducible. In this case, the stretched Brownian motion M is given by

$$M_t = \mathbb{E}[\nabla\psi(B_1)|B_t] \text{ where } B_0 \sim \nabla(P_1\psi)^*_\# \mu.$$

Proof. The duality follows from Theorem 2.16. The second part is more technical. □

4. The Skorokhod embedding problem from an optimal transport perspective

In this section we briefly introduce the Skorokhod embedding problem and explain how it can be investigated from an optimal transport perspective. In analogy to the previous sections, versions of the basic results on existence, duality and the characterization of optimality through cyclical monotonicity hold here. The proofs for this result are based on the ideas we have seen previously but are technically tedious as it is necessary to bridge the gap to the theory of stochastic processes. We therefore refer to [BCH17] for the proofs of these results and focus on the application to the Skorokhod embedding problem.

4.1. Overview.

Let B be a Brownian motion started in 0 and consider a probability μ on the real line which is centered and has second moment. The Skorokhod embedding problem is to construct a stopping time τ embedding μ into Brownian motion in the sense that

$$B_\tau \text{ is distributed according to } \mu, \quad \mathbb{E}[\tau] < \infty. \tag{SEP}$$

Here, the second condition is imposed to exclude certain undesirable solutions, and can be modified to extend to measures without a second moment. As already demonstrated by Skorokhod [Sko61, Sko65] in the early 1960s, it is always possible to construct solutions to the problem. Indeed, the survey article [Obł04] of Obłój classifies 21 distinct solutions to (SEP), although this list (from 2004) misses many more recent contributions. A common inspiration for many of these papers is to construct solutions to (SEP) that exhibit additional desirable properties or a distinct internal structure.

Recently it has been observed that (SEP) profits from an optimal transport perspective: it allows to obtain many previous developments as applications of one unifying principle (Theorem 4.3) and several difficult problems are rendered tractable. Moreover, it allows to easily handle a number of more general versions of the problem: for example, integrable measures, general starting distributions, and $\mathbb{R}^d$-valued Feller processes.

To illustrate the transport approach we introduce Root's construction, [Roo69], which is one of the earliest solutions to (SEP), and it is prototypical for many further solutions to (SEP) in that it has a simple *geometric description* and possesses a certain *optimality property* in the class of all solutions.

Root established that there exists a *barrier* $\mathcal{R}$ (which is essentially unique) such that the Skorokhod embedding problem is solved by the stopping time

$$\tau_{\mathrm{Root}} = \inf\{t \geq 0 : (t, B_t) \in \mathcal{R}\}. \tag{4.1}$$

A barrier is a Borel set $\mathcal{R} \subseteq \mathbb{R}_+ \times \mathbb{R}$ such that $(s,x) \in \mathcal{R}$ and $s < t$ implies $(t,x) \in \mathcal{R}$. The Root construction is distinguished by the following optimality property: among all solutions to (SEP) for a fixed terminal distribution μ, it minimizes $\mathbb{E}[\tau^2]$. For us, the optimality property

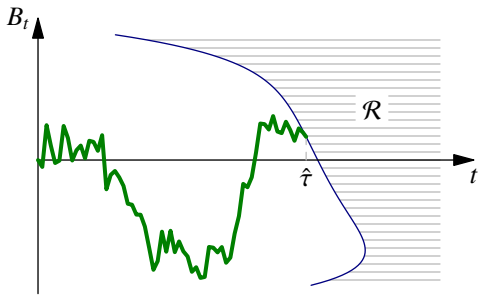


FIGURE 2. Root's solution of (SEP).

will be the starting point from which we deduce a geometric characterization of τ_{Root}. To this end, we now formalize the corresponding optimization problem.

We consider the set of *stopped paths*

$$S = \{(f, s) : f : [0, s] \to \mathbb{R} \text{ is continuous, } f(0) = 0\}. \tag{4.2}$$

Throughout the section we consider a function

$$\gamma : S \to \mathbb{R}.$$

We fix a stochastic basis $\Omega = (\Omega, \mathcal{G}, (\mathcal{G}_t)_{t\geq 0}, \mathbb{P})$ which is sufficiently rich to support a Brownian motion B and a uniformly distributed $\mathcal{G}_0$-random variable, independent of B. The *optimal Skorokhod embedding problem* is to construct a stopping time τ on Ω which optimizes

$$P_\gamma = \inf\left\{\mathbb{E}[\gamma((B_t)_{t\leq\tau}, \tau)] : \ \tau \text{ solves (SEP)}\right\}. \tag{OptSEP}$$

We emphasize that (OptSEP) does not depend on the particular choice of the underlying basis as long as it is rich enough in the above sense, cf. [BCH17, Lemma 3.11]. We will usually assume that (OptSEP) is *well posed* in the sense that $\mathbb{E}[\gamma((B_t)_{t\leq\tau}, \tau)]$ exists with values in $(-\infty, \infty]$ for all τ which solve (SEP) and is finite for one such τ.

The Root stopping time solves (OptSEP) in the case where $\gamma(f, s) = s^2$. Other examples where the solution is known include functions depending on the running maximum $\gamma((f, s)) := \bar{f}(s) := \max_{t\leq s} f(t)$ or functions of the local time at 0.

The solutions to (SEP) have their origins in many different branches of probability theory, and in many cases, the original derivation of the embedding occurred separately from the proof of the corresponding optimality properties. Moreover, the optimality of a given construction is often not immediate; for example, the optimality property of the Root embedding was first conjectured by Kiefer [Kie72] and subsequently established by Rost [Ros76].

In contrast, we will start with the optimization problem (OptSEP) and we seek a systematic method to determine the minimizer for a given function γ. To develop a general theory for this optimization problem we interpret stopping times in terms of a transport plan from the Wiener space $(C_0(\mathbb{R}_+), \mathbb{W})$ to the target measure μ, i.e. we want to think of a stopping time τ as transporting the mass of a trajectory $(B_t(\omega))_{t\in\mathbb{R}_+}$ to the point $B_{\tau(\omega)}(\omega) \in \mathbb{R}$. Note that this is *not a coupling* between $\mathbb{W}$ and μ in the usual sense and one cannot directly apply optimal transport theory. Nevertheless the transport perspective provides a useful intuition.

As in optimal transport, it is crucial to consider (OptSEP) in a suitably relaxed form, i.e. in (OptSEP) one has to optimize over *randomized stopping times* (see [BCH17, Definition 3.7]). These can be viewed as usual stopping times on a possibly enlarged probability space but in our context it is more natural to interpret them as stopping times of 'Kantorovich-type' (in the sense of optimal transport), i.e. stopping times which terminate a given path not at a single deterministic time instance but according to a distribution.

Exactly as in classical transport theory, (OptSEP) can be viewed as a linear optimization problem. The set of couplings in mass transport is compact and similarly the set of all

randomized stopping times solving (SEP) on Wiener space is compact in a natural sense. Under the standing assumption that B is defined on a sufficiently rich stochastic basis, these considerations yield:

Theorem 4.1. *Let $\gamma : S \to \mathbb{R}$ be lsc and bounded from below. Then* (OptSEP) *admits a minimizing stopping time τ.*

Here we can talk about the continuity properties of γ since S possesses a natural Polish topology.

As in the previous chapters, we have a natural dual problem.

Theorem 4.2. *Let $\gamma : S \to \mathbb{R}$ be lsc and bounded from below, and set*

$$D_\gamma = \sup\left\{ \int \psi(y)\, d\mu(y) : \psi \in C(\mathbb{R}), \exists M, \begin{array}{l} M \text{ is a continuous } \mathcal{G}\text{-martingale}, M_0 = 0 \\ \mathbb{P}-a.s., \forall t \geq 0, M_t + \psi(B_t) \leq \gamma((B_s)_{s\leq t}, t) \end{array} \right\}$$

where M, ψ satisfy $|M_t| \leq a + bt + cB_t^2$, $|\psi(y)| \leq a + by^2$ for some $a, b, c > 0$. Then we have the duality relation

$$P_\gamma = D_\gamma. \tag{4.3}$$

Next we consider a *monotonicity principle* which links the optimality of a stopping time τ with 'geometric' properties of τ. Combined with Theorem 4.1, this principle will turn out to be surprisingly powerful. For the first time, *all* the known solutions to (SEP) with optimality properties can be established through one unifying principle. Moreover, the monotonicity principle allows to treat the optimization problem (OptSEP) in a systematic manner, generating further embeddings as a by-product.

Theorem 4.3 (Monotonicity Principle)**.** *Let $\gamma : S \to \mathbb{R}$ be Borel measurable. Suppose that* (OptSEP) *is well posed and τ is an optimizer. Then there exists a γ-*monotone *(cf. Definition 4.5 below) Borel set $\Gamma \subseteq S$ such that $\mathbb{P}$-a.s.*

$$((B_t)_{t\leq\tau}, \tau) \in \Gamma\,. \tag{4.4}$$

If (4.4) holds, we will loosely say that Γ *supports* τ. The significance of Theorem 4.3 is that it links the optimality of the stopping time τ with a particular property of the set Γ, i.e. γ-monotonicity. In applications, the latter turns out to be much more tangible. We emphasize that we do not require continuity assumptions on γ in this result.

To link the optimality of a stopping time with properties of the set Γ we consider the minimization problem (OptSEP) on a pathwise level. Consider two paths $(f, s), (g, t) \in S$ which end at the same value, i.e. $f(s) = g(t)$. We want to determine which of the two paths should be *stopped* and which one should be allowed to *go on* further, bearing in mind that we try to minimize $\mathbb{E}[\gamma((B_s)_{s\leq\tau}, \tau)]$. To make this definition formal, we need to perform an operation at the level of individual paths. We will write $f \oplus h$ for the concatenation of the two paths $(f, s), (h, u) \in S$, specifically:

$$(f \oplus h)(r) := \begin{cases} f(r) & r \leq s \\ f(s) + h(r - s) & r \in (s, s + u] \end{cases}.$$

Then we set

$$\gamma^{(f,s)\oplus}(h, u) := \gamma(f \oplus h, s + u). \tag{4.5}$$

We will call $((f, s), (g, t))$ a *stop-go* pair if it is advantageous to *stop* (f, s) and to *go on* after (g, t) in the following sense:

Definition 4.4. *The pair $((f, s), (g, t)) \in S \times S$ is a* stop-go pair*, written $((f, s), (g, t)) \in \mathsf{SG}$, iff $f(s) = g(t)$ and*

$$\mathbb{E}\left[\gamma^{(f,s)\oplus}\left((B_u)_{u\leq\sigma}, \sigma\right)\right] + \gamma(g, t) \quad > \quad \gamma(f, s) + \mathbb{E}\left[\gamma^{(g,t)\oplus}\left((B_u)_{u\leq\sigma}, \sigma\right)\right] \tag{4.6}$$

for every *$(\mathcal{F}_t^B)_{t\geq 0}$-stopping time σ which satisfies $0 < \mathbb{E}[\sigma] < \infty$ and for which both sides of* (4.6) *are well defined and the left hand side is finite.*

Here $(\mathcal{F}_t^B)_{t\geq 0}$ denotes the natural filtration generated by the Brownian motion B. A consequence of considering only $(\mathcal{F}_t^B)_{t\geq 0}$-stopping times is that the set SG does not depend on the particular choice of the underlying stochastic basis.

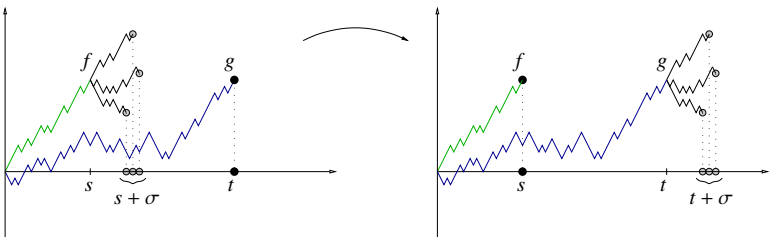


FIGURE 3. The left hand side of (4.6) corresponds to averaging the function γ over the stopped paths on the left picture; the right hand side to averaging the function γ over the stopped paths on the right picture.

The idea to relate a swapping of paths to Skorokhod embedding (as illustrated in Figure 3) was used by Hobson [Hob11, p 34] to provide a heuristic derivation of the optimality properties of the Root embedding.

Recalling (4.4), we see that the set $\Gamma \subseteq S$ contains the *stopped* paths: that is, a path (g, t) is in Γ if there is some possibility that the optimal stopping rule decides to stop at time t having observed the path $(g(u))_{u\in[0,t]}$. In addition, we need to consider those paths which we observe as the initial section of a longer, stopped, path: these are the *going* paths

$$\Gamma^< := \{(f, s) : \exists(\tilde{f}, \tilde{s}) \in \Gamma, s < \tilde{s} \text{ and } f \equiv \tilde{f} \text{ on } [0, s]\} . \tag{4.7}$$

We can now formally introduce γ-monotonicity.

Definition 4.5. *A set $\Gamma \subseteq S$ is called γ-monotone iff $\Gamma^< \times \Gamma$ contains no stop-go pairs, i.e.*

$$\mathsf{SG} \cap (\Gamma^< \times \Gamma) = \emptyset. \tag{4.8}$$

By the monotonicity principle, Theorem 4.3, an optimal stopping time is supported by a set Γ such that $\Gamma^< \times \Gamma$ contains no stop-go pair $((f, s), (g, t))$. Intuitively, such a pair gives rise to a possible modification, improving the given stopping rule: as $f(s) = g(t)$, we can imagine stopping the path (f, s) at time s, and allowing (g, t) to go on by transferring all paths which extend (f, s), the 'remaining lifetime', onto (g, t), which is now *going* (see Figure 3). By (4.6) this guarantees an improved value of P_γ, contradicting the optimality of our stopping rule. Observe that the condition $f(s) = g(t)$ is what guarantees that a modified stopping rule still embeds the measure μ. In Section 4.2 below we will briefly indicate how the monotonicity principle can be used to derive existing solutions to the Skorokhod embedding problem as well as a whole family of novel solutions to the Skorokhod embedding problem; many further examples are provided in [BCH17].

Importantly, the transport-based approach readily admits a number of strong generalizations and extensions. With only minor changes the existence result, Theorem 4.1, the duality result, Theorem 4.2, and the monotonicity principle, Theorem 4.3 below, extend to general starting distributions and Brownian motion in $\mathbb{R}^d$, and more generally to sufficiently regular Markov processes. This is notable since other constructions usually exploit rather specific properties of Brownian motion.

4.2. **Particular embeddings.** In this section we explain how Theorem 4.3 can be used to derive particular solutions to the Skorokhod embedding problem, (SEP), using the optimization problem (OptSEP). We only consider (SEP) for measures μ where $\int x^2\,\mu(dx) < \infty$. This constraint can be weakened to require only the first moment to be finite, subject to the restriction that the stopping time is *minimal*: that is, if τ is a stopping time such that $B_\tau \sim \mu$, then for any stopping time τ',

$$B_{\tau'} \sim \mu \text{ and } \tau' \leq \tau \text{ implies } \tau' = \tau \text{ a.s.} \tag{4.9}$$

In the case where μ has a second moment, minimality and $\mathbb{E}[\tau] < \infty$ are equivalent.

4.3. **The Root embedding.** We recall the definition of the Root embedding, τ_{Root}, from (4.1), and we wish to recover Root's result ([Roo69]) from an optimization problem. Remember that, according to Root's terminology, a (closed) set $\mathcal{R} \subseteq \mathbb{R}_+ \times \mathbb{R}$ is a *barrier* if $(s,x) \in \mathcal{R}$ implies $(t,x) \in \mathcal{R}$ whenever $t > s$. Then Root's construction of a solution to the Skorokhod embedding problem can be summarized as follows:

Theorem 4.6. *Let $\gamma(f,t) = h(t)$, where $h : \mathbb{R}_+ \to \mathbb{R}$ is a strictly convex function such that* (OptSEP) *is well posed. Then a minimizer of* (OptSEP) *exists, and moreover for any minimizer $\hat{\tau}$, there exists a barrier $\mathcal{R}$ such that $\hat{\tau} = \inf\{t \geq 0 : (t, B_t) \in \mathcal{R}\}$. In particular the Skorokhod embedding problem has a solution of barrier type as in* (4.1).

Proof. **Step 1.** We first pick — by Theorem 4.1 — a stopping time $\hat{\tau}$ which attains P_γ. By Theorem 4.3 there exists a set $\Gamma \subseteq S$ such that $((B_s)_{s \leq \hat{\tau}}, \hat{\tau}) \in \Gamma$ almost surely, and such that $(\Gamma^< \times \Gamma) \cap \mathsf{SG} = \emptyset$.
Step 2. Next, consider paths $(f,s),(g,t) \in S$ such that $f(s) = g(t)$. We consider when $((f,s),(g,t)) \in \mathsf{SG}$, i.e. under which conditions (f,s) should be stopped and Brownian motion should continue to go after (g,t). In the present case (4.6) amounts to

$$\mathbb{E}[h(s+\sigma)] + h(t) \quad > \quad h(s) + \mathbb{E}[h(t+\sigma)]. \tag{4.10}$$

Thus, by strict convexity of h, $((f,s),(g,t)) \in \mathsf{SG}$ iff $t < s$. We define two barriers by

$$\mathcal{R}_{\text{CL}} := \{(s,x) : \exists (g,t) \in \Gamma, g(t) = x, t \leq s\},$$
$$\mathcal{R}_{\text{OP}} := \{(s,x) : \exists (g,t) \in \Gamma, g(t) = x, t < s\}.$$

Fix $(g,t) \in \Gamma$. Then we have $(t,g(t)) \in \mathcal{R}_{\text{CL}}$. Suppose for contradiction that $\inf\{s \in [0,t] : (s,g(s)) \in \mathcal{R}_{\text{OP}}\} < t$. Then there exists $s < t$ such that $(f,s) := (g_{\restriction [0,s]}, s) \in \Gamma^<$ and $(s,f(s)) \in \mathcal{R}_{\text{OP}}$. By definition of $\mathcal{R}_{\text{OP}}$, it follows that there exists another path $(k,u) \in \Gamma$ such that $k(u) = f(s)$ and $u < s$. But then $((f,s),(k,u)) \in \mathsf{SG} \cap (\Gamma^< \times \Gamma)$ which cannot be the case. Hence,

$$(g,t) \in \Gamma \implies \inf\{s \in [0,t] : (s,g(s)) \in \mathcal{R}_{\text{CL}}\} \leq t \leq \inf\{s \in [0,t] : (s,g(s)) \in \mathcal{R}_{\text{OP}}\}.$$

Step 3. Now consider $\omega \in \Omega$ such that $(g,t) = \left((B_s(\omega))_{s \leq \hat{\tau}(\omega)}, \hat{\tau}(\omega)\right) \in \Gamma$. Then it follows immediately that:

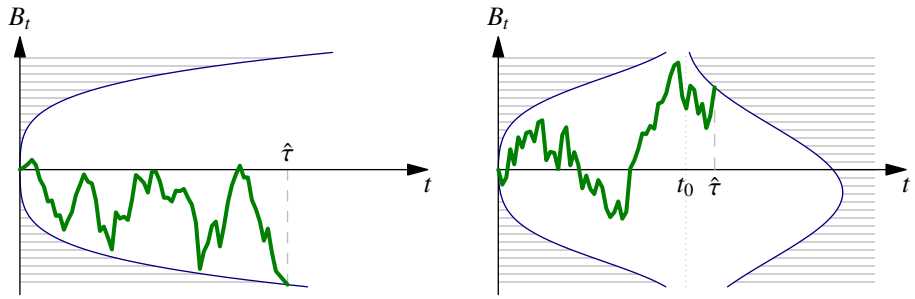
$$\tau_{\text{CL}}(\omega) := \inf\{s : (s, B_s(\omega)) \in \mathcal{R}_{\text{CL}}\} \leq \hat{\tau}(\omega) \leq \inf\{s : (s, B_s(\omega)) \in \mathcal{R}_{\text{OP}}\} =: \tau_{\text{OP}}(\omega). \tag{4.11}$$

We finally observe that $\tau_{\text{CL}} = \tau_{\text{OP}}$ a.s. by the strong Markov property, and the fact that one-dimensional Brownian motion immediately returns to its starting point. □

A consequence of this proof is that (on a given stochastic basis) there exists exactly one solution of the Skorokhod embedding problem which minimizes $\mathbb{E}[h(\tau)]$; this property was first established in [Ros76], together with the optimality property of Root's solution. To see this, assume that minimizers τ_1 and τ_2 are given. Then we can use an independent coin-flip to define a new minimizer $\bar{\tau}$ which is with probability $1/2$ equal to τ_1 and with probability $1/2$ equal to τ_2. By Theorem 4.6, $\bar{\tau}$ is of barrier type and hence $\tau_1 = \tau_2$.

Remark 4.7. We highlight here the nature of the proof of Theorem 4.6. The proof divides into three steps, two of these steps (Steps 1 and 3) being probabilistic in nature, making arguments about random variables on a particular probability space. The second step, however, is purely a pointwise argument about the properties of subsets of Γ in relation to the function γ which we look to optimize. The latter arguments are *not* probabilistic in nature.

Remark 4.8. The following argument, due to Loynes [Loy70], can be used to argue that barriers are unique in the sense that if two barriers solve (SEP), then their hitting times must be equal. Suppose that $\mathcal{R}$ and $\mathcal{S}$ are both *closed* barriers which embed μ. Note that we can take the closed barriers without altering the stopping properties. Consider the barrier $\mathcal{R} \cup \mathcal{S}$: let $A \subseteq \Omega_{\mathcal{R}} := \{x : (t,x) \in \mathcal{S} \implies (t,x) \in \mathcal{R}\}$. Then $\mathbb{P}(B_{\tau_{\mathcal{R}\cup\mathcal{S}}} \in A) \leq \mathbb{P}(B_{\tau_{\mathcal{R}}} \in A) = \mu(A)$. Similarly, for $A' \subseteq \Omega_{\mathcal{S}} := \{x : (t,x) \in \mathcal{R} \implies (t,x) \in \mathcal{S}\}$, $\mathbb{P}(B_{\tau_{\mathcal{R}\cup\mathcal{S}}} \in A') \leq \mathbb{P}(B_{\tau_{\mathcal{S}}} \in A') = \mu(A')$. Since $\mu(\Omega_{\mathcal{R}} \cup \Omega_{\mathcal{S}}) = 1$, $\tau_{\mathcal{R}\cup\mathcal{S}}$ embeds μ.



(A) The Rost construction (B) The Cave construction

FIGURE 4. The barriers corresponding to the Rost and Cave embeddings

It is known (see Monroe [Mon72]) that, when μ has a second moment, the second condition in (SEP), $\mathbb{E}[\tau] < \infty$ is equivalent to minimality of the stopping time (recall (4.9)). It immediately follows from the argument above that if the barriers $\mathcal{R}$ and $\mathcal{S}$ solve (SEP), then $\tau_{\mathcal{R}} = \tau_{\mathcal{S}}$ a.s. With minor modifications the argument of Loynes also applies to the Rost solution discussed below as well as to a number of further classical embeddings exhibiting optimality properties.

4.4. **The Rost embedding.** A set $\mathcal{R} \subseteq \mathbb{R}_+ \times \mathbb{R}$ is an *inverse barrier* if $(s, x) \in \mathcal{R}$ and $s > t$ implies that $(t, x) \in \mathcal{R}$. It has been shown by Rost [Ros76] that under the condition $\mu(\{0\}) = 0$ there exists an inverse barrier such that the corresponding hitting time (in the sense of (4.1)) solves the Skorokhod problem. It is not hard to see that without this condition some additional randomization is required. We derive this using an argument almost identical to the one above.

Theorem 4.9. *Suppose $\mu(\{0\}) = 0$. Let $\gamma(f, t) = h(t)$, where $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly* concave *function such that* (OptSEP) *is well posed. Then a minimizer $\hat{\tau}$ of* (OptSEP) *exists, and moreover for any minimizer $\hat{\tau}$, there exists an inverse barrier $\mathcal{R}$ such that $\hat{\tau} = \inf\{t \geq 0 : (t, B_t) \in \mathcal{R}\}$. In particular the Skorokhod embedding problem has a solution which is the hitting time of an inverse-barrier.*

Proof. Our proof follows closely the proof of Theorem 4.6. In particular, Steps 1 and 2 can be carried out almost verbatim to get an optimizer $\hat{\tau}$ and a *γ-monotone* set $\Gamma \subseteq S$ such that $\mathbb{P}(((B_t)_{t \leq \hat{\tau}}, \hat{\tau}) \in \Gamma) = 1$. By concavity of h, the set of stop-go pairs is now given by

$$\mathsf{SG} = \{((f, s), (g, t)) \in S \times S : f(s) = g(t), s < t\}.$$

We remove all paths (f, s) with $f(s) = 0$ from Γ, as $\mu(\{0\}) = 0$ this does not alter the full support property (or the γ-monotone property). Next we define inverse barriers by

$$\mathcal{R}_{\text{OP}} := \{(s, x) : \exists (g, t) \in \Gamma, g(t) = x, s < t\},$$
$$\mathcal{R}_{\text{CL}} := \{(s, x) : \exists (g, t) \in \Gamma, g(t) = x, s \leq t\}.$$

Denoting the respective hitting times by τ_{OP} and τ_{CL} the argument familiar from the Root case yields $\tau_{\text{CL}} \leq \hat{\tau} \leq \tau_{\text{OP}}$ a.s. and it remains to show $\tau_{\text{CL}} = \tau_{\text{OP}}$ a.s. The argument is slightly more involved than in the Root case but again entirely probabilistic:

We define $b(t) := \inf\{x > 0 : (t, x) \in \mathcal{R}_{\text{CL}}\}$, $c(t) := \sup\{x < 0 : (t, x) \in \mathcal{R}_{\text{CL}}\}$ and note that

$$\inf\{t > 0 : B_t \notin (c(t), b(t))\} \leq \tau_{\text{CL}} \leq \tau_{\text{OP}} \leq \inf\{t > 0 : B_t \notin [c(t), b(t)]\}.$$

Concentrating on the function b, we have for $\varepsilon > 0$

$$\underbrace{\inf\{t > 0 : B_t \geq b(t)\}}_{=:\sigma_b} \leq \underbrace{\inf\{t > 0 : B_t > b(t)\}}_{=:\sigma_b^+} \leq \underbrace{\inf\{t > 0 : B_t - \varepsilon t \geq b(t)\}}_{=:\sigma_b^\varepsilon}.$$

By Girsanov's Theorem, $\lim_{\varepsilon \to 0} \mathbb{P}(\sigma_b^\varepsilon \leq t) = \mathbb{P}(\sigma_b \leq t)$ for each $t \in \mathbb{R}_+$ hence $\sigma_b^+ = \sigma_b$ a.s.

Arguing likewise on c, we obtain $\tau_{\text{CL}} = \tau_{\text{OP}}$ a.s. □

As in the case of the Root embedding we obtain that the minimizer of $\mathbb{E}[h(\tau)]$ is unique.

4.5. **The cave embedding.** In this section we give an example of a new embedding that can be derived from Theorem 4.3. It can be seen as a unification of the Root and Rost embeddings. A set $\mathcal{R} \subseteq \mathbb{R}_+ \times \mathbb{R}$ is a *cave barrier* if there exists $t_0 \in \mathbb{R}_+$, an inverse barrier $\mathcal{R}^0 \subseteq [0, t_0] \times \mathbb{R}$ and a barrier $\mathcal{R}^1 \subseteq [t_0, \infty) \times \mathbb{R}$ such that $\mathcal{R} = \mathcal{R}^0 \cup \mathcal{R}^1$. We will show that there exists a cave barrier such that the corresponding hitting time (in the sense of (4.1)) solves the Skorokhod problem. We derive this using an argument similar to the one above:

Fix $t_0 \in \mathbb{R}$ and pick a continuous function $\varphi : \mathbb{R}_+ \to [0, 1]$ such that

- $\varphi(0) = 0, \lim_{t\to\infty} \varphi(t) = 0, \varphi(t_0) = 1$
- φ is strictly concave on $[0, t_0]$
- φ is strictly convex on $[t_0, \infty)$.

It follows that φ is strictly increasing on $[0, t_0]$ and strictly decreasing on $[t_0, \infty)$.

Theorem 4.10 (Cave embedding)**.** *Suppose $\mu(\{0\}) = 0$. Let $\gamma(f, t) = \varphi(t)$. Then a minimizer $\hat{\tau}$ of* (OptSEP) *exists, and moreover for any minimizer $\hat{\tau}$, there exists a cave barrier $\mathcal{R}$ such that $\hat{\tau} = \inf\{t \geq 0 : (t, B_t) \in \mathcal{R}\}$. In particular the Skorokhod embedding problem has a solution which is the hitting time of a cave barrier.*

Since this construction does not already appear in the literature, we emphasize that the result remains true for integrable (centered) measures μ (see Section 7).

Proof of Theorem 4.10. Note that since φ is bounded, the problem (OptSEP) is well posed. Following the steps of the proofs of Theorems 4.6 and 4.9, we find an optimizer $\hat{\tau}$ and a γ-monotone set $\Gamma \subseteq S$ such that $\mathbb{P}(((B_t)_{t\leq\hat{\tau}}, \hat{\tau}) \in \Gamma) = 1$. The set of stop-go pairs is given by

$$\mathsf{SG} = \{((f, s), (g, t)) \in S \times S : f(s) = g(t); s < t \leq t_0 \text{ or } t_0 \leq t < s\}.$$

Indeed, for $s < t \leq t_0$ and any $(h, r) \in S$ we have

$$\begin{aligned} \gamma((f \oplus h, s + r)) + \gamma((g, t)) \;&>\; \gamma((f, s)) + \gamma((g \oplus h, t + r)) \\ \Leftrightarrow \quad \varphi(s + r) - \varphi(s) \;&>\; \varphi(t + r) - \varphi(t) \end{aligned}$$

which holds iff $t \mapsto \varphi(t + r) - \varphi(t)$ is strictly decreasing on $[0, t_0]$ for all $r > 0$. If $t + r, t \in [0, t_0]$ this follows from concavity of φ. In the case that $t \leq t_0, t + r > t_0$ this follows since φ' is strictly positive on $[0, t_0)$ and strictly negative on (t_0, ∞). The case $t_0 \leq t < s$ can be established similarly.

Then, we define an 'open' cave barrier by

$$\mathcal{R}^0_{\text{OP}} := \{(t, x) : \exists (f, s) \in \Gamma, t < s \leq t_0\}, \quad \mathcal{R}^1_{\text{OP}} := \{(t, x) : \exists (f, s) \in \Gamma, t_0 \leq s < t\}$$

and $\mathcal{R}_{\text{OP}} := \mathcal{R}^0_{\text{OP}} \cup \mathcal{R}^1_{\text{OP}}$ (resp. a 'closed' cave barrier where we allow $t \leq s$ and $s \leq t$ in $\mathcal{R}^0_{\text{CL}}$ and $\mathcal{R}^1_{\text{CL}}$ resp.). We denote the corresponding hitting time by $\tau_{\mathcal{R}_{\text{OP}}} = \tau_{\mathcal{R}^0_{\text{OP}}} \wedge \tau_{\mathcal{R}^1_{\text{OP}}}$ (resp. $\tau_{\mathcal{R}_{\text{CL}}}$).

By the same argument as for the Root and Rost embeddings it then follows that $\tau_{\mathcal{R}_{\text{CL}}} \leq \hat{\tau} \leq \tau_{\mathcal{R}_{\text{OP}}}$ a.s. and also that $\tau_{\mathcal{R}_{\text{CL}}} = \tau_{\mathcal{R}_{\text{OP}}}$ a.s., proving the claim. □

Other recent approaches to the Root and Rost embeddings can be found in [GMO15, GOdR15, CP15, CW13]. These papers largely exploit PDE techniques, and as a consequence, are able to produce more explicit descriptions of the barriers, however the methods tend to be highly specific to the problem under consideration.

References

[AC11] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

[AG13] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, volume 2062 of *Lecture Notes in Math.*, pages 1–155. Springer, Heidelberg, 2013.

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[AH96] D. R. Adams and L. I. Hedberg. *Function spaces and potential theory*, volume 314 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996.

[AP03] L. Ambrosio and A. Pratelli. Existence and stability results in the L^1 theory of optimal transportation. In *Optimal transportation and applications (Martina Franca, 2001)*, volume 1813 of *Lecture Notes in Math.*, pages 123–160. Springer, Berlin, 2003.

[AY79] J. Azéma and M. Yor. Une solution simple au problème de Skorokhod. In *Séminaire de Probabilités, XIII (Univ. Strasbourg, Strasbourg, 1977/78)*, volume 721 of *Lecture Notes in Math.*, pages 90–115. Springer, Berlin, 1979.

[BB00] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[BC10] S. Bianchini and L. Caravenna. On optimality of c-cyclically monotone transference plans. *C. R. Math. Acad. Sci. Paris*, 348(11-12):613–618, 2010.

[BCH17] M. Beiglböck, A. Cox, and M. Huesmann. Optimal transport and Skorokhod embedding. *Invent. Math.*, 208(2):327–400, 2017.

[Bei15] Mathias Beiglböck. Cyclical monotonicity and the ergodic theorem. *Ergodic Theory Dynam. Systems*, 35(3):710–713, 2015.

[BGMS09a] M. Beiglböck, M. Goldstern, G. Maresch, and W. Schachermayer. Optimal and better transport plans. *J. Funct. Anal.*, 256(6):1907–1927, 2009.

[BGMS09b] M. Beiglböck, M. Goldstern, G. Maresch, and W. Schachermayer. Optimal and better transport plans. *J. Funct. Anal.*, 256(6):1907–1927, 2009.

[BHP13] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent bounds for option prices: A mass transport approach. *Finance Stoch.*, 17(3):477–501, 2013.

[Bil99] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[BL78] Douglas T Breeden and Robert H Litzenberger. Prices of state-contingent claims implicit in option prices. *The Journal of Business*, 51(4):621–51, 1978.

[BN15] B. Bouchard and M. Nutz. Arbitrage and duality in nondominated discrete-time models. *The Annals of Applied Probability*, 25(2):823–859, 2015.

[BS11] M. Beiglböck and W. Schachermayer. Duality for Borel measurable cost functions. *Trans. Amer. Math. Soc.*, 363(8):4203–4224, 2011.

[BVBP19] Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmund Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):1–28, 2019.

[BVBP20] Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmund Pammer. Weak monotone rearrangement on the line. *Electronic Communications in Probability*, 25, 2020.

[CP15] A.M.G. Cox and G Peskir. Embedding laws in diffusions by functions of time. *The Annals of Probability*, 43(5):2481–2510, 2015.

[CW13] A. M. G. Cox and J. Wang. Root's Barrier: Construction, Optimality and Applications to Variance Options. *Ann. Appl. Probab.*, 23(3):859–894, 2013.

[DMW90] R. C. Dalang, A. Morton, and W. Willinger. Equivalent martingale measures and no-arbitrage in stochastic securities market models. *Stochastics Stochastics Rep.*, 29(2):185–201, 1990.

[DS06] F. Delbaen and W. Schachermayer. *The mathematics of arbitrage*. Springer Finance. Springer-Verlag, Berlin, 2006.

[DS14a] Y. Dolinsky and H. M. Soner. Martingale optimal transport and robust hedging in continuous time. *Probab. Theory Relat. Fields*, 160(1-2):391–427, 2014.

[DS14b] Y. Dolinsky and M. H. Soner. Robust hedging with proportional transaction costs. *Finance Stoch.*, 18(2):327–347, 2014.

[DS15] Y. Dolinsky and H. M. Soner. Martingale optimal transport in the Skorokhod space. *Stochastic Processes and their Applications*, 125(10):3893–3931, 2015.

[FG21] Alessio Figalli and Federico Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. EMS Textbooks in Mathematics. EMS Press, Berlin, [2021] ©2021.

[FS16] Hans Föllmer and Alexander Schied. Stochastic finance. In *Stochastic Finance*. de Gruyter, 2016.

[GHT14] A. Galichon, P. Henry-Labordère, and N. Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Ann. Appl. Probab.*, 24(1):312–336, 2014.

[GM96] W. Gangbo and R. McCann. The geometry of optimal transportation. *Acta Math.*, 177(2):113–161, 1996.

[GMO15] P. Gassiat, A. Mijatović, and H. Oberhauser. An integral equation for Root's barrier and the generation of Brownian increments. *Ann. Appl. Probab.*, 25(4):2039–2065, 2015.

[GOdR15] P. Gassiat, H. Oberhauser, and G. dos Reis. Root's barrier, viscosity solutions of obstacle problems and reflected FBSDEs. *Stochastic Processes and their Applications*, 125(12):4601–4631, 2015.

[HLOST16] P. Henry-Labordère, J. Obłój, P. Spoida, and N. Touzi. The maximum maximum of a martingale with given n marginals. *Ann. Appl. Probab.*, 26(1):1–44, 2016.

[HN12] D. Hobson and A. Neuberger. Robust bounds for forward start options. *Math. Finance*, 22(1):31–56, 2012.

[Hob98] D. Hobson. Robust hedging of the lookback option. *Finance and Stochastics*, 2:329–347, 1998.

[Hob11] D. Hobson. The Skorokhod embedding problem and model-independent bounds for option prices. In *Paris-Princeton Lectures on Mathematical Finance 2010*, volume 2003 of *Lecture Notes in Math.*, pages 267–318. Springer, Berlin, 2011.

[HPRY11] F. Hirsch, C. Profeta, B. Roynette, and M. Yor. *Peacocks and associated martingales, with explicit constructions*, volume 3 of *Bocconi & Springer Series*. Springer, Milan; Bocconi University Press, Milan, 2011.

[Jac88] S. Jacka. Doob's inequalities revisited: A maximal h^1-embedding. *Stochastic processes and their applications*, 29(2):281–290, 1988.

[Kec95] A. S. Kechris. *Classical descriptive set theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[Kel84] H.G. Kellerer. Duality theorems for marginal problems. *Z. Wahrsch. Verw. Gebiete*, 67(4):399–432, 1984.

[Kie72] J. Kiefer. Skorohod embedding of multivariate RV's, and the sample DF. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 24(1):1–35, 1972.

[KS84] M. Knott and C. S. Smith. On the optimal mapping of distributions. *J. Optim. Theory Appl.*, 43(1):39–49, 1984.

[Lis09] Stefano Lisini. Nonlinear diffusion equations with variable coefficients as gradient flows in wasserstein spaces. *ESAIM: Control, Optimisation and Calculus of Variations*, 15(3):712–740, 2009.

[LMT14] G. Last, P. Mörters, and H. Thorisson. Unbiased shifts of Brownian motion. *Ann. Probab.*, 42(2):431–463, 2014.

[Loy70] R. M. Loynes. Stopping times on Brownian motion: Some properties of Root's construction. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 16:211–218, 1970.

[Mon81] G. Monge. Memoire sur la theorie des deblais et des remblais. *Histoire de l'académie Royale des Sciences de Paris*, 1781.

[Mon72] I. Monroe. On embedding right continuous martingales in Brownian motion. *Ann. Math. Statist.*, 43:1293–1311, 1972.

[Obł04] J. Obłój. The Skorokhod embedding problem and its offspring. *Probab. Surv.*, 1:321–390, 2004.

[OST15] J. Obłój, P. Spoida, and N. Touzi. Martingale inequalities for the maximum via pathwise arguments. In *In Memoriam Marc Yor-Séminaire de Probabilités XLVII*, pages 227–247. Springer, 2015.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[Per86] E. Perkins. The Cereteli-Davis solution to the H^1-embedding problem and an optimal embedding in Brownian motion. In *Seminar on stochastic processes, 1985*, pages 172–223. Springer, 1986.

[Roo69] D. H. Root. The existence of certain stopping times on Brownian motion. *Ann. Math. Statist.*, 40:715–718, 1969.

[Ros76] H. Rost. Skorokhod stopping times of minimal variance. In *Séminaire de Probabilités, X (Première partie, Univ. Strasbourg, Strasbourg, année universitaire 1974/1975)*, pages 194–208. Lecture Notes in Math., Vol. 511. Springer, Berlin, 1976.

[Rüs91] L. Rüschendorf. Fréchet-bounds and their applications. In *Advances in probability distributions with given marginals (Rome, 1990)*, volume 67 of *Math. Appl.*, pages 151–187. Kluwer Acad. Publ., Dordrecht, 1991.

[Rüs95] L. Rüschendorf. Optimal solutions of multivariate coupling problems. *Appl. Math. (Warsaw)*, 23(3):325–338, 1995.

[San15] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[Sko61] A. V. Skorohod. *Issledovaniya po teorii sluchainykh protsessov (Stokhasticheskie differentsialnye uravneniya i predelnye teoremy dlya protsessov Markova)*. Izdat. Kiev. Univ., Kiev, 1961.

[Sko65] A. V. Skorokhod. *Studies in the theory of random processes*. Translated from the Russian by Scripta Technica, Inc. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1965.

[Str65] V. Strassen. The existence of probability measures with given marginals. *Ann. Math. Statist.*, 36:423–439, 1965.

[Str85] H. Strasser. *Mathematical theory of statistics*, volume 7 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1985. Statistical experiments and asymptotic decision theory.

[Val83] P. Vallois. Le probleme de Skorokhod sur $\mathbb{R}$: une approche avec le temps local. In *Séminaire de Probabilités XVII 1981/82*, pages 227–239. Springer, 1983.

[Vil03] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.