STOCHASTIC MASS TRANSFER

JULIO BACKHOFF-VERAGUAS AND MARTIN HUESMANN

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. martingale optimal transport (MOT) and causal/adapted optimal transport (COT).

The aim of this lecture 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 its martingale variant which is motivated by intriguing questions from robust/model independent finance. In the second part of the lecture we will complement the worst case point of view of MOT on robust finance by a "local" approach. This will naturally lead us to adapted versions of the OT problem, the COT, which we will explore in detail. Our discussion will be guided by examples from finance and stochastic analysis.

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 $\lambda \in \mathcal{P}(\mathsf{X})$ we denote the image measure of λ under T by $T(\lambda) = T_\sharp \lambda = \lambda \circ T^{-1}$
- The Lebesgue measure will be denoted by Leb.
- The set of all couplings between two probability measures μ, ν will be denoted by $\mathsf{Cpl}(\mu, \nu)$.
- $C_b(\mathsf{X})$ denotes the set of 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. At the end of this section we will shortly hint at some of the fascinating applications of optimal transport in analysis of PDEs, geometry, and beyond.

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

1.1. The Monge and Kantorovich optimal transport problem.

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 $\mathcal{P}(\mathsf{X}), \mathcal{P}(\mathsf{Y})$. Fix a Borel measurable function $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R} \cup \{\infty\}$. 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.

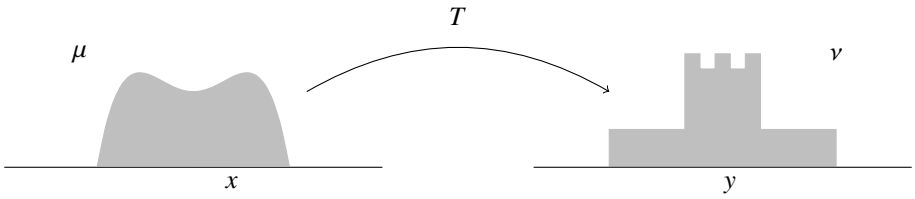


FIGURE 1. A possible transport from a distribution μ to a distribution ν via a map T. The cost from x to y is given via some cost function $c(x, y)$; total cost: $\int c(x, T(x))\mu(dx)$.

Given two distributions of mass $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ 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*

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

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.

Definition 1.3. *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.*

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. His motivation originated from engineering and he considered the special case of $c(x, y) = |x - y|$ in $\mathbb{R}^d$.

This problem is very difficult due to various reasons. It is a non-linear problem in the unknown T. More seriously, it can be ill-posed:

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 any map T s.t. $T(\mu) = \nu$.

Example 1.5. 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

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

as follows by change of variables. 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, it is difficult to find conditions ensuring the existence of at least one (optimal) map T transporting μ to ν. Moreover, the constraint $T(\mu) = \nu$ is not weakly sequentially closed w.r.t. any reasonable topology. For instance, let $f : \mathbb{R} \to \mathbb{R}$ be a one-periodic function which equals 1 on $[0, 1/2)$ and -1 on $[1/2, 1)$. Let $f_n(x) = f(nx)$. Put $\mu = \mathsf{Leb}_{|[0,1]}$ and $\nu = \frac{1}{2}(\delta_1 + \delta_{-1})$. Then, $f_n(\mu) = \nu$ for all n but f_n converges weakly to the zero function $\bar{f} \equiv 0$ so that $\lim_n f_n(\mu) = \nu \neq \bar{f}(\mu) = \delta_0$. These observations show two severe limitations which one would like to overcome in a relaxed version of the Monge-problem. To get an idea on how this might look like let us consider the following example.

Example 1.6. Let $\mathsf{X} = \mathsf{Y} = \mathbb{R}^2$ and $c(x,y) = |x-y|^2$. Let $\mu = \mathsf{Leb}^1_{|\{0\}\times[0,1]}$ and $\nu = \frac{1}{2}\left(\mathsf{Leb}^1_{|\{-1\}\times[0,1]} + \mathsf{Leb}^1_{|\{1\}\times[0,1]}\right)$. Then, $P_c^M = 1$ but no optimal transport map exists. Indeed, for any candidate map T it holds that $\int |x - T(x)|^2 \, \mu(dx) \geq 1$. By looking at dyadic subdivisions of $\{0\} \times [0,1]$ it is not difficult to construct transport maps T_n with $|x - T_n(x)|^2 \leq 1 + 2^{-2n}$. Integrating w.r.t. μ shows that $P_c^M = 1$. However, there cannot be any map $T : \mathbb{R}^2 \to \mathbb{R}^2$ such that $T(\mu) = \nu$ and $|T(x) - x| = 1$ μ-a.e. so that no optimal map exists. (Indeed, assume optimal T exists, s.t. $T(0,y) \in \{(-1,y),(1,y)\}$ μ a.s. Put $A^{\pm} = \{y : T(0,y) = (\pm 1, y)\}$. Then, $\mathsf{Leb}^1(A^+) + \mathsf{Leb}^1(A^-) = 1$ and $T(\mu) = \mathsf{Leb}^1_{|\{-1\}\times A^-} + \mathsf{Leb}^1_{|\{1\}\times A^+}$. Assume $\mathsf{Leb}^1(A^+) > 0$. It then follows, that $T(\mu)(\{1\}\times A^+) = \mathsf{Leb}^1(A^+) \neq \frac{1}{2}\mathsf{Leb}^1(A^+) = \nu(\{1\}\times A^+)$, a contradiction.)

In the last example it is very intuitive how the optimal solution should look like. It should split the mass at each point $(0,y)$ into two equal pieces and send one piece to the "right" and one to the "left". Hence, we should relax the constraint that each point $x \in \mathsf{X}$ is transported to exactly one $y \in \mathsf{Y}$ and allow for multivalued transport maps or even "continuously"-valued transport maps. This is best captured in the language of couplings of probability measures.

Definition 1.7. *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$. A coupling of μ and ν is a measure $q \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$ with marginals μ and ν, i.e.*

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

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

Stochastically, a coupling q of μ and ν is a joint law of two random variables (X,Y) such that $\mathsf{Law}_q(X) = \mu$ and $\mathsf{Law}_q(Y) = \nu$. In particular, conditioning on $X = x$ we can interpret the regular conditional probability $q(\cdot|X = x)$ as a plan on how to transport the mass at x. Therefore, we will often call couplings by the name *transport plans*. Analytically, this corresponds to disintegrating q w.r.t. its first marginal μ to obtain a family of probability measures $(q_x(dy))_{x\in\mathsf{X}}$ (see Theorem A.1). Writing $\mathrm{proj}_{\mathsf{X}} : \mathsf{X}\times\mathsf{Y} \to \mathsf{X}, (x,y) \mapsto x, \mathrm{proj}_{\mathsf{Y}} : \mathsf{X}\times\mathsf{Y} \to \mathsf{Y}, (x,y) \mapsto y$ a measure $q \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$ is an element of $\mathsf{Cpl}(\mu,\nu)$ iff $\mathrm{proj}_{\mathsf{X}}(q) = \mu$ and $\mathrm{proj}_{\mathsf{Y}}(q) = \nu$.

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

Definition 1.8. *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ and c a cost function. The Kantorovich problem is to solve*

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

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

Remark 1.9. The optimal coupling in Example 1.6 is given by $q = \frac{1}{2}(q_{T_+} + q_{T_-})$ where $T_{\pm}(0,y) = (\pm 1, y)$.

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

Remark 1.10.
- The set $\mathsf{Cpl}(\mu,\nu)$ is always non-empty. The product coupling (stochastically, the independent coupling) $\mu\otimes\nu \in \mathsf{Cpl}(\mu,\nu)$.
- The set $\mathsf{Cpl}(\mu,\nu)$ is convex.
- The map $q \mapsto \int c\, dq$ 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 $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}(\mathsf{Z})$ on a Polish space Z converges weakly to $\mu \in \mathcal{P}(\mathsf{Z})$ iff

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

where $C_b(\mathsf{Z})$ denotes the continuous and bounded functions on Z. We call the induced topology on $\mathcal{P}(\mathsf{Z})$ the weak topology. For us, Z will usually be X, Y or $\mathsf{X}\times\mathsf{Y}$.

Theorem 1.11 (Prokhorov)**.** *Let Z be a Polish space. A family $A \subseteq \mathcal{P}(\mathsf{Z})$ of probability measures on Z is relatively compact w.r.t. the weak topology iff it is tight, i.e. for every $\varepsilon > 0$ there exists $K_\varepsilon \subseteq \mathsf{Z}$ compact such that*

$$\sup_{\mu\in A} \mu(\mathsf{Z} \setminus K_\varepsilon) \leq \varepsilon.$$

For a proof we refer to [Bil99].

Lemma 1.12. *If $A_1 \subseteq \mathcal{P}(\mathsf{X}), A_2 \subseteq \mathcal{P}(\mathsf{Y})$ are tight so is $A_3 := \{q \in \mathcal{P}(\mathsf{X}\times\mathsf{Y}) : \text{proj}_{\mathsf{X}}(q) \in A_1$ and $\text{proj}_{\mathsf{Y}}(q) \in A_2\}$.*

Proof. Let $q \in A_3$ 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

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

□

Corollary 1.13. *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 1.12. It remains to show that it is closed. Pick $(q_n)_{n\in\mathbb{N}} \subseteq \mathsf{Cpl}(\mu,\nu)$ with limit q. We have to show that q 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

$$\int \varphi dq = \int \bar\varphi dq = \lim_n \int \bar\varphi dq_n = \lim_n \int \varphi dq_n = \int \varphi d\mu$$

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

As an important consequence of this corollary we have the following result on existence of optimal couplings. For the definition of lower semi-continuity see Appendix B.

Theorem 1.14. *Assume that c is lower semi-continuous and bounded from below. Then there exists a minimizer q^* to* (KP)*, i.e. $q^* \in \arg\min_{q\in\mathsf{Cpl}(\mu,\nu)} \int c dq$.*

Proof. The proof follows by the direct method of the calculus of variations.

Observe that the map $q \mapsto \int c dq$ is lower semi-continuous by the Portmanteau theorem. (If you have not seen this, see Lemma B.1.) Take a minimizing sequence $(q_n)_{n\in\mathbb{N}}$ i.e. $\int c dq_n \to P_c^K(\mu,\nu)$. By Corollary 1.13, there is $q \in \mathsf{Cpl}(\mu,\nu)$ such that up to passing to a subsequence $q_n \to q$. By the first part of the proof it follows that

$$\int c dq \leq \liminf_n \int c dq_n = P_c^K(\mu,\nu).$$

Hence, q is an optimal coupling. □

In Theorem 1.14 the condition of lower boundedness can be suitably relaxed. However, lower semi-continuity is important. Consider for instance $\mathsf{X} = \mathsf{Y} = [0,1]$, $\mu = \nu$ uniformly distributed, and $c(x,y) = 1$ on $\{y \geq x\}$ and 0 otherwise: then $P_c^K = 0$, so if q is optimal we must have $q(y < x) = 1$, leading to a contradiction since $\int |y-x| q(dx,dy) = \int x\mu(dx) - \int y\nu(dy) = 0$ would imply $q(y = x) = 1$.

Remark 1.15. The Monge optimizer, if it exists, can be strictly worse than the Kantorovich optimizer. Indeed, consider the cost function $c(x, y) = |x - y|$ for $\mathsf{X} = \mathsf{Y} = \mathbb{R}$, and the measures $\mu = \frac{1}{3}\delta_0 + \frac{2}{3}\delta_1$, $\nu = \frac{2}{3}\delta_0 + \frac{1}{3}\delta_1$. Then the only admissible Monge map is $T(0) = 1$, $T(1) = 0$. Observe that $c(x, T(x)) = 1$. On the other hand the coupling $q = \frac{1}{3}\delta_{0,0} + \frac{1}{3}\delta_{1,0} + \frac{1}{3}\delta_{1,1}$ has marginals μ and ν but its cost is only $\frac{1}{3}$.

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 monotonicity (more precisely, c-cyclical monotonicity) and the dual problem.

1.2. **The dual problem and characterization of optimal couplings.** We aim to find necessary and sufficient conditions for a coupling to be optimal. Let us first look at a discrete example.

Example 1.16. Let $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. Let $q = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)} \in \mathsf{Cpl}(\mu, \nu)$. Fix a cost function c. Then, q is optimal iff for each permutation $\sigma \in S_n$ it holds that

$$\int c dq = \frac{1}{n}\sum_{i=1}^n c(x_i, y_i) \leq \frac{1}{n}\sum_{i=1}^n c(x_i, y_{\sigma(i)}).$$

The necessity is obvious since otherwise we can easily construct a competitor by rerouting the transport $(x_1 \mapsto y_{\sigma(1)}, x_2 \mapsto y_{\sigma(2)}, ...)$. The sufficiency follows from Choquet's theorem as in Problem 4.(c) at the end of this section.

This observation leads us to the following definition.

Definition 1.17. *A set $\Gamma \subseteq \mathsf{X} \times \mathsf{Y}$ is called c-cyclically monotone iff for all $n \in \mathbb{N}$ and for all n-tuples $(x_1, y_1), \ldots, (x_n, y_n) \in \Gamma$*

$$\sum_{i=1}^n c(x_i, y_i) \leq \sum_{i=1}^n c(x_i, y_{\sigma(i)}) \quad \text{for all } \sigma \in S_n. \tag{1.1}$$

A coupling $q \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ is called c-cyclically monotone iff it is concentrated on a c-cyclically monotone set, i.e. there is a c-cyclically monotone set Γ such that $q(\Gamma) = 1$.

Observe that c-cyclical monotonicity is a pointwise definition. As we will see several times during this course this often induces geometric constraints on Γ. Observe also that the definition remains the same if in (1.1) we take only one fixed permutation (which is not the identity): for instance the shift $\sigma(i) = i + 1 \bmod n$. Finally, note that in general a c-cyclically monotone coupling need not have a c-cyclically monotone support. Recall that the support of a measure q is defined as $\{x : q(U) > 0, \text{ for all neighbourhoods } U \text{ of } x\}$. If q is a finite measure the support is the smallest closed set of full measure. We write $\mathrm{supp}(q)$ for the support of q. If c is continuous a c-cyclically monotone coupling has a c-cyclically monotone support.

Lemma 1.18 (Necessary condition for optimality)**.** *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$. Assume that c is continuous and bounded from below such that $c(x, y) \leq a(x) + b(y)$ for $a \in L^1(\mu), b \in L^1(\nu)$.*[1] *If $q^* \in \mathsf{Cpl}(\mu, \nu)$ is an optimal coupling (w.r.t. the cost function c), then* $\mathrm{supp}(q^*)$ *is c-cyclically monotone (so in particular q^* is c-cyclically monotone).*

[1]Note that there is a small subtlety here since L^1 is defined via equivalence classes of functions. The condition $a \in L^1$ should be interpreted as a is an integrable measurable function.

Proof. Since for $q \in \mathsf{Cpl}(\mu,\nu)$ it holds that

$$\int c(x,y)dq(x,y) \leq \int a(x)+b(y)dq(x,y) = \mu(a)+\nu(b) < \infty,$$

it follows that $c \in L^1(q)$ for all $q \in \mathsf{Cpl}(\mu,\nu)$. In particular P_c^K is finite.

We will argue by contradiction leveraging the observation from Example 1.16 to the general case using the continuity of c.

Assume, there exists $n > 1, \sigma \in S_n$ and $(x_1,y_1),\dots,(x_n,y_n) \in \mathrm{supp}(q^*)$ such that

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

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

$$\sum_{i=1}^n c(u_i,v_i) > \sum_{i=1}^n c(u_i,v_{\sigma(i)}).$$

In the next step, we will use this property to construct a competitor $\bar{q}$ of q^* with strictly lower transport cost. To this end, consider $\Omega = \Pi_{i=1}^n U_i \times V_i, m_i = q^*(U_i \times V_i) > 0$ (since $(x_i,y_i) \in \mathrm{supp}(q^*)$), $P = \bigotimes_{i=1}^n \frac{1}{m_i} q^*_{|U_i \times V_i}$ and define

$$\bar{q} = q^* + \frac{\min_j m_j}{n} \sum_{i=1}^n \left((\mathrm{proj}_{U_i}, \mathrm{proj}_{V_{\sigma(i)}})(P) - (\mathrm{proj}_{U_i}, \mathrm{proj}_{V_i})(P) \right).$$

Observe, that $\bar{q} \in \mathsf{Cpl}(\mu,\nu)$ and $\int c d\bar{q} < \int c dq^*$ by construction so that q^* is not optimal. □

We have already seen that the functional $q \mapsto \int c dq$ in the Kantorovich optimization problem (KP) is linear in the variable q so that the Kantorovich problem is a linear optimization problem over the compact and convex set $\mathsf{Cpl}(\mu,\nu)$. As such there is a dual point of view, a dual problem, which is very useful to pin down optimizers.

Let us first take an intuitive approach:

Example 1.19. We interpret the measure μ as the distribution of breweries and ν as the distribution of pubs and supermarkets. Then, $c(x,y)$ can be interpreted as the cost of transporting a good (here beer) from the brewery at x to the pub at y. There is a dual economic problem. Assume there is an agent who offers to do the transport for the brewers. They charge a price $\varphi(x)$ for picking up the goods at location x and charge a price $\psi(y)$ to drop off this good at location y. Then, this offer is competitive iff $\varphi(x)+\psi(y) \leq c(x,y)$. This leads us to the following dual problem and a number of natural questions: Are there competitive dual strategies? If yes, what are the best and what can we learn from them?

Definition 1.20 (Dual problem)**.** *For $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ the dual problem is to maximize*

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

among all $\varphi \in C_b(\mathsf{X}), \psi \in C_b(\mathsf{Y})$ 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 $D_c^K := D_c^K(\mu,\nu)$.

In the inequality $\varphi(x)+\psi(y) \leq c(x,y)$ one can also allow functions $\varphi \in L^1(\mu), \psi \in L^1(\nu)$. It is immediate that $D_c^K \leq P_c^K$ since for any candidates φ, ψ, q it follows from the marginal constraint on q that

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

In fact the other inequality holds as well.

Theorem 1.21 (Duality)**.** *Let $\mu \in \mathcal{P}(\mathsf{X}), \nu \in \mathcal{P}(\mathsf{Y})$ and c be continuous and bounded from below. Assume that there exists $a \in L^1(\mu), b \in L^1(\nu)$ s.t. $c(x,y) \le a(x) + b(y)$. Then,*

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

The previous theorem states that there is *no duality gap* in the Kantorovich problem under the stated conditions. We will only give a sketch of a proof here. Below this theorem will be a consequence of another result. However, this sketch can be made rigorous using convex analysis, see [San15, Section 1.6.3]. Denote by

$$\mathcal{D}(c) := \{(\varphi,\psi) \in C_b(\mathsf{X}) \times C_b(\mathsf{Y}) : \varphi(x) + \psi(y) \le c(x,y)\}$$

the set of admissible dual variables.

Sketch of proof of duality. As in Lemma 1.18 we obtain that P_c^K is finite. Write

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

Then, we can (formally) write assuming " inf sup = sup inf "

$$\begin{aligned} &\inf_{q\in\mathsf{Cpl}(\mu,\nu)} \int c dq = \inf_{q\in\mathcal{M}(\mathsf{X}\times\mathsf{Y})} \int c dq + \chi(q) \\ &= \inf_{q\in\mathcal{M}(\mathsf{X}\times\mathsf{Y})} \sup_{(\varphi,\psi)\in C_b(\mathsf{X})\times C_b(\mathsf{Y})} \int c(x,y) - \varphi(x) - \psi(y) dq + \mu(\varphi) + \nu(\psi) \\ &= \sup_{(\varphi,\psi)\in C_b(\mathsf{X})\times C_b(\mathsf{Y})} \inf_{q\in\mathcal{M}(\mathsf{X}\times\mathsf{Y})} \int c(x,y) - \varphi(x) - \psi(y) dq + \mu(\varphi) + \nu(\psi) \end{aligned}$$

Observe that

$$\inf_{q\in\mathcal{M}(\mathsf{X}\times\mathsf{Y})} \int c(x,y) - \varphi(x) - \psi(y) dq = \begin{cases} 0 & \text{if } (\varphi,\psi) \in \mathcal{D}(c) \\ -\infty & \text{else} \end{cases}.$$

Hence, the claim follows. □

Remark 1.22. The duality result holds well beyond the case of continuous cost functions c. Using standard approximation methods it is possible to deduce duality for l.s.c. cost functions bounded from below. But the result holds even for Borel measurable cost functions, see [BS11]. In the literature one often proves duality first and then derives c-cyclical monotonicity of the optimizers as a corollary. To illustrate this, suppose that (φ,ψ) are optimal for D_c^K and q is optimal for P_c^K. It follows that $\varphi(x) + \psi(y) = c(x,y)$, q-a.s. It is easy to check that $\Gamma := \{(x,y) : \varphi(x) + \psi(y) = c(x,y)\}$ is c-cyclically monotone and so q must likewise be c-cyclically monotone.

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

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

Then, $(\tilde{\varphi},\psi) \in \mathcal{D}(c)$ and since $\varphi \le \tilde{\varphi}$ it follows that $\mu(\varphi) \le \mu(\tilde{\varphi})$ 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 $f(x) + \psi(y) \le c(x,y)$. 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*

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

We define the $\bar{c}$-transform of $\psi : \mathsf{Y} \to \mathbb{R}$ by

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

We say a function $\psi : \mathsf{Y} \to \mathbb{R}$ is c-concave if $\psi = \varphi^c$ for some φ and analogously for $\bar{c}$-concave functions. (If $\mathsf{X} = \mathsf{Y}$ and c is symmetric the distinction between c and $\bar{c}$ plays no role and we will drop it.)

From the previous considerations it is clear that in the dual problem we can restrict to pairs of $\bar{c}$- and 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 $\varphi^{c\bar{c}} := (\varphi^c)^{\bar{c}} \geq \varphi$. We have $\varphi^{c\bar{c}} = \varphi$ iff φ is $\bar{c}$-concave (i.p. for any $\varphi : \mathsf{X} \to \mathbb{R}$ it holds that $\varphi^{c\bar{c}c} := ((\varphi^c)^{\bar{c}})^c = \varphi^c$.); in general, $\varphi^{c\bar{c}}$ is the smallest $\bar{c}$-concave function larger than φ.*

Proof. Exercise. □

Importantly, c-concave functions share many properties of convex (concave) functions. Therefore, also the terminology is inspired by convex analysis.

Definition 1.25. *Let $\varphi : \mathsf{X} \to \mathbb{R}$ be $\bar{c}$-concave. Its $\bar{c}$-superdifferential is defined as*

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

We also write $\partial^{\bar{c}}\varphi(x) = \{y : (x,y) \in \partial^{\bar{c}}\varphi\}$. Similarly, we define the c-superdifferential of a c-concave function $\psi : \mathsf{Y} \to \mathbb{R}$.

Remark 1.26. We have $y \in \partial^{\bar{c}}\varphi(x)$ iff $\varphi(x) = c(x,y) - \varphi^c(y)$ and $\varphi(z) \leq c(z,y) - \varphi^c(y)$ for all $z \in \mathsf{X}$.

Example 1.27. All these examples will be "symmetric" so that we do not distinguish between c-concave and $\bar{c}$-concave.

a) Let $\mathsf{X} = \mathsf{Y}$, $c(x,y) = d(x,y)$ be a distance. Then φ is c-concave iff φ is 1-Lipschitz, i.e. $|\varphi(x) - \varphi(y)| \leq |x - y|$, and $\varphi^c = -\varphi$.
b) $\mathsf{X} = \mathsf{Y} = \mathbb{R}^n$, $c(x,y) = -x \cdot y$, the standard Euclidean inner product. Then, φ is c-concave iff φ is concave and u.s.c. In particular, the c-superdifferential of φ is precisely the classical superdifferential $\partial\varphi$ of φ from convex analysis.
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.

The following result is a crucial step towards an identification of the geometric structure of optimal couplings. It links the geometric concept of c-cyclical monotonicity with the property of c-concavity of functions that show up in the dual problem.

Proposition 1.28. *Assume c is real valued. A non-empty set $\Gamma \subseteq \mathsf{X} \times \mathsf{Y}$ is c-cyclically monotone iff $\Gamma \subseteq \partial^{\bar{c}}\varphi$ for some $\bar{c}$-concave function φ.*

Proof. First observe that $\partial^{\bar{c}}\varphi$ is c-cyclical monotone. Indeed, for any $n \in \mathbb{N}$ and tuples $(x_1,y_1),\dots,(x_n,y_n) \in \partial^{\bar{c}}\varphi$ it follows for $\sigma \in S_n$ by definition of $\bar{c}$-concavity that

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

To show the converse, we will explicitly construct a $\bar{c}$-concave function φ with the desired properties. To this end, fix $(x_0,y_0) \in \Gamma$. For $x \in \mathsf{X}$ set:[2]

$$\begin{aligned}\varphi(x) := \inf\{&c(x,y_n) - c(x_n,y_n) + c(x_n,y_{n-1}) - c(x_{n-1},y_{n-1}) + \dots + \\ &+ c(x_1,y_0) - c(x_0,y_0) : n \in \mathbb{N}; (x_1,y_1),\dots,(x_n,y_n) \in \Gamma\}\end{aligned} \tag{1.2}$$

[2]One can argue measurability via continuity of c or via $\bar{c}$-concavity of φ. In general this is more difficult.

Since c is real valued and $\Gamma \neq \emptyset$ we have $\varphi < \infty$. First note that φ is not identical $-\infty$ since $\varphi(x_0) = 0$. Indeed, by choosing $n = 1$, $(x_1, y_1) = (x_0, y_0)$ in (1.2) we see $\varphi(x_0) \leq 0$. However, the c-cyclical monotonicity of Γ implies that

$$c(x_0, y_n) - c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + \ldots + c(x_1, y_0) - c(x_0, y_0) \geq 0$$

so that $\varphi(x_0) \geq 0$. Next, writing

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

we see that

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

(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 $\Gamma \subseteq \partial^{\bar{c}}\varphi$ it is sufficient to show $\varphi(x) + \varphi^c(y) \geq c(x, y)$ on Γ since the other inequality follows from $\bar{c}$-concavity. Since $\varphi^c = \psi^{\bar{c}c} \geq \psi$ (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 $\varphi = \psi^{\bar{c}}$ 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. □

This 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 1.29 (Fundamental theorem of OT; characterization of optimizers; monotonicity principle of OT)**.** *Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ be continuous and bounded from below such that $c(x, y) \leq a(x) + b(y)$ for some $a \in L^1(\mu), b \in L^1(\nu)$. Let $q \in \mathsf{Cpl}(\mu, \nu)$. Then the following are equivalent:*

i) *q is an optimal coupling;*
ii) *the support $\mathrm{supp}(q)$ of q is c-cyclical monotone;*
iii) *there exists a $\bar{c}$-concave function φ with $\varphi \vee 0 \in L^1(\mu)$ s.t. $\mathrm{supp}(q) \subseteq \partial^{\bar{c}}\varphi$.*

Proof. **i) ⇒ ii):** This follows by Lemma 1.18.

ii) ⇒ iii): By Proposition 1.28 there is a $\bar{c}$-concave φ such that $\mathrm{supp}(q) \subseteq \partial^{\bar{c}}\varphi$. Moreover, from the construction it follows that (with the notation from the proof of Proposition 1.28)

$$\varphi(x) \leq c(x, y_0) - c(x_0, y_0) \leq a(x) + b(y_0) - c(x_0, y_0).$$

Hence, $\varphi \vee 0 \in L^1(\mu)$.

iii) ⇒ i): Pick any $\tilde{q} \in \mathsf{Cpl}(\mu, \nu)$. We will show that $\int c dq \leq \int c d\tilde{q}$. By construction of φ it holds that

$$\begin{aligned} \varphi(x) + \varphi^c(y) &= c(x, y), \quad \text{for all } (x, y) \in \mathrm{supp}(q) \\ \varphi(x) + \varphi^c(y) &\leq c(x, y), \quad \text{for all } x \in \mathsf{X}, y \in \mathsf{Y}. \end{aligned}$$

Hence,

$$\begin{aligned} \int c(x, y) dq(x, y) &= \int \varphi(x) + \varphi^c(y) dq(x, y) = \int \varphi(x) d\mu(x) + \int \varphi^c(y) d\nu(y) \\ &= \int \varphi(x) + \varphi^c(y) d\tilde{q}(x, y) \leq \int c(x, y) d\tilde{q}(x, y). \end{aligned}$$

□

Remark 1.30. One can strengthen the result of Theorem 1.29. In fact, the support of any optimal coupling $q^* \in \mathsf{Cpl}(\mu, \nu)$ is contained in $\partial^{\bar{c}}\varphi$. Indeed, with q and φ as in Theorem

1.29 we can argue

$$\int \varphi(x)d\mu(x)+\int \varphi^c(y)d\nu(y) = \int \varphi(x)+\varphi^c(y)dq^*(x,y) \leq \int c(x,y)dq^*(x,y)$$
$$= \int c(x,y)dq(x,y) = \int \varphi(x)+\varphi^c(y)dq(x,y) = \int \varphi(x)d\mu(x)+\int \varphi^c(y)d\nu(y),$$

so that equality holds throughout. By $\bar{c}$-concavity of φ this immediately implies that $\varphi(x)+\varphi^c(y) = c(x,y)$ for q^*-a.e. (x,y), i.e. $\text{supp}(q^*) \subseteq \partial^{\bar{c}}\varphi$.

Remark 1.31. The crucial insight of Theorem 1.29 is that

optimality of a given coupling is solely a property of the geometry of its support

i.e. how the mass is exactly distributed over the support is of less importance (as long as the marginals are matched). In particular, if q is optimal and $\tilde{q}$ another probability with $\text{supp}(\tilde{q}) \subseteq \text{supp}(q)$, then it is an optimal coupling between its own 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 $\bar{c}$-concave function φ such that for all $x \in \mathsf{X}$ it holds that $T(x) \in \partial^{\bar{c}}\varphi(x)$, then for any $\mu \in \mathcal{P}(\mathsf{X})$ the map T is optimal between μ and $T(\mu)$ (up to integrability issues 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.

Another immediate consequence of the above theorem (see Problems 18-19) is that the limit of a convergent sequence of optimizers (between their marginals) is an optimizer (between its marginals). This is of course important for practical implementations.

Theorem 1.32 (Duality)**.** *Let $c : \mathsf{X}\times\mathsf{Y} \to \mathbb{R}$ be continuous and bounded from below such that $c(x,y) \leq a(x)+b(y)$ for some $a \in L^1(\mu), b \in L^1(\nu)$. Then,*

$$\inf_{q\in\mathsf{Cpl}(\mu,\nu)} \int cdq = \sup\{\mu(\varphi)+\nu(\psi)\},$$

where the supremum runs over all $\varphi \in L^1(\mu), \psi \in L^1(\nu)$ s.t. $\varphi(x)+\psi(y) \leq c(x,y)$. Furthermore, the supremum is attained for a pair (φ,φ^c) for some $\bar{c}$-concave function φ.

Proof. **$\geq$:** This follows from the observation

$$\int \varphi d\mu + \int \psi d\nu = \int (\varphi(x)+\psi(y))q(dx,dy) \leq \int cdq$$

valid for any admissible φ, ψ and q.

$\leq$: Together with attainment this is a direct consequence of Theorem 1.29. □

Definition 1.33. *A $\bar{c}$-concave function φ such that the pair (φ,φ^c) is a maximizing pair for the dual problem is called a $\bar{c}$-concave Kantorovich potential, or Kantorovich potential, of the measures μ,ν.*

As mentioned in Remark 1.22 the duality for the optimal transport problem holds for quite general cost functions (see Problem 8). However, the existence of primal or dual optimizers is not guaranteed in general.

Example 1.34.

a) *(Kantorovich-Rubinstein formula)* Let $\mathsf{X} = \mathsf{Y}$, $c(x,y) = d(x,y)$ a distance. Assume that $\int d(x,x_0)\mu(dx) < \infty$, and the same for ν, for some (and then all) $x_0 \in \mathsf{X}$. Then:

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

Note that the right hand side immediately implies that $\inf_{q\in\mathsf{Cpl}(\mu,\nu)} \int cdq =: W_1(\mu,\nu)$ 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 $W_1(\mu+\eta,\nu+\eta) = W_1(\mu,\nu)$.

b) $\mathsf{X} = \mathsf{Y}$, $c(x,y) = \begin{cases} 0 & \text{if } x = y; \\ 1 & \text{else .} \end{cases}$ Then,

$$\inf_{q \in \mathsf{Cpl}(\mu,\nu)} \int c dq = \sup_{0 \le \varphi \le 1} \int \varphi (d\mu - d\nu) = \|\mu - \nu\|_{TV},$$

where $\|\mu - \nu\|_{TV} := \sup_{A \in \mathcal{B}(\mathsf{X})} |\mu(A) - \nu(A)|$.

Similarly, the implication that optimal couplings are concentrated on c-cyclically monotone sets is true in a remarkable generality. The reverse implication is more intricate. For instance, there are transport plans concentrated on c-cyclically monotone sets which are not optimal (see [AP03]). The proofs of these results go slightly beyond the scope of this course. However, if c is real valued and Borel measurable there is a nice and comparably short probabilistic argument (see [Bei12]) showing that transport plans concentrated on c-cyclically monotone sets are optimal. The proof can be skipped on a first reading:

Theorem 1.35. *Let $c : \mathsf{X} \times \mathsf{Y} \to [0, \infty)$ be a Borel measurable cost function and $q \in \mathsf{Cpl}(\mu, \nu)$ be concentrated on a c-cyclically monotone set Γ satisfying $\int c dq < \infty$. Then q is optimal.*

The key tool is the pointwise ergodic theorem (see e.g. [Kal02, Theorem 9.6]):

Theorem 1.36. *Let (Z, κ) be a probability space and $\sigma : \mathsf{Z} \to \mathsf{Z}$ measure preserving, i.e. $\sigma(\kappa) = \kappa$. Then, for every $f \in L^1(\kappa)$ the limit*

$$f^* = \lim_n \frac{1}{n} \sum_{i=0}^{n-1} f \circ \sigma^i \tag{1.3}$$

exists almost surely and in $L^1(\kappa)$.

Ergodic theorems are powerful tools to obtain limit results. They can be interpreted as generalizations of the law of large numbers for iid or mixing random variables. However note that f^* need not be a constant in general which is why we will integrate (1.3) below.

Proof of Theorem 1.35. Let $q, \tilde{q} \in \mathsf{Cpl}(\mu, \nu)$ be finite-cost transport plans with $q(\Gamma) = 1$. We will show that $\int c dq \le \int c d\tilde{q}$. Put $\mathsf{Z} = (\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}$ and consider the shift mapping

$$\sigma : \mathsf{Z} \to \mathsf{Z}; \quad (x_i, y_i)_{i=1}^{\infty} \mapsto (x_{i+1}, y_{i+1})_{i=1}^{\infty}.$$

Define the projections $P, Q : \mathsf{Z} \to \mathsf{X} \times \mathsf{Y}$ by

$$P((x_i, y_i)_{i=1}^{\infty}) = (x_1, y_1), \quad Q((x_i, y_i)_{i=1}^{\infty}) = (x_1, y_2).$$

We claim that there exists a measure κ on Z such that $\sigma(\kappa) = \kappa, P(\kappa) = q$, and $Q(\kappa) = \tilde{q}$. Indeed, identify Z with the product

$$\mathsf{Y}^{(1)} \times \mathsf{X}^{(1)} \times \mathsf{Y}^{(2)} \times \mathsf{X}^{(2)} \times \dots,$$

where $\mathsf{Y}^{(i)}, \mathsf{X}^{(i)}$ are copies of Y and X. Let $(q_y)_{y \in \mathsf{Y}}$ be a disintegration of q w.r.t. ν and $(\tilde{q}_x)_{x \in \mathsf{X}}$ a disintegration of $\tilde{q}$ w.r.t. μ. Then, we let κ be the distribution of the Markov chain starting according to ν and with transition kernel $(q_y)_{y \in \mathsf{Y}}$ to move from $\mathsf{Y}^{(i)}$ to $\mathsf{X}^{(i)}$ and transition kernel $(\tilde{q}_x)_{x \in \mathsf{X}}$ to move from $\mathsf{X}^{(i)}$ to $\mathsf{Y}^{(i+1)}$.

Put $f := c \circ Q - c \circ P \in L^1(\kappa)$. We need to show that $\int f d\kappa = \int c d\tilde{q} - \int c dq \ge 0$. Applying the ergodic theorem to f and integrating over (1.3) yields

$$\int f d\kappa = \int f^* d\kappa = \int \lim_n \frac{1}{n} \sum_{i=0}^{n-1} f \circ \sigma^i d\kappa.$$

Unravelling the definition of f and σ this gives

$$\int c d\tilde{q} - \int c dq = \int \left(\lim_n \frac{1}{n} \sum_{i=0}^{n-1} c(x_i, y_{i+1}) - c(x_i, y_i) \right) d\kappa((x_i, y_i)_i). \tag{1.4}$$

Hence, it suffices to show that the integrand on the r.h.s. is κ-a.s. non-negative. Observe that $\kappa(\Gamma \times (\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}) = q(\Gamma) = 1$. Moreover, $\sigma^{-n}(\Gamma \times (\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}) = (\mathsf{X} \times \mathsf{Y})^n \times \Gamma \times (\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}$

so that $\Gamma^{\mathbb{N}} = \bigcap_{n\geq 0} \sigma^{-n}(\Gamma \times (\mathsf{X}\times\mathsf{Y})^{\mathbb{N}})$. By σ-invariance of κ it follows that $\kappa(\Gamma^{\mathbb{N}}) = 1$ it is sufficient to restrict to sequence $(x_i, y_i)_i \in \Gamma^{\mathbb{N}}$. If c is bounded c-cyclical monotonicity of Γ immediately implies that

$$\begin{aligned}&\liminf_n \frac{1}{n}\sum_{i=1}^{n} c(x_i, y_{i+1}) - c(x_i, y_i)\\ &= \liminf_n \frac{1}{n}\left(\sum_{i=1}^{n-1} c(x_i, y_{i+1}) - c(x_i, y_i) + c(x_n, y_1) - c(x_n, y_n) + c(x_n, y_{n+1}) - c(x_n, y_1)\right) \geq 0\end{aligned}$$

for any $(x_i, y_i)_i \in \Gamma^{\mathbb{N}}$. In the general case fix $(x_0, y_0) \in \Gamma$. For $(x_i, y_i)_{i\geq 1} \in \Gamma^{\mathbb{N}}$ we have by cyclical monotonicity (thinking $y_{n+2} = y_0$) that for any n using $c \in [0, \infty)$

$$c(x_0, y_1) + \sum_{k=1}^{n} (c(x_k, y_{k+1}) - c(x_k, y_k)) + c(x_{n+1}, y_0) \geq c(x_0, y_0) + c(x_{n+1}, y_{n+1}) \geq 0.$$

Since $c(x_0, y_1)/n \to 0$ this implies that

$$\liminf_n \frac{1}{n}\sum_{k=1}^{n} (c(x_k, y_{k+1}) - c(x_k, y_k)) + \frac{c(x_{n+1}, y_0)}{n} \geq 0. \tag{1.5}$$

In particular, if we are able to show that $\liminf_n \frac{c(x_{n+1},y_0)}{n} = 0$ κ-a.s. equation (1.5) implies the desired κ-a.s. positivity of the integrand in (1.4). To this end, put $g((x_i, y_i)_i) := c(x_1, y_0)$ so that we have $g \circ \sigma^n = c(x_{n+1}, y_0)$. Since g is finitely valued $g/n \to 0$ in $\kappa-$measure and, since σ is measure preserving, $g \circ \sigma^n/n \to 0$. Passing to a subsequence if necessary, the convergence holds κ-a.s. concluding the proof. □

1.3. **Uniqueness and existence of optimal transport maps.** In the last section we have seen that there is always an optimal coupling as soon as the cost function is l.s.c. and sufficiently bounded. In Theorem 1.29 we have seen that optimal couplings are concentrated on c-cyclical monotone sets. In this section we want to use this information to find conditions under which the optimal coupling is induced by an optimal transport map. In particular, in these cases the solutions to the Kantorovich problem and the Monge problem coincide and we are sometimes able to deduce uniqueness of the optimal coupling.

Remark 1.37. The cost function must play a role in these questions! For instance, if $\mathsf{X} = \mathsf{Y} = \mathbb{R}$, μ and ν are arbitrary, but the supremum of the support of μ is smaller than the infimum of the support of ν, and $c(x, y) = |x - y|$, then any coupling with these marginals is optimal for the Kantorovich problem. On the other hand, the marginal measures must also play a role in these questions! For instance, if now $\mathsf{X} = \mathsf{Y} = \mathbb{R}^2$, $c(x, y) = |x - y|^2$, $\mu = 1/2\delta_{(0,-1)} + 1/2\delta_{(0,1)}$ and $\nu = 1/2\delta_{(-1,0)} + 1/2\delta_{(1,0)}$, then any coupling with these marginals is optimal for the Kantorovich problem. Hence we must pay attention to all the data of the problem.

By Theorem 1.29 we know that an optimal coupling is concentrated on the superdifferential $\partial^{\bar{c}}\varphi$ of a $\bar{c}$-concave function φ. In particular, if we can show that for μ-a.e. $x \in \mathsf{X}$ the set $\partial^{\bar{c}}\varphi(x)$ is single-valued any optimal coupling q needs to be induced by a transport map.

Theorem 1.38 (Brenier, Rüschendorf)**.** *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, i.e.* $\int |x|^2 d\mu(x) + \int |x|^2 d\nu(x) < \infty$. *Moreover, assume that* $\mu \ll \mathsf{Leb}$.

Then, there is a unique optimal coupling $q^* \in \mathsf{Cpl}(\mu, \nu)$. *This optimizer is of the form* $q* = (\mathrm{Id}, \nabla\bar{\varphi})(\mu)$ *for some convex function* $\bar{\varphi} : \mathsf{X} \to \mathbb{R}$.

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* $\tilde{T}(\mu)$.

Proof. The assumptions of Theorem 1.29 are satisfied with $a(x) = b(x) = |x|^2$. Hence, there exists a c-concave function φ such that $\mathrm{supp}(q) \subseteq \partial^c\varphi$ for any optimal coupling q. By Example 1.27, the function $\bar{\varphi}(x) = \frac{|x|^2}{2} - \varphi(x)$ is convex. 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 $\nabla\bar{\varphi}(x) = x - \nabla\varphi(x)$ exists μ-a.e. Furthermore, on the set of differentiability of $\bar{\varphi}$ it holds that $\nabla\bar{\varphi}(x) = y$ iff $y \in \partial^c\varphi(x)$. Hence, any optimal coupling q is concentrated on the graph of $\nabla\bar{\varphi}$.

This immediately implies that the optimal coupling is unique. Indeed, assume there are two optimal couplings q_1, q_2 both concentrated on the graph of some maps T_1, T_2. Then, $q_3 = \frac{1}{2}(q_1 + q_2)$ is optimal again by linearity of the Kantorovich problem. By the first part of the theorem, q_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, any $\nabla\bar{\varphi}$ with $\nabla\bar{\varphi}(\mu) = \nu$ is optimal since the graph defines a c-cyclical monotone set. By uniqueness, we can conclude. □

Remark 1.39 (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.

Remark 1.40. 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 1.41. The question of regularity of the optimal transport maps is an interesting story in itself which goes way beyond the scope of this course. Obviously the least we must assume is that the marginals have a convex support. If additionally μ and ν have $\alpha-$Hölder continuous densities then the optimal map is $C^{1,\alpha}$. If these 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 1.42 (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 $h \geq 0$.

Theorem 1.43 (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 $q^* = q_d^* + q_o^*$ with $q_d^* = q^*_{|\{(x,x):x\in\mathbb{R}^d\}}$. Then, $q_d^* = (\mathrm{Id},\mathrm{Id})(\mu\wedge\nu)$ 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.*

Remark 1.44. If μ, ν are two measures, then $\mu\wedge\nu$ denotes their minimum (or common mass) in the lattice sense: $\mu\wedge\nu$ is a measure, $\mu\wedge\nu(A) \leq \min\{\mu(A), \nu(A)\}$ for all measurable A, and if ρ satisfies this then $\rho(\cdot) \leq \mu\wedge\nu(\cdot)$. It can be checked that $\mu\wedge\nu(A) = \inf_B\{\mu(B) + \nu(A\backslash B)\}$ with the inf running over the measurable subsets of A. If μ, ν have densities then so does $\mu\wedge\nu$ and its density is the pointwise minimum of these.

The crucial idea to prove the above theorem 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, $q_d \leq (\mathrm{Id},\mathrm{Id})(\mu\wedge\nu)$. 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. $h(r) = |r|^p$, $p > 1$, and assume that the Kantorovich problem is finite. 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 $x_1 \leq x_2$ or $x_2 \leq x_1$? 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., (1.1) with $n = 2$ implies

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

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 q living on Γ to be the quantile coupling/monotone rearrangement between its marginals (see Problem 11). More precisely, if q has marginals μ and ν with cumulative distribution functions F_μ and F_ν, then we must have

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

Notice that in this one-dimensional case we essentially did not have to assume anything on the marginal distributions to obtain uniqueness. However, if e.g. μ is atomless the optimizer is of Monge type with an explicit description:

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

Observe, that in the one-dimensional case we only considered cyclical monotonicity using 2-cycles. A set $\Gamma \subseteq \mathsf{X}\times\mathsf{Y}$ satisfying (1.1) 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. Furthermore, in higher dimensions the optimizer(s) will typically depend on the function h.

1.4. **Wasserstein distance and interpolation of probability measures.** For various applications of optimal transport a key object are the Kantorovich-Wasserstein distances W_p. They inherit various geometric properties of the base space and induce an useful interpolation of probability measures.

We consider a Polish space X with a compatible metric d and cost functions $c(x,y) = d^p(x,y)$, $p \geq 1$. We will consider product spaces X^n and denote by $\mathrm{proj}_i : \mathsf{X}^n \to \mathsf{X}$ the projection onto the i-th coordinate. Similarly, $\mathrm{proj}_{i,j}$ denotes the projection onto the i-th and j-th coordinate.

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 L^p Wasserstein distance W_p is defined for $\mu, \nu \in \mathcal{P}_p(\mathsf{X})$ as*

$$\begin{aligned} W_p(\mu,\nu) &= \left(\inf_{q \in \mathsf{Cpl}(\mu,\nu)} \int d^p(x,y) q(dx,dy) \right)^{\frac{1}{p}} \\ &= \left(\int d^p(x,y) q^*(dx,dy) \right)^{\frac{1}{p}} \end{aligned}$$

for any optimal coupling q^*.

Let us show that W_p is in fact a distance. The most difficult part is the triangle inequality. To show this we will rely on the following result.

Lemma 1.48 (Glueing lemma)**.** *Let $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ be three Polish spaces and $q_1 \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$, $q_2 \in \mathcal{P}(\mathsf{Y} \times \mathsf{Z})$ be two probability measures such that $\mathrm{proj}_{\mathsf{Y}}(q_1) = \mathrm{proj}_{\mathsf{Y}}(q_2)$. Then there exists a measure $q_3 \in \mathcal{P}(\mathsf{X} \times \mathsf{Y} \times \mathsf{Z})$ such that*

$$\mathrm{proj}_{\mathsf{X},\mathsf{Y}}(q_3) = q_1, \quad \text{and} \quad \mathrm{proj}_{\mathsf{Y},\mathsf{Z}}(q_3) = q_2.$$

Proof. Use the disintegration theorem to write with $\mu(dy) = \mathrm{proj}_{\mathsf{Y}}(q_1)(dy)$ the measures $q_1(dx,dy) = (q_1)_y(dx)\mu(dy)$, $q_2(dy,dz) = (q_2)_y(dz)\mu(dy)$ and conclude putting

$$q_3(dx,dy,dz) = (q_1)_y(dx)\mu(dy)(q_2)_y(dz).$$

□

This allows us to show that W_p are distances:

Theorem 1.49. *W_p defines a distance on $\mathcal{P}_p(\mathsf{X})$.*

Proof. Since $d(x,y) = d(y,x)$ it follows that $W_p(\mu,\nu) = W_p(\nu,\mu)$ and $W_p(\mu,\mu) = 0$. If $W_p(\mu,\nu) = 0$ there is a coupling q^* (note that d^p is continuous and bounded from below so that we have existence of optimal couplings) 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. By the gluing lemma there exists a measure $q_3 \in \mathcal{P}(\mathsf{X}^3)$ such that $\mathrm{proj}_{1,2}(q_3) = q_1, \mathrm{proj}_{2,3}(q_3) = q_2$. Moreover,

$\text{proj}_{1,3}(q_3) \in \mathsf{Cpl}(\mu_1, \mu_3)$. Hence, it follows from the triangle inequality in $L^p(q_3)$ that

$$\begin{aligned}
W_p(\mu_1,\mu_3) \leq& \left(\int d^p(x,z)\,\text{proj}_{1,3}(q_3)(dx,dz)\right)^{\frac{1}{p}}\\
=&\left(\int d^p(x,z) q_3(dx,dy,dz)\right)^{\frac{1}{p}}\\
\leq&\left(\int (d(x,y)+d(y,z))^p\, q_3(dx,dy,dz)\right)^{\frac{1}{p}}\\
\leq&\left(\int d(x,y)^p q_3(dx,dy,dz)\right)^{\frac{1}{p}} + \left(\int d(y,z)^p q_3(dx,dy,dz)\right)^{\frac{1}{p}}\\
=&\left(\int d(x,y)^p q_1(dx,dy)\right)^{\frac{1}{p}} + \left(\int d(y,z)^p q_2(dy,dz)\right)^{\frac{1}{p}}\\
=&W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3).
\end{aligned}$$

Finally, we need to show that W_p is real valued. From the triangle inequality we obtain

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

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

Remark 1.50. From the Kantorovich-Rubinstein formula one can directly show that W_1 defines a distance. On the other hand, observe that if $\mathsf{X} = \mathbb{R}$, $d = |\cdot|$, then

$$W_p(\mu,\nu) = \left(\int_{\mathbb{R}} |F_\mu^{-1}(x) - F_\nu^{-1}(x)|^p dx\right)^{\frac{1}{p}}.$$

In general there is no closed form solution for this distance.

Example 1.51. Let $\mu_i = \mathcal{N}(m_i, \sigma_i)$ for $i = 1, 2$. If $X \sim \mu_1$, then $Y := (X - m_1)(\sigma_2/\sigma_1) + m_2 \sim \mu_2$. Now $x \mapsto (\sigma_2/\sigma_1)(x - m_1) + m_2$ is increasing so it is optimal between μ_1 and μ_2 and we can calculate $W_2(\mu_1,\mu_2)^2$, assuming $m_1 = m_2 = 0$ for simplicity:

$$\int \left|x(1-\frac{\sigma_2}{\sigma_1})\right|^2 \mu_1(dx) = (1-\frac{\sigma_2}{\sigma_1})^2\sigma_1^2 = |\sigma_1 - \sigma_2|^2,$$

so in particular $W_2(\mu_1,\mu_2) = |\sigma_1 - \sigma_2|$ if μ_1 and μ_2 are centered.

Let us explore the connection between convergence in Wasserstein distance and weak convergence:

Lemma 1.52. *Suppose that the metric d is bounded. Then convergence in Wasserstein distance, and weak convergence, coincide.*

Proof. Evidently p plays here no role so take $p = 1$. Fix some $\bar{x} \in \mathsf{X}$ and observe that if $f : \mathsf{X} \to \mathbb{R}$ is 1-Lipschitz w.r.t. d and $f(\bar{x}) = 0$ then $\|f\|_\infty \leq \sup_{x,y} d(x,y) < \infty$. By the Kantorovich-Rubinstein formula

$$W_1(\mu,\nu) = \sup_{f\,1-Lipschitz} \int f d(\mu-\nu) = \sup_{f\in K} \int f d(\mu-\nu),$$

with K being the set of 1-Lipschitz functions with value zero at $\bar{x}$. By the Arzela-Ascoli theorem K is uniformly continuous, and in fact compact for the supremum norm. It is then easy to see that if $\mu_n \to \mu$ weakly, then $W_1(\mu_n,\mu) \to 0$. Conversely, if $W_1(\mu_n,\mu) \to 0$ then $\int f d\mu_n \to \int f d\mu$ for all f Lipschitz, and then by an approximation argument (exercise), for all f continuous and bounded. □

The question arises, as to what is the precise connection between weak convergence and convergence in Wasserstein distance for unbounded metrics. The answer is:

Theorem 1.53. *$W_p(\mu_n, \mu) \to 0$ iff both $\mu_n \to \mu$ weakly and*

$$\int d^p(x, x_0) d\mu_n(x) \to \int d^p(x, x_0) d\mu(x),$$

for some $x_0 \in \mathsf{X}$.

For the proof we refer to e.g. [San15, Section 5]. We also stress that this result can be strengthened by replacing the condition "$\int d^p(x, x_0) d\mu_n(x) \to \int d^p(x, x_0) d\mu(x)$ for some $x_0 \in \mathsf{X}$" by "$\int f(x) d\mu_n(x) \to \int f(x) d\mu(x)$ for all f continuous such that $|f(\cdot)| \leq C[1 + d^p(\cdot, x_0)]$". We also stress that if the metric d is complete, then also $(\mathcal{P}_p(\mathsf{X}), W_p)$ is a complete metric space.

To recap: the number $W_p(\mu, \nu)$ measures the distance between μ and ν. Now we want to take one step further and ask what is the optimal path from μ to ν? This leads us to a time dependent problem. We restrict ourselves to the case of $\mathsf{X} = \mathsf{Y} = \mathbb{R}^n$ and will write $(T_t x)_{0 \leq t \leq 1}$ for the trajectory of x between time 0 and time 1. We will write $C[T_t x]$ for the associated cost of this trajectory.

Definition 1.54. *The time-dependent Monge problem is given by*

$$\inf \left\{ \int C[T_t x] d\mu(x), T_0 = \mathrm{Id}, T_1(\mu) = \nu \right\}, \tag{1.6}$$

where the infimum runs over all continuous and piecewise C^1 curves $(T_t x)_{0 \leq t \leq 1}$ for μ almost all x.

Observe, that (1.6) is compatible with (MP) for the cost $c(x, y)$ (in the sense that the optimizer are equivalent, i.e. each $(T_t)_t$ gives rise to an optimal $T = T_1$ for (MP) and vice versa) if for all x and y

$$c(x, y) = \inf\{C[(z_t)_{0 \leq t \leq 1}], z_0 = x, z_1 = y\}. \tag{1.7}$$

Example 1.55. Let $p \geq 1$ and $C[(z_t)] = \int_0^1 |\dot{z}_t|^p dt$, with $\dot{z}_t$ denoting the time derivative of the curve z_t. Then, defining $c(x, y)$ as in (1.7) yields the cost $c(x, y) = |x - y|^p$. In fact, more generally as a consequence of Jensen's inequality it holds that if $c : \mathbb{R}^n \to \mathbb{R}$ is convex, then

$$\inf \left\{ \int_0^1 c(\dot{z}_t), z_0 = x, z_1 = y \right\} = c(y - x).$$

This means that straight lines with constant speed are optimal.

Example 1.56. In the setting of the previous example, it is often useful to consider the probabilistic notation: One minimizes over continuous-time stochastic processes $\{Z_t\}_{t \in [0,1]}$ with absolutely continuous paths such that $Z_0 \sim \mu$ and $Z_1 \sim \nu$. The cost function is then $\mathbb{E}\left[\int_0^1 |\dot{Z}_t|^p dt \right]$.

Combining these observations with Theorem 1.42 we arrive at the following result.

Theorem 1.57. *Let $c(x, y) = h(y - x)$ for some superlinear and strictly convex $h : \mathbb{R}^n \to \mathbb{R}$. Let $\mu \ll \mathsf{Leb}$ and let $C[(z_t)] = \int_0^1 h(\dot{z}_t) dt$. Then, there is a unique $\nabla\varphi$ for some c-concave function φ such that the solution to* (1.6) *is given by*

$$T_t(x) = x - t \nabla h^*(\nabla \varphi(x)), \quad 0 \leq t \leq 1.$$

Remark 1.58. Of course one can formulate also time dependent versions of (KP) in the obvious way. For instance if a coupling q^* is optimal for a static problem then q_t defined by $(\mathrm{Id}, \mathrm{Id} + t(\mathrm{proj}_2 - \mathrm{proj}_1))(q^*)$ is a sensible time-denpendent version which can be shown to be optimal for a suitable time-dependent problem if the cost functions are compatible as above. We will not dive into this now.

We now want to focus on a very special case which is maybe the most important case since it carries a lot of geometric information. We will consider now the case of $\mathsf{X} = \mathsf{Y} = \mathbb{R}^n$ (you can also think of X being a Riemannian manifold) and $c(x, y) = |x - y|^2$ (in case of the

Riemannian manifold $c(x,y) = d(x,y)^2$ where d denotes the geodesic distance). This leads to a remarkable interpolation between probability measures which is called *displacement interpolation*. It was invented by Robert McCann. We will discuss the special situation where we can grant existence of optimal transport maps. However, this is not necessary for the results which we present and with slightly more care one can formulate corresponding results using couplings, see also Remark 1.64.

Assume that μ and ν do not charge small sets. By Theorem 1.38, there exists convex function $\bar{\varphi}$ such that $\nabla\bar{\varphi}(\mu) = \nu$ (and $\nabla\bar{\varphi}^*(\nu) = \mu$ where $\bar{\varphi}^*$ is the Legendre dual of $\bar{\varphi}$). Define

$$\rho_t := [\mu,\nu]_t := ((1-t)\,\mathrm{Id} + t\nabla\bar{\varphi})(\mu), \quad 0 \le t \le 1.$$

(Observe that $\rho_t = T_t(\mu)$ for the map T_t induced by Theorem 1.57. To see this recall the functions φ and $\bar{\varphi}$ appearing in in the proof of Theorem 1.38.)

Lemma 1.59. *With notation and assumption as above there holds*

(i) $[\mu,\nu]_0 = \mu, [\mu,\nu]_1 = \nu$;
(ii) $W_2(\mu,\mu_t) = tW_2(\mu,\nu)$, *i.e. the path* $[0,1] \ni t \mapsto \mu_t$ *is a geodesic between* μ *and* ν *for the distance* W_2;
(iii) $[\mu,\nu]_t = [\nu,\mu]_{1-t}$;
(iv) $[[\mu,\nu]_t, [\mu,\nu]_{t'}]_s = [\mu,\nu]_{(1-s)t+st'}$
(v) *if* $\mu \ll \mathsf{Leb}$ *(or* $\nu \ll \mathsf{Leb}$ *by symmetry), then* $[\mu,\nu]_t \ll \mathsf{Leb}$ *for all* $t \in (0,1)$.

Proof. Item (*i*) is immediate. To show (*ii*) observe that $(1-t)\,\mathrm{Id} + t\nabla\bar{\varphi} = \nabla\left(\frac{(1-t)|\cdot|^2}{2} + t\bar{\varphi}\right)$ is the gradient of a convex function. Hence, it is the optimal transport map from μ to ρ_t by Theorem 1.38. Then,

$$W_2^2(\mu,\rho_t) = \int |x - ((1-t)x + t\nabla\bar{\varphi}(x)|^2\,\mu(dx) = t^2\int |x - \nabla\bar{\varphi}(x)|^2\,\mu(dx) = t^2 W_2^2(\mu,\nu).$$

To show (*iii*) observe that

$$\begin{aligned}[\mu,\nu]_t =& ((1-t)\,\mathrm{Id} + t\nabla\bar{\varphi})(\mu) = ((1-t)\,\mathrm{Id} + t\nabla\bar{\varphi})(\nabla\bar{\varphi}^*(\nu))\\ =& (((1-t)\,\mathrm{Id} + t\nabla\bar{\varphi}) \circ \nabla\bar{\varphi}^*)(\nu) = ((1-t)\nabla\bar{\varphi}^* + t\,\mathrm{Id})(\nu)\end{aligned}$$

Item (*iv*) is a direct computation. To show the last item (*v*) define

$$\bar{\varphi}_t(x) = \frac{(1-t)|x|^2}{2} + t\bar{\varphi}(x)$$

so that $\rho_t = \nabla\bar{\varphi}_t(\mu)$. Using $\mu \ll \mathsf{Leb}$ it can be shown that $\rho_t(A) = \mu(\partial\bar{\varphi}_t^*(A))$ (e.g. [Vil03, Lemma 4.6]). Then, it is sufficient to show that $\bar{\varphi}_t^*$ is Leb-a.e. differentiable with Lipschitz constant bounded from above since this implies that if A is a null set then $\partial\bar{\varphi}_t^*(A)$ is also a null set proving the claim. To this end, observe that

$$\langle \nabla\bar{\varphi}_t(x) - \nabla\bar{\varphi}_t(y), x - y\rangle \ge (1-t)|x-y|^2$$

so that by Cauchy-Schwarz

$$|\nabla\bar{\varphi}_t(x) - \nabla\bar{\varphi}_t(y)| \ge (1-t)|x-y|. \tag{1.8}$$

Hence, $\bar{\varphi}_t$ is uniformly convex and its Legendre dual $\bar{\varphi}_t^*$ is everywhere differentiable with $\nabla\bar{\varphi}_t^* = (\nabla\bar{\varphi}_t)^{-1}$ (e.g. [Roc97, Theorem 26.3]. In particular, $\nabla\bar{\varphi}_t^*$ is Lipschitz with constant bounded by $\frac{1}{1-t}$. This concludes the proof. □

Remark 1.60. Of course one can also consider displacement interpolation w.r.t. other cost functions than $c(x,y) = |x-y|^2$. For these interpolations one uses the maps from Theorem 1.57 and it is possible to derive corresponding properties. We will not pursue this direction but refer to e.g. [San15, Section 5]. However, for many applications the L^2 case is sufficient.

Remark 1.61. In dimension one, we have that $[\mu,\nu]_t$ is the measure whose quantile is equal to $(1-t)F_\mu^{-1} + tF_\nu^{-1}$.

Observe that the displacement interpolation has a considerably different nature than usual linear interpolation

$$(1-t)\mu + t\nu.$$

Since the measure are transported along curves in X this interpolation is able to capture geometric properties of the base space. An important example is the convexity of functionals on the space $\mathcal{P}_2(\mathsf{X})$ w.r.t. displacement interpolation. This convexity is called *displacement convexity*. For instance, this is useful to show uniqueness of minimizers to certain functionals (in fact, this was the motivation to introduce this concept in [McC94]).

Definition 1.62 (Displacement convexity)**.** *Denote the the set of absolutely continuous probability measures $\rho \ll \mathsf{Leb}$ by $\mathcal{P}_{ac}(\mathbb{R}^n)$.*

(i) *A subset $\mathcal{P} \subseteq \mathcal{P}_{ac}(\mathbb{R}^n)$ is called displacement convex if it is closed under displacement interpolation: for all $\mu, \nu \in \mathcal{P}, t \in [0,1]$ also $[\mu,\nu]_t \in \mathcal{P}$.*

(ii) *A functional $F : \mathcal{P} \to \mathbb{R}$ is called displacement convex if for all $\mu, \nu \in \mathcal{P}$ and $t \in [0,1]$ it holds that*

$$F([\mu,\nu]_t) \leq (1-t)F(\mu) + tF(\nu).$$

Example 1.63. Denote by $\mathcal{P}_m$ the set of all probability measures with mean m. Then, $\mathcal{P}_m$ is displacement convex.

Remark 1.64. Note that if q^* is the optimal coupling between μ and ν and $e_t : \mathsf{X}\times\mathsf{X} \to \mathsf{X}$ is given by $(x,y) \mapsto (1-t)x + ty$ then, $q_t^* := [\mu,\nu]_t := e_t(q^*)$ is a geodecic in the sense of Lemma 1.59 (ii). Also, the other items of Lemma 1.59 extend to this setup. However, note that since the optimal coupling need not be unique this interpolation (potentially) depends on the choice of coupling (e.g. if $\mu = \frac{1}{2}(\delta_{(0,0)} + \delta_{(1,1)})$, $\nu = \frac{1}{2}(\delta_{(1,0)} + \delta_{(0,1)})$ every coupling is optimal, each leading to a different interpolant).

Still one can also talk about displacement convexity in this setup if one asks functionals F to be convex for any possible interpolation.

There are three basic examples of displacement convex functionals:

a) internal energy: $\mathcal{U}(\rho) = \int_{\mathbb{R}^n} U(\rho(x))dx$, where $U : \mathbb{R}_+ \to \mathbb{R}\cup\{\infty\}$ is measurable, $\rho(dx) = \rho(x)dx \in \mathcal{P}_{ac}$ (if $\rho \notin \mathcal{P}_{ac}$ then $\mathcal{U}(\rho) = \infty$);

b) potential energy: $\mathcal{V}(\rho) = \int_{\mathbb{R}^n} V(x)d\rho(x)$, where $V : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is measurable;

c) interaction energy: $\mathcal{W}(\rho) = \frac{1}{2}\int_{\mathbb{R}^n\times\mathbb{R}^n} W(x-y)\rho(dx)\rho(dy)$, where $W : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is measurable.

Examples for functionals U that have been of major interest in the last years are $U(\rho) = \rho^\gamma$, for $\gamma \geq 1$ and $U(\rho) = \rho\log\rho$ so that $\mathcal{U}$ is the entropy. There is a general theorem granting displacement convexity of these functionals (or arbitrary sums of these functionals) for which we refer to [Vil03, Theorem 5.15]. We only prove the easiest case:

Theorem 1.65 (Criteria for displacement convexity)**.** *Let $\mathcal{P}$ (e.g. $\subseteq \mathcal{P}_{ac}$) be a displacement convex set on which $\mathcal{V}$ is well defined with values in $\mathbb{R}^n \cup \{\infty\}$. Then, $\mathcal{V}$ is displacement convex if V is convex. Conversely, if $\mathcal{V}$ is displacement convex on $\mathcal{P}(\mathbb{R}^d)$, then V is convex.*

Proof. Assume that V is convex. We argue for the case that $\mathcal{P} \subseteq \mathcal{P}_{ac}$, the general is similar with a bit more notation/analysis. Let $\rho_0, \rho_1 \in \mathcal{P}_{ac}$ and for $t \in [0,1]$ put $\rho_t = [\rho_0,\rho_1]_t = T_t(\rho_0)$ with $T_t(x) = (1-t)\,\mathrm{Id} + t\nabla\bar{\varphi}$ for some convex function $\bar{\varphi} =: T$ with $T(\rho_0) = \rho_1$. Then,

$$\begin{aligned}\mathcal{V}(\rho_t) &= \int V(x)\rho_t(dx) = \int V(T_t(x))\rho_0(dx)\\ &= \int V((1-t)x + tT(x))\rho_0(dx) \leq \int (1-t)V(x) + tV(T(x))\rho_0(dx)\\ &= \int (1-t)V(x)\rho_0(dx) + tV(x)T(\rho_0)(dx) = (1-t)\mathcal{V}(\rho_0) + t\mathcal{V}(\rho_1).\end{aligned}$$

For the converse direction it is sufficient to consider the displacement interpolation between two Dirac-measures δ_x and δ_y which directly yields the assertion. □

Observe, that for this theorem to hold it is not necessary that we consider a map T of the form $\nabla\bar{\varphi}$ for some convex $\bar{\varphi}$. The important thing is that we transport along geodesics on the base space (straight lines on $\mathbb{R}^n$).

Also note that for the linear interpolation $(1-t)\rho_0 + t\rho_1$ we only get that $\rho \mapsto \mathcal{V}(\rho)$ is linear for any function V. In particular, we cannot capture any convexity properties of V. For instance if V is strictly convex displacement convexity allows to argue the existence of a unique minimizer of the functional $\mathcal{V}$ (assume there are two, consider a displacement interpolation and use the *strict* convexity of V to derive a contradiction).

Before closing this section we want to take another differentiable point of view on the Kantorovich-Wasserstein distance W_2. To this end, let us consider as above a family of maps T_t such that the maps $(t,x) \mapsto T_t(x)$ and $T_t^{-1}(x)$ are Lipschitz on $[0,1]\times K$ for all compact $K \subseteq \mathbb{R}^n$, and $\rho_t := T_t(\mu)$ defines a curve of probability measures connecting μ and ρ_1. Since $T_t(x)$ is assumed to be locally Lipschitz we can associate a velocity field $v(t,x) = \frac{d}{dt}T_t(x)$ to these trajectories. We then have the following result:

Theorem 1.66. *The curve ρ_t is a solution to the linear transport equation*

$$\partial_t\rho_t + \nabla\cdot(v_t\rho_t) = 0, \rho_0 = \mu. \tag{1.9}$$

Remark 1.67. The divergence operator $\nabla\cdot$ is defined via duality $\int f d\nabla\cdot m := -\int \nabla f dm$ where f is a smooth test function and m a vector valued measure (loosely speaking sth that can be written as $m(dx) = (m_1(dx),\dots,m_n(dx))$). Similarly, the time derivative $\partial_t\rho_t$ should be understood in a weak form. One can also show that ρ_t is unique but we do not show this here.

Remark 1.68. Equation (1.9) is known in physics as the identity of conservation of mass and also called continuity equation.

Proof. Disclaimer: We will be short on some analytical details since this is not the core of this lecture.

Let $f \in C_c^\infty$ be a smooth test function. Then, $\int f d\rho_t = \int f\circ T_t d\mu$. By assumption T_t^{-1} is continuous so that $f\circ T_t$ is compactly supported uniformly in $0\le t\le 1$, Lipschitz so that for almost all x,t

$$\frac{d}{dt} f\circ T_t = (\nabla f\circ T_t)\cdot\frac{dT_t}{dt} = (\nabla f\circ T_t)\cdot(v_t\circ T_t).$$

Hence, for $h>0$ we obtain

$$\frac{1}{h}\left(\int f d\rho_{t+h} - \int f d\rho_t\right) = \int \frac{f\circ T_{t+h}(x) - f\circ T_t(x)}{h}\mu(dx)$$

By the Lipschitz property the integrand is uniformly bounded and we can pass to the limit to see that $t\mapsto \int f d\rho_t$ is differentiable for almost all t and we obtain

$$\frac{d}{dt}\int f d\rho_t = \int (\nabla f\circ T_t)\cdot(v_t\circ T_t)d\mu = \int \nabla f\cdot v_t d\rho_t.$$

which is precisely our claim. □

As a next step we want to picture μ as a distribution of gas particles which are supposed to move in space over the time interval $[0,1]$. At time 1 they are distributed according to ν. If they move along trajectories $T_t x$ with velocity v_t and distribution ρ_t at time t their kinetic energy at time t is given by

$$E_t = \int |v_t|^2\rho_t(x)dx$$

and (ρ_t, v_t) have to satisfy (1.9). To each pair $(\rho_t,v_t)_{t\in[0,1]}$ we can associate the action

$$A(\rho,v) = \int_0^1\int |v_t|^2\rho_t(x)dxdt.$$

This is the total effort needed to move particles around with the velocity field v. Then, the natural question is what is the minimal action needed to move μ to ν? This leads us the Benamou-Brenier minimization problem:

Definition 1.69 (Benamou-Brenier minimization problem)**.** *Minimize $A(\rho, v)$ among all $(\rho, v) = (\rho_t, v_t)_{0\le t\le 1} \in V(\mu, \nu)$ where $V(\mu, \nu)$ consists of all (ρ, v) such that $\rho \in C([0, 1], \mathcal{P}_{ac}(\mathbb{R}^n))$, $v_t \in L^2(d\rho_t(x)dt)$, (ρ, v) satisfies the continuity equation* (1.9) *in a weak sense, and $\rho_0 = \mu, \rho_1 = \nu$.*

We have the following remarkable result:

Theorem 1.70 (Benamou-Brenier)**.** *For $\mu, \nu \in \mathcal{P}_{ac}(\mathbb{R}^n) \cap \mathcal{P}_2(\mathbb{R}^n)$ there holds*

$$W_2^2(\mu, \nu) = \inf_{(\rho,v)\in V(\mu,\nu)} A(\rho, v).$$

We only give a sketch of the proof leaving aside all technical problems coming from potential non-smoothness of candidates (ρ, v).

Sketch of the proof. We first show $\inf_{(\rho,v)\in V(\mu,\nu)} A(\rho, v) \ge W_2^2(\mu, \nu) = \inf_{T,T(\mu)=\nu} \int |T(x) - x|^2\mu(dx)$ since $\mu \ll \mathsf{Leb}$. If (ρ, v) is sufficiently smooth we can define $T_t(x)$ as the solution of $\frac{dT_t(x)}{dt} = v_t(T_t(x))$ with initial datum $T_0(x) = x$. Since (ρ, v) solves the continuity equation one can show that $T_t(\mu) = \rho_t$ (essentially as a consequence of uniqueness in Theorem 1.66 which we did not discuss). In particular,

$$\int |v_t(x)|^2 d\rho_t(x) = \int |v_t(T_t(x))|^2 d\rho_0(x) = \int \left|\frac{d}{dt}T_t(x)\right|^2 d\rho_0(x)$$

Integrating this equality in t and using Jensen's inequality we obtain

$$\begin{aligned} A(\rho, v) &= \int_0^1 \int |v_t(x)|^2 \rho_t(dx) = \int \int_0^1 \left|\frac{d}{dt}T_t(x)\right|^2 dt\, d\rho_0(x) \\ &\ge \int \left|\int_0^1 \frac{d}{dt}T_t(x)dt\right|^2 d\rho_0(x) = \int |T_1(x) - T_0(x)|^2 \rho_0(dx) \\ &= \int |T_1(x) - x|^2 \rho_0(dx). \end{aligned}$$

Since, $T_1(\mu) = \nu$ we obtain $A(\rho, v) \ge W_2^2(\mu, \nu)$. Hence, $\inf_{(\rho,v)\in V(\mu,\nu)} A(\rho, v) \ge W_2^2(\mu, \nu)$.

To show the other inequality, we will explicitly construct a pair (ρ, v) attaining equality. To this end, let $T = \nabla\bar\varphi$ be the optimal transport map transporting μ to ν. As above set

$$T_t(x) = (1 - t)x + t\nabla\bar\varphi(x) =: \nabla\bar\varphi_t(x), \quad \rho_t := T_t(\mu).$$

We can then define the velocity field (defined ρ_t almost everywhere) by

$$v_t = \left(\frac{d}{dt}T_t\right) \circ T_t^{-1} = (T - \mathrm{Id}) \circ T_t^{-1}.$$

Similarly, as for Theorem 1.66 one can see that (ρ_t, v_t) solves the continuity equation (1.9). Moreover, for any nonnegative measurable function Φ we have

$$\int \Phi(v_t) d\rho_t = \int \Phi(v_t \circ T_t) d\rho_0 = \int \Phi(T - \mathrm{Id}) d\rho_0.$$

This applies in particular to $\Phi(v) = |v|^2$, so that for any t we have

$$\int |v_t(x)|^2 d\rho_t(x) = \int |T(x) - x|^2 d\rho_0(x) = W_2^2(\mu, \nu),$$

which implies the result. □

Remark 1.71. The Benamou-Brenier formulation of W_2 as least action functional is not only physically pleasing but also much better suited for generalizations of the differential point of view of W_2 to singular spaces like for instance graphs (see below for remarks on the usefulness of the differential point of view).

The attentive reader might have wondered as to what the continuous-time counterpart to the dual problem ought to be. We only sketch here the result in the particular case of quadratic costs:

Lemma 1.72. *We have*

$$\frac{1}{2}W^2(\mu,\nu) = \sup \int u(1,y)d\nu(y) - \int u(0,x)d\mu(x), \tag{1.10}$$

where the supremum runs over the (viscosity) solutions u to the quadratic Hamilton-Jacobi Equation with free boundary conditions:

$$\partial_t u + \frac{1}{2}|\nabla_x u|^2 = 0.$$

Proof. By duality the problem at hand is

$$\sup_{\psi \in C_b(\mathbb{R}^n)} \int \psi(y)d\nu + \int \psi^c(x)d\mu,$$

for $c(\cdot) = \|\cdot\|^2/2$. Given ψ bounded and differentiable define $u(t,x) = \inf_y\{(1-t)c\left(\frac{y-x}{1-t}\right) - \psi(y)\}$. Remark that $u(1,x) = -\psi(y)$ and $u(0,x) = \psi^c(x)$. Suppose now that there is a unique minimizer $y(t,x)$ attaining the infimum for $u(t,x)$. Then we leave it as an exercise to show

- $u(\cdot,\cdot)$ is differentiable,
- $\nabla_x u(t,x) = -\nabla c\left(\frac{y(t,x)-x}{1-t}\right)$,
- $\partial_t u(t,x) = -c\left(\frac{y(t,x)-x}{1-t}\right) + \frac{y(t,x)-x}{1-t} \cdot \nabla c\left(\frac{y(t,x)-x}{1-t}\right)$,

and to derive from this that $-\partial_t u + c^*(-\nabla_x u) = 0$, where c^* is the Fenchel (convex) conjugate of c. Finally redefining u to $-u$ finishes the proof. □

Remark 1.73. The proof is written in such a way that you can guess the dynamic dual problem when the cost function c is more general than the square one (though one still needs convexity and super-linearity). To make these arguments rigorous, even in the quadratic case, one must accept the fact that u in the proof need not be differentiable, and provide a suitable interpretation for the Hamilton-Jacobi PDE.

1.5. **Outlook.** In this section, we will give a very short outlook to possible directions one could take now, all of which we will not pursue in this lecture (e.g. see [San15, Vil03, Vil09, AG13, Gal18, PC19]).

- Detailed study of displacement convexity leads to proofs of functional/geometric inequalitiy like log-Sobolev inequality, Poincare-inequality, Brunn-Minkowski inequality...
- Benamou-Brenier formulation of W_2 resembles the formulation of a Riemannian distance leading to a formal interpretation of $(\mathcal{P}_2(\mathsf{X}), W_2)$ as an infinite dimensional Riemannian manifold. Hence, one can study gradient flows of various functionals on $\mathcal{P}_2$. This leads to gradient flow representation of PDEs such as heat flow, porous medium equation or Fokker-Planck equations.
- The study of displacement convexity properties of entropy functionals leads to synthetic notions of curvature and dimension and thereby to lower "Ricci"-curvature and upper "dimension" bounds of metric measure spaces. These spaces share various properties with classical Riemannian manifolds.
- Building numerical solver (interesting in itself) for the optimal transport problem leads to entropic variants of the problems which in turn are linked to the Schrödinger problem.
- The optimal transport problem is used in image analysis and machine learning for example as a "loss function".
- The optimal transport problem has various applications in the economic literature, for example in the context of mechanism design.
- Multimarginal versions of the problem are very natural objects. In a similar note, rather than displacement interpolation between two measures, one may be interested in finding barycenters of multiple measures.

1.6. **Exercises.**

I: Basic Material

Problem 1. (Absolute Continuity)

Ist folgendes Wahr?:
Für alle Wahrscheinlichkeitsmaße μ und ν auf X bzw. Y, gilt

$$\pi \in \Pi(\mu, \nu) \Rightarrow \pi \ll \mu \otimes \nu,$$

das heißt, der unabhängige Transportplan dominiert (im Sinne von absoluter Stetigkeit von Massen) jeden anderen Transportplan. Wenn nicht, muss es immer einen anderen dominierenden Plan $\pi^* \in \Pi(\mu, \nu)$ geben?

Problem 2. (Formulae)

Sei $X = Y = \mathbb{R}^n$ und μ, ν absolut stetig bzgl. dem Lebuesguemass auf $\mathbb{R}^n$, mit den Dichten f bzw. g. Finden Sie eine (heuristische) Bedingung/Formel für $T : X \to Y$, die nur von f und g abhängt, so dass $T(\mu) = \nu$ gilt (d.h. ν ist gleich dem Bildmass von μ unter T). Finden Sie Beispiele für die $T(\mu) = \nu$ gilt aber wo die gefundene Bedingung/Formel sinnlos ist.

Problem 3. (On the Monge Case)

Gegeben T^1, T^2 mit $T^1(\mu) = T^2(\mu) = \nu$, und $T_\lambda := \lambda T^1 + (1 - \lambda)T^2$ mit $\lambda \in [0, 1]$, muss immer $T_\lambda(\mu) = \nu$ gelten? Ist zu erwarten, dass die Bedingung $T(\mu) = \nu$ im Allgemeinen "kompakt in T" ist?

Problem 4. (Discrete Case)

Seien $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$ und μ,ν die Uniformverteilungen auf X,Y. Zu beweisen ist die folgende Behauptung: Für eine Kostenfunktion c auf $X \times Y$ haben das Kantorovich-Transportproblem und das Monge-Transportproblem (zwischen μ und ν) den selben Wert, und es gibt mindestens eine optimale Monge Transport Map. Beweisen Sie dazu die folgenden Argumente:

a) Das Kantorovich-Problem lässt sich als Optimierungsproblem über Doppelstochastische Matrizen (von Grösse $n \times n$) schreiben. Die Monge Maps entprechen genau den Permutationsmatrizen.
b) Beweisen Sie Problem 4 in [[Vil03]: Warm-up Exercises], nähmlich, dass die Menge der extremalen Punkte der Menge der Doppelstochastische Matrizen gleich der Menge der Permutationsmatrizen ist.
c) Choquets Theorem (wie in [[Vil03]: Warm-up Exercises, Problem 3]) sagt, dass ein lineares Optimierungsproblem auf einer kompaten Menge von mindestens einem extremalen Punkt der Menge gelöst wird. Nutzen Sie (ohne Beweis) diesen Satz um unsere Behauptung zu beweisen.

II: Duality

Problem 5. (Approximation Argument)

Sei (X, d) ein metrischer Raum und c eine nicht-negative unterhalbstetige Abbildung. Zu beweisen ist, die Existenz einer Folge $\{c_n\}_n$ die wachsend nach c punktweise konvergiert, wobei jede c_n nicht-negativ, beschränkt und gleichmäßig Stetig ist. (Hinweis: untersuchen Sie die Abbildung $x \mapsto \inf_{y \in X}[c(y) + nd(x, y)]$).

Problem 6. (... Continuation ...)

Mithilfe von Frage 5, beweisen Sie dass die Funktion $\pi \in \mathcal{P}(X) \mapsto \int c d\pi$ unterhalbstetig bezüglich schwache Konvergenz ist.

Problem 7. (Duality Gap)

Wir zeigen dass für meßbare Kostenfunktionen, die den Wert $+\infty$ aufnehmen durfen, die Kantorovich Dualität scheitern kann. Sei $X = Y = [0, 1]$, und $\mu = \nu = \lambda$ die Lebesguemaße auf $[0, 1]$. Als Kostenfunktion wählen wir

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

Dann gilt:

a) Die Funktion c ist meßbar und nicht unterhalbstetig.
b) Der Wert vom Kantorovich (primal) Problem zwischen μ, ν und mit Kosten c, gleicht 1.
c) Der Wert vom entsprechenden dualen Problem, gleicht 0.

Problem 8. (Duality for l.s.c. costs)

Let μ and ν be probability measures on X and Y. Moreover, Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ be a l.s.c. function that is bounded from below and such that there exists $a \in L^1(\mu)$ and $b \in L^1(\nu)$ with $c(x, y) \leq a(x) + b(y)$ for all x, y. Prove that

$$\inf_{\pi \in \mathsf{Cpl}(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{Y}} c d\pi = \sup_{\substack{\varphi \in L^1(\mu),\, \psi \in L^1(\nu) \\ \varphi(x)+\psi(y) \leq c(x,y)}} \int_{\mathsf{X}} \varphi d\mu + \int_{\mathsf{Y}} \psi d\nu,$$

leveraging on the known result for continuous costs.

Hint: You can use Baire's theorem on semi-continuous functions which states that any lower-semicontinuous function (on a Polish space) is the pointwise limit of an increasing sequence of continuous functions (cf. Problem 5).

III: The One-Dimensional Case

Problem 9. (Distance Cost)

Seien $A, B \subseteq \mathbb{R}$ disjunkte Intervalle, und seien $\mu, \nu \in \mathcal{P}(\mathbb{R})$ mit $\text{supp}(\mu) \subseteq A$ und $\text{supp}(\nu) \subseteq B$. Als Kostenfunktion betrachten wir $c(x, y) = |x - y|$. Zu zeigen ist, dass für das entsprechende Kantorovich Problem jedes $\pi \in \Pi(\mu, \nu)$ optimal ist.

Problem 10. (One-Dimensional Segments)

Sei $X = Y = [-1, 1] \times [0, 1] \subseteq \mathbb{R}^2$. Sei $\mu \in \mathcal{P}(X)$ gegeben durch das ein-dimensionale Lebesgue-Maß auf $\{(0, y) \in \mathbb{R}^2 : y \in [0, 1]\}$. Sei $\nu \in \mathcal{P}(Y)$ gleich $1/2$ dem ein-dimensionale Lebesgue-Maß auf $\{(-1, y) \in \mathbb{R}^2 : y \in [0, 1]\}$ plus $1/2$ dem ein-dimensionale Lebesgue-Maß auf $\{(1, y) \in \mathbb{R}^2 : y \in [0, 1]\}$. Wir betrachten das Transport Problem mit Kostenfunktion $c((x_1, x_2), (y_1, y_2)) = (x_1 - y_1)^2 + (x_2 - y_2)^2$; also der quadratische Fall. Zu beweisen ist:

a) Weder μ noch ν ist absolut stetig bezüglich des zweidimensionalen Lebesgue-Maßes.
b) Es existiert eine Monge Transport Map von μ nach ν. Genau genommen existiert eine Folge $\{T^n\}_{n\in\mathbb{N}}$ von solchen Maps, deren Kosten gegen 1 konvergieren.
c) Der Wert des Kantorovich Problems ist 1, und es gibt keine zulässige Monge Map, die ebenfalls diesen Wert erreicht.

Problem 11. (Monotone Rearrangement)

Wir untersuchen ein Transport Problem auf $\mathbb{R}$. Seien $\mu, \nu \in \mathcal{P}(\mathbb{R})$ mit Verteilungsfunktionen F_μ, F_ν. Als Kostenfunktion wählen wir $c(x, y) = h(y - x)$ mit $h : \mathbb{R} \to \mathbb{R}_+$ strikt konvex. Sei λ das Lebesgue-Maß auf $[0, 1]$. Zu zeigen ist:

a) Definiert man die Pseudoinverse $F_\mu^{-1}(x) := \inf\{t \in \mathbb{R} : F_\mu(t) \geq x\}$, dann ist F_μ^{-1} wachsend, rechts-stetig und es gilt $(F_\mu^{-1})_\# \lambda = \mu$.

b) Man definiere $\pi_{mon} := (F_\mu^{-1}, F_\nu^{-1})_\# \lambda$. Das heißt π_{mon} entspricht der Verteilung des Zufallsvektors $(F_\mu^{-1}(U), F_\nu^{-1}(U))$, wobei U eine uniform auf $[0,1]$-verteilte Zufallsvariable ist. Dann gilt $\pi_{mon} \in \Pi(\mu,\nu)$ und $\pi_{mon}((-\infty,a] \times (-\infty,b]) = \min\{F_\mu(a), F_\nu(b)\}$.

c) Sei $\pi \in \Pi(\mu,\nu)$, für das die folgende Eigenschaft gilt:

Wenn $(x_1,y_1),(x_2,y_2) \in supp(\pi)$ mit $x_1 < x_2$, dann gilt auch $y_1 \leq y_2$.

Dann muss $\pi = \pi_{mon}$ gelten. (Hinweis: Betrachen Sie die Mengen $(-\infty,a] \times (-\infty,b]$.)

d) Angenommen das Kantorovich Problem hat endlichen Wert, dann es hat eine einzige eindeutige Lösung, nämlich π_{mon}.

Bemerkung (nichts zu beweisen): Man nennt π_{mon} die "quantile transform", "Monotone rearrangement" oder auch "Fréchet coupling". Wenn h nur konvex ist, dann ist auch π_{mon} optimal, aber nicht mehr eindeutig (siehe Frage 1). Wenn μ keine Atome hat, dann ist π_{mon} von Monge-Art und explizit von der Tansport Map $T := F_\nu^{-1} \circ F_\mu$ induziert.

Problem 12. (Concave Cost)

a) Let $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a strictly concave function, set $c(x,y) := h(|x-y|)$ and let μ and ν be probability measures on $\mathbb{R}$. Suppose $\pi^* \in \mathsf{Cpl}(\mu,\nu)$ is an optimizer of

$$\inf_{\pi \in \mathsf{Cpl}(\mu,\nu)} \int_{\mathbb{R}^2} h(|x-y|) d\pi(x,y).$$

Show that there exists a set $\Gamma \subseteq \mathbb{R}^2$ with $\pi^*(\Gamma) = 1$ such that for all pairs (x_1,y_1), $(x_2,y_2) \in \Gamma$ the *"corresponding arches do not cross"*, i.e. the intersection

$$[\min\{x_1,y_1\}, \max\{x_1,y_1\}] \cap [\min\{x_2,y_2\}, \max\{x_2,y_2\}]$$

is either empty, a single point or equal to one of the two intervals.

b) We denote by δ_x the point mass in $x \in \mathbb{R}$. Let $\mu = \frac{1}{2}(\delta_{-9} + \delta_1)$, $\nu = \frac{1}{2}(\delta_{-1} + \delta_9)$. For each $p \in (0,1)$, identify all optimizer of

$$\inf_{\pi \in \mathsf{Cpl}(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^p d\pi(x,y)$$

and compute the minimal transport cost. What do you observe?

IV: Fundamental Theorem of Optimal Transport, c-transforms and c-cyclical monotonicity

Problem 13. (Example of monotonicity)

Let h be a strictly convex function, set $c(x,y) := h(y-x)$ for all $x,y \in \mathbb{R}$. Let $\Gamma \subseteq \mathbb{R}^2$. Show that Γ is c-cyclically monotone if and only if

$$\forall (x_1,y_1),(x_2,y_2) \in \Gamma : x_1 < x_2 \Rightarrow y_1 \leq y_2.$$

Hint: Let $h : \mathbb{R} \to \mathbb{R}$ be a strictly convex function and let $\varepsilon > 0$. There holds $h(a+\varepsilon) - h(a) < h(b+\varepsilon) - h(b)$ if and only if $a < b$.

Problem 14. (Properties of the c-transform)

The goal of this exercise is to prove Lemma 1.20 from the lecture. Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ and $\varphi : \mathsf{X} \to \mathbb{R} \cup \{-\infty\}$ be functions.

a) Show that $\varphi^{c\bar{c}} := (\varphi^c)^{\bar{c}} \geq \varphi$.

b) Prove that $\varphi^{c\bar{c}} = \varphi$ if and only if φ is $\bar{c}$-concave.

Hint: Use that for any function $\psi : \mathsf{Y} \to \mathbb{R} \cup \{-\infty\}$ there holds $\psi^{\bar{c}c} \geq \psi$ (the proof is very similar to a)).

c) Show that $\varphi^{c\bar{c}}$ is the smallest $\bar{c}$-concave function larger than φ, i.e. show that any $\bar{c}$-concave function $\varphi' : \mathsf{X} \to \mathbb{R} \cup \{-\infty\}$ with $\varphi' \geq \varphi$ satisfies $\varphi' \geq \varphi^{c\bar{c}}$.

Problem 15. (Examples of c-concave Functions)

Let $c : \mathsf{X}^2 \to \mathbb{R}$ and $\varphi : \mathsf{X} \to \mathbb{R}$ be functions.

a) Suppose $c = d$ is a metric on X. Show that the following are equivalent:
   (i) φ is c-concave.
   (ii) φ is 1-Lipschitz, i.e. $|\varphi(x) - \varphi(y)| \leq d(x, y)$ for all $x, y \in \mathsf{X}$.
   (iii) φ satisfies $\varphi^c = -\varphi$.

b) Suppose $\mathsf{X} = \mathbb{R}^n$ and $c(x, y) := -x \cdot y$ is the standard Euclidean inner product. Prove that φ is c-concave if and only if φ is u.s.c. and concave. Moreover, show that

$$\partial^c \varphi(x) = \{a \in \mathbb{R}^n : \varphi(y) \leq \varphi(x) + a \cdot (y - x) \text{ for all } y \in \mathbb{R}^n\}.$$

c) Suppose $\mathsf{X} = \mathbb{R}^n$ and $c(x, y) := \frac{1}{2}\|x - y\|^2$. Show that φ is c-concave if and only if the function $\bar{\varphi}(x) := \frac{\|x\|^2}{2} - \varphi(x)$ is l.s.c. and convex.

Problem 16. (An Application)

Lösen Sie Problem 11.(d) mithilfe vom Fundamentaler Satz optimaler Transport (Hinweis: man versucht die strikt konkavität von der Kostenfunktion zu widersprechen, in dem man die Negation der Eigenschaft von Problem 11.(c) annimmt).

Problem 17. (Conjugacy and Inverses)

Show by hand that if φ is convex and $\nabla\varphi$ exists and is invertible, then $(\nabla\varphi)^{-1} = \nabla\varphi^*$, with φ^* the convex conjugate. Now obtain this type of result, using the fundamental theorem of optimal transport under the assumption that μ, ν have a density.

Problem 18. (Stability of Optimizers)

Consider c be continuous and bounded. Let q_n be optimal between its marginals for the cost function c, and suppose $q_n \to q$. Then q is optimal between its marginals for the cost function c.

Problem 19. (Stability (Continued))

Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ be continuous and bounded from below. Moreover, let $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ be sequences of probability measures that converge weakly to the probability measures μ and ν and let $(c_n)_{n\in\mathbb{N}}$ be a sequence of continuous cost functions on $\mathsf{X} \times \mathsf{Y}$ that converge uniformly to c. For each $n \in \mathbb{N}$, we denote by π_n an optimizer of c_n w.r.t. μ_n and ν_n. Suppose that

$$\liminf_{n\to\infty} \int_{\mathsf{X}\times\mathsf{Y}} c_n d\pi_n < +\infty.$$

Prove that there exists subsequence of $(\pi_n)_{n\in\mathbb{N}}$ that converges weakly to an optimizer of

$$\inf_{\pi\in\mathsf{Cpl}(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{Y}} c(x, y) d\pi(x, y).$$

V: Uniqueness, Wasserstein Distances and Time-Dependent Version of OT

Problem 20. (Gaussians)

Find explicitly the optimal map between $\mu = \mathcal{N}(m_1, \Sigma_1)$ and $\nu = \mathcal{N}(m_2, \Sigma_2)$ for the quadratic cost function, as well as their W_2 distance. Can you extend this to the case when $\mu, \nu \in \{\text{Law}(AX + b)\}$ where X has a fixed distribution and A ranges over a class of matrices and b ranges over all vectors?

Problem 21. (Distance Functional)

We set $\mathcal{P}_2(\mathbb{R}^n) := \left\{\mu \in \mathcal{P}(\mathbb{R}^n) : \int_{\mathbb{R}^n} \|x\|^2 d\mu(x) < +\infty\right\}$. Fix $\mu \in \mathcal{P}_2(\mathbb{R}^n)$ with $\mu \ll \mathsf{Leb}$ and define the functional $F_\mu : \mathcal{P}_2(\mathbb{R}^n) \to \mathbb{R}$ by

$$F_\mu(\nu) := \inf_{\pi \in \mathsf{Cpl}(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 d\pi(x, y).$$

Show that $\mathcal{P}_2(\mathbb{R}^n)$ is a convex set and that F_μ is strictly convex, i.e. for all $\nu \neq \nu'$ in $\mathcal{P}_2(\mathbb{R}^n)$ and $\alpha \in (0, 1)$ there holds $F_\mu((1-\alpha)\nu + \alpha\nu') < (1-\alpha)F_\mu(\nu) + \alpha F_\mu(\nu')$.

Problem 22. (Relative compactness in Wasserstein space)

Show that if $\{\mu\}_n$ is tight and

$$\lim_{C\to\infty} \limsup_n \int_{d(x,x_0)\geq C} d(x, x_0)^p \mu_n(dx) = 0, \tag{1.11}$$

then $\{\mu\}_n$ is relatively compact in W_p. Find a sufficient condition for (1.11).

Problem 23. (Composition)

Show that the composition of optimal transport maps is in general not optimal. On the other hand, show that the composition of the optimal transport map from $[\mu, \nu]_r$ to $[\mu, \nu]_s$, with the optimal map from $[\mu, \nu]_s$ to $[\mu, \nu]_t$, is optimal from $[\mu, \nu]_r$ to $[\mu, \nu]_t$ $(r < s < t)$.

Problem 24. (Interaction Energy)

Show that $\rho \mapsto \int\int W(x - y)\rho(dx)\rho(dy)$ is displacement convex on $\mathcal{P}_2(\mathbb{R}^n) \cap \mathcal{P}_{ac}(\mathbb{R}^n)$ if $W : \mathbb{R}^n \to \mathbb{R}$ is convex and lower bounded.

Problem 25. (Interpolation of Gaussian Distributions)

Let $\mathcal{V} \subseteq \mathcal{P}(\mathbb{R})$ be the set of all Gaussian distributions on $\mathbb{R}$, i.e.

$$\mathcal{V} := \{\mathcal{N}(m, v) \,:\, m \in \mathbb{R}, v > 0\}.$$

a) Show that $\mathcal{V}$ is not a convex set.
b) Prove that $\mathcal{V}$ is displacement convex.

Problem 26. (Geometric Applications)

Let $n \geq 1$ and λ be the Lebesgue measure on $\mathbb{R}^n$. The internal energy $\mathcal{F}$ defined by

$$\mathcal{F}(\mu) := \begin{cases} -\int_{\mathbb{R}^n} (\rho(x))^{1-\frac{1}{n}} dx & \mu \ll \lambda, \rho = \frac{d\mu}{d\lambda} \\ +\infty & \text{else} \end{cases}$$

is displacement convex.

a) Deduce from the displacement convexity of $\mathcal{F}$ the Brunn-Minkowski inequality, i.e. show that for all compact sets $A, B \subseteq \mathbb{R}^n$ there holds

$$\lambda(A)^{\frac{1}{n}} + \lambda(B)^{\frac{1}{n}} \leq \lambda(A + B)^{\frac{1}{n}}$$

where $A + B := \{a + b : a \in A, b \in B\}$ is the Minkowski sum of two sets.

Hint: You can use that if μ and ν have bounded densities w.r.t. λ, then $[\mu, \nu]_t$ has a bounded density w.r.t. λ for all $t \in [0, 1]$.

b) Let $K \subseteq \mathbb{R}^n$ be a compact and convex set. We define the surface area of K as the differential rate of volume increase if we enlarge the set, i.e.

$$S(K) := \limsup_{\varepsilon\to 0} \frac{\lambda(\{x \in \mathbb{R}^n \,:\, \exists y \in K \text{ s.t. } |x - y| \leq \varepsilon\}) - \lambda(K)}{\varepsilon}$$

Use a) to prove the isopermetric inequality

$$S(K) \geq n\lambda(K)^{\frac{n-1}{n}} \lambda(B_1)^{\frac{1}{n}}$$

where B_1 denotes the ball with radius 1 around 0.

c) Show that the ball B_1 has the maximal volume among all compact convex bodies in $\mathbb{R}^n$ that have the same surface area as B_1. Similarly, show that B_1 has the smallest surface area among all compact convex bodies with the same volume.

2. The martingale optimal transport problem

2.1. **Motivation: A very quick primer in (robust) finance.** In this section we motivate the martingale optimal transport problem by the problem of worst case bounds in robust finance. We start with a very quick review on finance. For a more detailed exposition of mathematical finance (without robustness) we refer to [DS06] (the first two chapters give a very nice and quick introduction without the technical subtleties from the general theory).

Let us assume we are in the following setup:

- we are given a risky asset $S \equiv (S_0, \dots, S_T)$, seen as a canonical process on $\mathbb{R}^{T+1}$ (in general one could of course also consider continuous time processes).
- there are no trading cost, no interest rate,..
- a "derivative/option" is a financial security whose value/payoff is derived from the evolution of S, i.e. a function of S. We write $f(S) = f((S_t)_{0\le t\le T}) = f(S_0, \dots, S_T)$.

The fundamental problem in mathematical finance is to find a fair price for f.

There are two important principles: Firstly,

$$f \le g \quad \Rightarrow \quad \mathsf{price}(f) \le \mathsf{price}(g) \quad \text{if } a \ge 0 \Rightarrow \mathsf{price}(af) = a \cdot \mathsf{price}(f).$$

In short, price should be a linear operator on the space of all derivatives. In particular, if we find g with $f = g$ then $\mathsf{price}(f) = \mathsf{price}(g)$. Secondly, since there are no transaction cost, the price for trading in S should be zero, i.e. $\mathsf{price}((H \cdot S)_T) = 0$ where $H \cdot S$ denotes the gain or loss from trading in S using the strategy H; mathematically, $H \cdot S$ is a stochastic integral, e.g. in discrete time

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

with H_s being $\mathcal{F}_s$-measurable ($(\mathcal{F}_s)_s$ being the natural filtration of S, namely $\mathcal{F}_s$ is the information of $S_0, S_1, \dots, S_s$).

We say that a probability measure $\mathbb{P}$ – modelling the evolution of S – satisfies the *no arbitrage* condition, short NA, if

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

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 $\mathbb{Q}$ w.r.t. its natural filtration. This simply means that S is $\mathbb{Q}$-integrable and that $\mathbb{E}_{\mathbb{Q}}[S_{t+1}|\mathcal{F}_t] = S_t$. There is the important fundamental theorem of asset pricing (FTAP) connecting (NA) to the theory of martingales.

Theorem 2.1 ([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 2.1 such that S is a $\mathbb{Q}$-martingale is called a (equivalent) martingale measure. Note that this reflects the intuitive idea that price should be a linear operator.

The combination of the first and second principle on pricing derivatives leads to the following superhedging result, which implies that the "maximal fair price" for a derivative is given by the expectation against an equivalent martingale measure. Note that there might be several fair prices attached to a prescribed evolution $\mathbb{P}$ of the asset model S.

Theorem 2.2. *Let $\mathbb{P}$ satisfy* (NA) *and let $f \in L^1(\mathbb{P})$. Then*

$$\sup\{\mathbb{E}_{\mathbb{Q}}[f] : \mathbb{Q} \sim \mathbb{P}, \mathbb{Q} \text{ martingale measure}\} = \inf\{a : \exists H, a + (H \cdot S)_T \ge f, \ \mathbb{P}-a.s.\},$$

where the process H on the r.h.s. is adapted to $(\mathcal{F}_s)_s$.

Similarly, the "minimal fair price" is given by

$$\inf\{\mathbb{E}_{\mathbb{Q}}[f] : \mathbb{Q} \sim \mathbb{P}, \mathbb{Q} \text{ martingale measure}\} = \sup\{a : \exists H, a + (H \cdot S)_T \le f, \ \mathbb{P}-\text{a.s.}\}.$$

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

As a next step we want to incorporate market data in order to reduce the size of the potentially big set of martingale measures. Denote $C_{t,k} = (S_t - k)_+$ 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 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 $p_t \geq 0$.
(2) Take $k_1 < k_2$ and set $k := (1-\lambda)k_1 + \lambda k_2$ for some $\lambda \in (0,1)$. Then, it holds that
$$(1-\lambda)C_{t,k_1} + \lambda C_{t,k_2} \geq C_{t,k}.$$
Consequently, we have
$$(1-\lambda)p_t(k_1) + \lambda p_t(k_2) \geq p_t(k) = p_t((1-\lambda)k_1 + \lambda k_2).$$
Hence p_t is convex in k.
(3) $C_{t,0} = S_t$ (stock prices are non-negative) $\Rightarrow\ p_t(0) = S_0$.
(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
$$0 \leq p_t(k_1) - p_t(k_2) \leq k_2 - k_1,$$
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 2.3 (Breeden-Litzenberger)**.** *Assume that $k \mapsto p(k)$ satisfies the properties (1)-(5) as above. Then, there exists a unique probability μ on $\mathbb{R}_+$ s.t.*

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

Moreover,

$$p(0) = \int x\mu(dx)$$
$$\mu((k,\infty]) = -p'(k+),$$

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 2.3, if the prices at a given maturity t for call options/derivatives 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 $f(S_t)$ there holds

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

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

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

Alternatively, Lemma 2.3 implies that the knowledge of the prices for all call options uniquely defines the distribution of S_t, i.e. the 1-dimensional marginal distribution of the stock price process S. Therefore, the prices for such options can be assumed to be known and we can use these options for hedging. Inserting this into the superreplication result yields:

Theorem 2.4 (Superhedging with market data)**.** *Assume that $\mathbb{P}$ satisfies* (NA) *and that all call prices are given from market data for every maturity $0 \le t \le T$. Denote the measures given by Lemma 2.3 by $\mu_1, \dots, \mu_T$. Then,*

$$\begin{aligned}&\sup\{\mathbb{E}_{\mathbb{Q}}[f] : \mathbb{Q} \sim \mathbb{P}, \mathbb{Q} \text{ martingale measure}, S_i \sim_{\mathbb{Q}} \mu_i\}\\ &= \inf\{a + \sum_{i=1}^{T} \int \varphi_i d\mu_i : \exists H, a + \sum_{i=1}^{T} \varphi_i(S_i) + (H \cdot S)_T \ge f(S)\ \mathbb{P}-a.s.\}.\end{aligned}$$

This is a nice result but there is an apparent problem: Different models $\mathbb{P}$ lead to potentially different martingale measures which will usually lead to different prices/price bounds. This is the "model risk" associated to pricing f with a specific model $\mathbb{P}$.

Summing up: While market data determines/restricts the distribution of the stock price process S at fixed time instances, $t = 0, 1, \dots, T$, we do not know how the stock price moves from one time instance to the next.

In *robust finance*, one aims at estimates (prices, hedging strategies) that are independent of the choice of the particular model (i.e. avoid fixing $\mathbb{P}$) only assuming S to be a martingale. This leads us to the following variant of the pricing problem

$$\sup / \inf\{\mathbb{E}_{\mathbb{Q}}[f] : \mathbb{Q} \text{ martingale measure}, S_i \sim_{\mathbb{Q}} \mu_i\}. \tag{2.1}$$

In analogy to (KP) we call this the martingale optimal transport problem. 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.

2.2. **Existence, duality, and geometry of optimizers: discrete time.** We want to analyze (2.1) and focus for (notational) simplicity on the case of one period, namely $T = 1$ so $t \in \{0, 1\}$. 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, write

$$\mathsf{X} = \mathsf{Y} = \mathbb{R},\ \mu := \mu_0 \in \mathcal{P}_1(\mathbb{R}),\ \nu := \mu_1 \in \mathcal{P}_1(\mathbb{R}),\ c(\cdot) := f(\cdot),$$

so that the object of study becomes:

$$P_c^{mg} := P_c^{mg}(\mu, \nu) := \inf_{Q \in \mathsf{MT}(\mu,\nu)} \int c\, dQ, \tag{MOT}$$

where $\mathsf{MT}(\mu, \nu) = \{Q \in \mathsf{Cpl}(\mu, \nu) : \mathbb{E}_Q[S_1|S_0] = S_0\}$.

Remark 2.5. Above we have used the probabilistic notation coming from math finance. The reader should keep in mind that $S = (x, y)$, $S_0 = x$ and $S_1 = y$, in the notation of the previous chapter. Similarly the condition $\mathbb{E}_Q[S_1|S_0] = S_0$ is equivalent to $\int y Q^x(dy) = x$ where Q^x is the conditional distribution of y given x.

Remark 2.6. In the literature on MOT one often considers the maximization problem instead of the minimization problem due to the relation to the superhedging result. Mathematically this is of course equivalent and we chose to work with the minimization to make it consistent with the rest of this lecture.

Definition 2.7. *Any solution to* (MOT) *is called optimal martingale coupling or optimal martingale transport.*

Remark 2.8. Due to the martingale constraint, there cannot be a Monge-type martingale coupling ($T(\mu) = \nu$) unless $\mu = \nu$. The "best" one can hope for is that the optimal martingale coupling is concentrated on the graph of two functions.

Remark 2.9. By a famous result of Strassen [Str65], there holds $\mathsf{MT}(\mu,\nu) \neq \emptyset$ iff μ is in convex order to ν, short $\mu \preceq_c \nu$, which by definition holds iff

$$\int f d\mu \leq \int f d\nu$$

for all convex functions $f : \mathbb{R} \to \mathbb{R}$ with linear growth.

Lemma 2.10. *Assume that $\mu \preceq_c \nu$ then $\mathsf{MT}(\mu,\nu)$ is compact w.r.t. the topology of weak convergence (in fact, it is compact w.r.t. the* 1*-Wasserstein topology).*

Proof. Since $\mathsf{MT}(\mu,\nu) \subseteq \mathsf{Cpl}(\mu,\nu)$ it is sufficient to show that $\mathsf{MT}(\mu,\nu)$ is closed by Corollary 1.13. Remark that $Q \in \mathsf{Cpl}(\mu,\nu)$ is a martingale iff $\int g(x)(y-x)\, dQ(x,y) = 0$ for all $g \in C_b(\mathbb{R})$. Let $Q_n \to Q$, each $Q_n \in \mathsf{MT}(\mu,\nu)$. Then $Q_n \to Q$ in the 1-Wasserstein topology, as the marginals are fixed and each of them has a finite first moment. For this, apply Theorem 1.53 with $d((x,y),(\bar{x},\bar{y})) = |x-\bar{x}| + |y-\bar{y}|$ and $p = 1$. Since $g(x)(y-x)$ has at most linear growth we conclude by the discussion after Theorem 1.53 that Q is a martingale.

The statement that $\mathsf{MT}(\mu,\nu)$ is in fact compact w.r.t. the 1-Wasserstein topology is similar proven. It suffices to observe that $\mathsf{Cpl}(\mu,\nu)$ is relatively compact w.r.t. the 1-Wasserstein topology. This can be shown as above, or by invoking Problem 22. □

Corollary 2.11. *Assume that c is l.s.c. and bounded from below. Then there exists an optimal martingale coupling.*

Proof. As for Theorem 1.14. □

Remark 2.12. If c is not l.s.c. in general there does not exist an optimal coupling. Lower boundedness can be suitably relaxed.

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 2.13. *Let c be l.s.c. and bounded from below. Assume that $P_c^{mg} < \infty$. Then, $P_c^{mg} = D_c^{mg} := D_c^{mg}(\mu,\nu)$ where*

$$D_c^{mg} = \sup\left\{ \int \varphi d\mu + \int \psi d\nu : \ \exists\, h \text{ s.t. } \varphi(x) + \psi(y) + h(x)(y-x) \leq c(x,y)\ \forall\ x,y \in \mathbb{R} \right\}.$$

Remark 2.14. The dual problem can be interpreted as a robust subhedging problem of the option with payoff c. In this sense, Theorem 2.13 is a robust version (i.e. independent of the reference model $\mathbb{P}$) of Theorem 2.4.

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

Theorem 2.13 will be derived from the following min-max theorem:

Theorem 2.15 (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 of Theorem 2.13. We first assume that c is continuous and bounded. We apply Theorem 2.15 with $K = \mathsf{Cpl}(\mu,\nu)$, $L = C_b(\mathbb{R})$ and $F(Q,h) = \int c(x,y) - h(x)(y-x)dQ(x,y)$ for $h \in C_b(\mathbb{R})$. As justified in the proof of Lemma 2.10 the set K is 1-Wasserstein compact, which is what we need since $F(\cdot,h)$ is 1-Wasserstein continuous. Now:

$$\begin{aligned}
P_c^{mg} &= \inf_{Q\in\mathsf{MT}(\mu,\nu)} \int c dQ \\
&= \inf_{Q\in\mathsf{Cpl}(\mu,\nu)} \sup_{h\in C_b(\mathbb{R})} \int c(x,y) - h(x)(y-x)dQ(x,y) \\
&\overset{Thm 2.15}{=} \sup_{h\in C_b(\mathbb{R})} \inf_{Q\in\mathsf{Cpl}(\mu,\nu)} \int c(x,y) - h(x)(y-x)dQ(x,y) \\
&\overset{Thm 1.32}{=} \sup_{h\in C_b(\mathbb{R})} \sup_{\varphi,\psi} \left\{ \int \varphi d\mu + \int \psi d\nu \mid \forall x,y : \varphi(x)+\psi(y) \le c(x,y) - h(x)(y-x) \right\} \\
&= D_c^{mg}.
\end{aligned}$$

In fact, observe that for any φ, ψ and h with $\varphi(x)+\psi(y)+h(x)(y-x) \le c(x,y)$ and $Q \in \mathsf{MT}(\mu,\nu)$, we obtain

$$\int \varphi d\mu + \psi d\nu = \int \varphi(x)+\psi(y)+h(x)(y-x)dQ(x,y) \le \int c(x,y)dQ(x,y)$$

and thus $P_c^{mg} \ge D_c^{mg}$ even if c is not continuous and bounded. Let us now suppose that c is l.s.c. and bounded from below. Then there exists a sequence $(c_n)_{n\in\mathbb{N}}$ of continuous and bounded functions such that $c_n \uparrow c$. As discussed, we have $P_c^{mg} \ge D_c^{mg}$. Since $c_n \le c$ we naturally have $D_{c_n}^{mg} \le D_c^{mg}$. For each n pick $q_n \in \mathsf{MT}(\mu,\nu)$ such that

$$P_{c_n}^{mg} \ge \int c_n dq_n - \frac{1}{n}.$$

Since $\mathsf{MT}(\mu,\nu)$ is compact, there exists $q \in \mathsf{MT}(\mu,\nu)$ and a subsequence still denoted by $(q_n)_{n\in\mathbb{N}}$ that converges to q. Finally,

$$\begin{aligned}
P_c^{mg} &\le \int c dq \overset{\substack{mon.\\ conv.}}{=} \lim_n \int c_n dq = \lim_n \lim_k \int c_n dq_k \overset{c_n\uparrow}{\le} \limsup \int c_k dq_k \\
&\le \limsup_k \left(P_{c_k}^{mg} + \frac{1}{k} \right) = \limsup_k D_{c_k}^{mg} \le D_c^{mg},
\end{aligned}$$

showing that $P_c^{mg} = D_c^{mg}$. □

As the following example shows, opposed to Theorem 1.32 one cannot go beyond l.s.c. cost functions c in Theorem 2.13:

Example 2.16. Let $\mu = \nu = \mathsf{Leb}_{|[0,1]}$. Then $\mathsf{MT}(\mu,\nu) = \{\hat{Q}\}$ 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 $P_c^{mg} = 1$ and we claim that $D_c^{mg} = 0$. 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 $x \ne y$. 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 $x_n \ne x$. 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 $T^\varepsilon(x) \ne x$, $T^\varepsilon(\mathsf{Leb}_{|[0,1]}) = \mathsf{Leb}_{|[0,1]}$, and $T^\varepsilon(x) - x$ is equal to ε on $[0, 1-\varepsilon]$ and equal to

$\varepsilon - 1$ on $(1-\varepsilon, 1]$. Hence $\varphi(x)+\psi(T^\varepsilon(x)) \leq -h(x)[T^\varepsilon(x)-x]$ and integrating w.r.t. Lebesgue we get

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

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

The reason that duality fails in Example 2.16 it that the dual problem, as defined in Theorem 2.13, is too restrictive. Financially speaking: In the dual problem we hedge against scenarios not present in the primal problem! To see this, we need the following.

Definition 2.17. *For a finite measure μ on $\mathbb{R}$ with $\int |x| d\mu < \infty$ we define its potential function $u_\mu : \mathbb{R} \to \mathbb{R}$ by*

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

Remark 2.18. The function $y \mapsto |x - y|$ is convex. If $\mu \preceq_c \nu$, then $u_\mu(x) \leq u_\nu(x)$ for all $x \in \mathbb{R}$. In fact, the reverse is true as well (c.f. Lemma 2.3). It is possible to read off various properties of μ from the behaviour of u_μ, i.e. total mass, mean, atoms, etc. See Exercise 32.

Pick $\mu \preceq_c \nu$ and assume that there exists a $z \in \mathbb{R}$ s.t. $u_\mu(z) = u_\nu(z)$. Then any $Q \in \mathsf{MT}(\mu, \nu)$ satisfies

$$\begin{aligned} u_\nu(z) \quad &= \quad \mathbb{E}_Q\left[\,|S_1 - z|\,\right] = \mathbb{E}_Q\left[\mathbb{E}_Q\left[\,|S_1 - z|\;|S_0\right]\right] \\ &\overset{\text{Jensen}}{\geq} \quad \mathbb{E}_Q\left[\,|\,\mathbb{E}_Q\left[S_1 - z\,|S_0\right]\,|\,\right] = \mathbb{E}_Q\left[\,|S_0 - z|\,\right] = u_\mu(z) = u_\nu(z). \end{aligned}$$

In particular, we get equality in Jensen's inequality, and hence any $Q \in \mathsf{MT}(\mu, \nu)$ satisfies

$$\begin{aligned} &Q(S_1 \geq z \mid S_0 > z) = 1, \\ &Q(S_1 \leq z \mid S_0 < z) = 1, \\ &Q(S_1 = z \mid S_0 = z) = 1 \qquad (\text{if } \mu(\{z\}) > 0\,). \end{aligned}$$

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.

Remark 2.19. Note that a variant of this argument implies that it is sufficient to test convex order against a single convex function which is nowhere flat, e.g. $\sqrt{1 + x^2}$.

We set $\mathrm{proj}_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, (x, y) \mapsto x$ and $\mathrm{proj}_2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, (x, y) \mapsto y$.

Lemma 2.20. *Let I be an open interval such that $u_\mu = u_\nu$ on ∂I. Set $\mu_I = \mu_{|I}$. Pick $Q \in \mathsf{MT}(\mu, \nu)$ and set $\nu_I = \mathrm{proj}_2(Q_{|I\times\mathbb{R}})$. Then ν_I is concentrated on $\bar{I}$ and does not depend on Q. Moreover, $u_{\nu_I} - u_{\mu_I} = 0$ on $\mathbb{R} \setminus I$ and $u_{\nu_I} - u_{\mu_I} = u_\nu - u_\mu$ on I.*

Proof. Pick $Q \in \mathsf{MT}(\mu, \nu)$ and apply the previous considerations to both $z \in \partial I$. Then

$$Q((I \times \bar{I}) \cup (\mathbb{R} \setminus I)^2) = 1,$$

i.e. no mass of μ is transported from $\mathbb{R} \setminus I$ into I and the mass of μ in I is transported to $\bar{I}$. Hence $\mu_I \preceq_c \nu_I$ and the rest follows from straightforward calculations with potential functions, which are left as an exercise. □

Definition 2.21. *Let μ and ν be finite measures on $\mathbb{R}$ with $\mu \preceq_c \nu$. The pair (μ, ν) is called irreducible if $I = \{u_\mu < u_\nu\}$ is connected and $\mu(I) = \mu(\mathbb{R})$. In this case, we denote by J the set $I \cup \{x \in \partial I : \nu(\{x\} > 0)\}$.*

Theorem 2.22. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with $\mu \preceq_c \nu$. Let $(I_k)_{1\leq k\leq N}$ be the (disjoint) open components of $\{u_\mu < u_\nu\}$ with $N \in \{0, 1, ..., \infty\}$. Set $I_0 = \mathbb{R} \setminus \bigcup_{k\geq 1} I_k$ and $\mu_k = \mu_{|I_k}$ for $k \geq 0$ s.t. $\mu = \sum_{k\geq 0} \mu_k$. Then there exists a unique decomposition $\nu = \sum_{k\geq 0} \nu_k$ s.t. $\mu_0 = \nu_0$ and $\mu_k \preceq_c \nu_k$ for all $k \geq 1$. Moreover, $I_k = \{u_{\mu_k} < u_{\nu_k}\}$ and each $Q \in \mathsf{MT}(\mu, \nu)$ can be uniquely decomposed as $Q = \sum_{k\geq 0} Q_k$ with $Q_k \in \mathsf{MT}(\mu_k, \nu_k)$ for all $k \geq 0$.*

Proof. This is a rather direct consequence of Lemma 2.20. □

Corollary 2.23. *The primal problem for MOT splits into (at most countably many) independent MOT problems on each irreducible component. More precisely, let $\mu = \sum_{k\geq 0}\mu_k$, $\nu = \sum_{k\geq 0}\nu_k$ as in Theorem 2.22 leading for all $Q \in \mathsf{MT}(\mu,\nu)$ to a decomposition $Q = \sum_{k\geq 0} Q_k$. Furthermore, let c be a cost function s.t. $P_c^{mg}(\mu,\nu) < \infty$. Then Q is optimal if and only if Q_k is optimal for the martingale transport between μ_k and ν_k for each k .*

Remark 2.24. Corollary 2.23 implies that for the primal problem only the set

$$\{(x,x) : x \in \mathbb{R}\} \cup \bigcup_{k\geq 1}(I_k \times J_k)$$

is relevant (notation from Definition 2.21). One can show that $B \subseteq \mathbb{R}^2$ is $\mathsf{MT}(\mu,\nu)$-polar, i.e. $Q(B) = 0$ for all $Q \in \mathsf{MT}(\mu,\nu)$, if and only if

$$B \subseteq (N_\mu \times \mathbb{R}) \cup (\mathbb{R} \times N_\nu) \cup \left(\{(x,x)\} \cup \bigcup_{k\geq 1}(I_k \times J_k)\right)^c \tag{2.2}$$

where N_μ is a μ-null set and N_ν is a ν-null set. Consequently, the constraint

$$\varphi(x) + \psi(y) + h(x)(y-x) \leq c(x,y) \qquad \text{for all } x,y \in \mathbb{R} \tag{2.3}$$

is too strong in the sense that it considers regions which are not seen by the primal problem. Instead one should replace 'for all $x,y \in \mathbb{R}$' by '$\mathsf{MT}(\mu,\nu)$-quasi surely', i.e. (2.3) should hold outside of $\mathsf{MT}(\mu,\nu)$ -polar sets. Put differently, we should consider the dual problem on each irreducible component separately and patch together.

Theorem 2.25. *Also the dual problem splits along irreducible components.*

Proof. Rather technical. Therefore we omit it. □

Theorem 2.26. *Let $c : \mathbb{R}^2 \to [0,\infty)$ be Borel and $\mu \preceq_c \nu$. Then*

$$\begin{aligned} P_c^{mg} &= \inf_{Q\in\mathsf{MT}(\mu,\nu)} \mathbb{E}_Q[c] \\ &= \sup\left\{\int^* \varphi d\mu + \int^* \psi d\nu \mid \exists h : \varphi(x) + \psi(y) + h(x)(y-x) \leq c(x,y)\ \mathsf{MT}(\mu,\nu)\text{-q.s.}\right\} \\ &= D_c^{mg}. \end{aligned}$$

Moreover, if $D_c^{mg} < \infty$, there exists a dual optimizer (φ,ψ,h).

Remark 2.27. $\int^*$ indicates that there is a technical subtlety defining these integrals which we omit.

Proof (Brutalist sketch). Prove the statement on each irreducible component for c l.s.c., and use Choquet's capacity Theorem to extend it to c Borel. Use Theorem 2.25 to patch the different components together. Arzela-Ascoli yields the existence of a dual optimizer. □

Theorem 2.28. *Let $c : \mathbb{R}^2 \to [0,\infty)$ be Borel, $\mu \preceq_c \nu$ probability measures and suppose that $P_c^{mg} < \infty$. Then there exists a Borel set $\Gamma \subseteq \mathbb{R}^2$ s.t.*

(i) $Q \in \mathsf{MT}(\mu,\nu)$ is concentrated on Γ if and only if Q is optimal w.r.t. the cost function c.

(ii) Let $\bar\mu \preceq_c \bar\nu$ be probability measures on $\mathbb{R}$. If $\bar Q \in \mathsf{MT}(\bar\mu,\bar\nu)$ is concentrated on Γ, then $\bar Q$ is optimal for the MOT problem between $\bar\mu$ and $\bar\nu$ w.r.t. the cost function c.

If (φ,ψ,h) is an optimizer of the dual problem (w.r.t. μ,ν) then we can take

$$\Gamma = \left\{(x,y) \in \mathbb{R}^2 : \varphi(x) + \psi(y) + h(x)(y-x) = c(x,y)\right\} \cap \left(\{(x,x) : x \in \mathbb{R}\} \cup \bigcup_{k\geq 1} I_k \times J_k\right).$$

Proof. (*i*): Let (φ, ψ, h) be an optimizer of the dual problem and pick Γ as above. For all $\tilde{Q} \in \mathsf{MT}(\mu, \nu)$ we get

$$\mathbb{E}_{\tilde{Q}}[c] \geq \mathbb{E}_{\tilde{Q}}\left[\varphi(S_0) + \psi(S_1) + h(S_0)(S_1 - S_0)\right] = D_c^{mg}$$

with equality if and only if $\tilde{Q}(\Gamma) = 1$.

(*ii*): We show that the Γ-defining triplet (φ, ψ, h) is also a dual optimizer w.r.t. $\bar{\mu}$ and $\bar{\nu}$ from which the result follows as before. By Remark 2.24 it suffices to show that $(\bar{\mu}, \bar{\nu})$ has "more" irreducible components than (μ, ν), i.e.

$$u_\mu(z) = u_\nu(z) \Rightarrow u_{\bar{\mu}}(z) = u_{\bar{\nu}}(z)$$

for all $z \in \mathbb{R}$. To this end, fix such a $z \in \mathbb{R}$. Since $\bar{Q} \in \mathsf{MT}(\bar{\mu}, \bar{\nu})$ is concentrated on $\{(x, x) : x \in \mathbb{R}\} \cup \bigcup_{k \geq 1} I_k \times J_k$, it holds $\bar{Q}(S_1 \geq z \mid S_0 > z) = 1$ and hence

$$\mathbb{E}_{\bar{Q}}\left[|S_1 - z| I_{\{S_0 > z\}}\right] = \mathbb{E}_{\bar{Q}}\left[|S_0 - z| I_{\{S_0 > z\}}\right].$$

The analogue statement is true for $I_{\{S_0 < z\}}$ and thus $u_{\bar{\mu}}(z) = u_{\bar{\nu}}(z)$. □

Remark 2.29. For Γ as in the previous Theorem set $\Gamma_x = \{y \in \mathbb{R} : (x, y) \in \Gamma\}$. W.l.o.g. we will assume that the existence of $y \in \Gamma_x$ with $y < x$ yields the existence of $y' \in \Gamma_x$ with $y' > x$.

Definition 2.30. *Let α be a finite measure on $\mathbb{R} \times \mathbb{R}$ with marginals α_0 and α_1 s.t. $\int |x| d\alpha_1 < \infty$. A measure α' on $\mathbb{R} \times \mathbb{R}$ is called a competitor of α if it has marginals α_0 and α_1, and for α_0-a.e. x it holds*

$$\int y \alpha_x(dy) = \int y \alpha'_x(dy)$$

where $(\alpha_x)_x$ and $(\alpha'_x)_x$ are disintegrations of α and α' w.r.t. α_0.

Corollary 2.31. *In the setup of Theorem 2.28 let α be a finite measure concentrated on Γ with $|\mathrm{supp}(\alpha)| < \infty$. For any competitor α' of α it holds*

$$\int c d\alpha \leq \int c d\alpha'.$$

Proof. First assume that α is a martingale coupling, i.e. $\int y d\alpha_x(y) = x$. Then we have

$$\begin{aligned}
\int c d\alpha &= \int \varphi(x) + \psi(y) + h(x)(y - x) d\alpha(x, y) \\
&= \int \varphi(x) d\alpha_0(x) + \int \psi d\alpha_1(y) + \int h(x) \left(\int y d\alpha_x(y) - x \right) d\alpha_0(x) \\
&= \int \varphi(x) + \psi(y) + h(x)(y - x) d\alpha'(x, y) \\
&\leq \int c d\alpha'
\end{aligned}$$

because the proof of Theorem 2.28 yields that α' is concentrated on $\{(x, x) : x \in \mathbb{R}\} \cup \bigcup_{k \geq 1} I_k \times J_k$. For the general case, we can pick a family $(p_x(dy))_x$ s.t.

$$\int y(\alpha_x(dy) + p_x(dy)) = \int y(\alpha'_x(dy) + p_x(dy)) = x$$

and then the result follows by considering

$$\begin{aligned}
\bar{\alpha}(dx, dy) &= (\alpha_x(dy) + p_x(dy))\alpha_0(dx) \qquad \text{and} \\
\bar{\alpha}'(dx, dy) &= (\alpha'_x(dy) + p_x(dy))\alpha_0(dx).
\end{aligned}$$

□

Remark 2.32. A set Γ that satisfies the assertion of the last corollary is called finitely monotone. Note that a finitely monotone set does not need to support any martingale (e.g. $\Gamma = \{0, 1\}$).

Example 2.33. Assume that $c_{xyy} < 0$, e.g. $c(x,y) = (y-x)^3$ or $c(x,y) = g(x)f(y)$ with g strictly decreasing and f strictly convex. Such cost functions are called generalized Spence-Mirrlees cost functions. Pick Γ from Theorem 2.28. Then Γ contains no pairs (x',y'), (x,y^+) and (x,y^-) s.t. $x < x'$ and $y^- < y' < y^+$.

To see this, let $\lambda \in (0,1)$ such that $y' = \lambda y^- + (1-\lambda)y^+$ and set

$$\begin{aligned} \alpha &= \delta_{(x',y')} + \lambda\delta_{(x,y^-)} + (1-\lambda)\delta_{(x,y^+)} \\ \alpha' &= \delta_{(x,y')} + \lambda\delta_{(x',y^-)} + (1-\lambda)\delta_{(x',y^+)} \end{aligned}$$

By contradiction, suppose that α is supported on Γ, and remark that α' is a competitor of α. However, since

$$y \mapsto c(x,y) - c(x',y) = \int_x^{x'} -c_x(z,y)dz$$

is by assumption a strictly convex function, we get

$$\begin{aligned} &\int cd\alpha - \int cd\alpha' \\ =\ & c(x',y') - c(x,y') + \left[\lambda\left(c(x,y^-) - c(x',y^-)\right) + (1-\lambda)\left(c(x,y^+) - c(x',y^+)\right)\right] \\ >\ & c(x',y') - c(x,y') + c(x,y') - c(x',y') = 0 \end{aligned}$$

which is a contradiction to the choice of Γ and Corollary 2.31.

Definition 2.34. *A Borel subset of $\mathbb{R}^2$ that contains no pairs (x',y'), (x,y^+) and (x,y^-) s.t. $x < x'$ and $y^- < y' < y^+$, is called left-monotone. A martingale transport giving measure 1 to a left-monotone set is called left-monotone.*

Hence Example 2.33 shows that Γ from Theorem 2.28 is a left-monotone set if $c_{xyy} < 0$. In light of Theorem 2.28, as soon as $P_c^{mg} < \infty$ we have that any optimal martingale transport is left-monotone. Now we want to show that if $\mu \ll \mathsf{Leb}$, then there exists a unique left-monotone transport which corresponds to (MOT) having a unique minimizer w.r.t. Spence-Mirrlees cost-functions. To this end, we need the following lemma:

Lemma 2.35. *Let $k \in \mathbb{N}$, $\Gamma \subseteq \mathbb{R}^2$ and assume there exist uncountable many $a \in \mathbb{R}$ s.t. $|\Gamma_a| \geq k$, where $\Gamma_a = \{y : (a,y) \in \Gamma\}$. Then, there exists $a \in \mathbb{R}$ and $b_1 < ... < b_k$ in Γ_a s.t. for all $\varepsilon > 0$ one can find $a' > a$ and $b_1' < ... < b_k'$ in $\Gamma_{a'}$ with*

$$\max\{|a-a'|, |b_1 - b_1'|, ..., |b_k - b_k'|\} < \varepsilon.$$

Proof. This is a cardinality argument. Call $A := \{a : |\Gamma_a| \geq k\}$ and for $a \in A$ choose $b_1^a < \cdots < b_k^a$ in Γ_a. Set $\Gamma_A = \{(a, b_1^a, \ldots, b_k^a) : a \in A\}$. Call $(a, b_1^a, \ldots, b_k^a)$ a right accumulation point if there is $A \ni a_n \searrow a$ with $b_i^{a_n} \to b_i^a$ for all $i = 1, \ldots, k$. Call $I_A \subseteq \Gamma_A$ the complement in Γ_A of the set of accumulation points. We leave it as an exercise to check that I_A is at most countable. It follows that there are uncountably many right accumulation points in Γ_A, and any such point serves to finish the proof. □

So assume there exist uncountable many a s.t. $|\Gamma_a| \geq 3$ where Γ is a left-monotone set. Pick a and $y_1 < y_2 < y_3$ in Γ_a as in the lemma above. For $\varepsilon < \min\{y_2 - y_1, y_3 - y_2\}$ there exists $a' > a$ and $y_1' < y_2' < y_3'$ with

$$\max\{|a-a'|, |y_1 - y_1'|, |y_2 - y_2'|, |y_3 - y_3'|\} < \varepsilon.$$

In particular, y_2' satisfies $y_1 < y_2' < y_3$ contradicting the left-monotone property of Γ.

Thus, if Γ is left-monotone, then there are at most countable many a with $|\Gamma_a| > 2$. As a consequence, we have $|\Gamma_a| \leq 2$ μ-a.s. for all $\mu \ll \mathsf{Leb}$. Define for $T_1(a) =$ the smallest of the (at most) two elements in Γ_a, and $T_2(a)$ the largest one. Up to redefinition on a μ-null set, these are Borel functions. (In reality we must apply here a so-called "measurable selection theorem", but this is beyond the scope of these notes.) Hence we obtained that any left-monotone martingale coupling (resp. any minimizer of (MOT) w.r.t. Spence-Mirrlees cost functions) is concentrated on the union of the graphs of the two measurable functions T_1 and T_2. By a straightforward variant of Remark 1.39 we have uniqueness (Exercise!). Since the existence of the derivative c_{xyy} implies continuity, we also have existence. Notice

that the specific structure of c did not play any role, similar to the convexity assumption in 1-dimensional optimal transport (c.f. Ex 1.46).

We can summarize the above discussion in the following proposition:

Proposition 2.36. *Assume the generalized Spence-Mirrlees condition $c_{xyy} < 0$, and that $P_c^{mg} < \infty$. Then a martingale transport is optimal for P_c^{mg} iff it is left-monotone. In case $\mu \ll \mathsf{Leb}$ the minimizer is unique, and it is concentrated on the graph of two functions T_1, T_2 satisfying*

- *$T_1(x) \le x \le T_2(x)$;*
- *$x < x'$ implies $T_2(x) < T_2(x')$ and $T_1(x') \notin (T_1(x), T_1(x'))$.*

Finally: the here described martingale couplings do not depend on the cost function c.

In fact we can say more about the case when μ has atoms, but we rather stop the discussion at this point. We leave it as an exercise to characterize optimizers in case of the opposite Spence-Mirrlees condition: $c_{xyy} > 0$. In this case optimizers are called right-monotone martingales.

We provide, for the curious reader, a more general statement too:

Theorem 2.37. *Let $h \in C^2(\mathbb{R})$, $c(x,y) = h(y-x)$, and assume $P_c^{mg} < \infty$. If affine functions intersect the graph of h' in at most k points and $Q \in \mathsf{MT}(\mu,\nu)$ is optimal, then there exists a disintegration $(Q_x)_x$ s.t. for all $x \in \mathbb{R}$ at least one of the following is true:*

(i) $\mu(\{x\}) > 0$.
(ii) $|\operatorname{supp}(Q_x)| \le k$.

Finally we remark that the important case of a cost function having a kink can be similarly studied:

Example 2.38. Let $c(x,y) = |x-y|$. If μ possesses no atoms, then there exists a unique optimal martingale coupling π concentrated on the graph of two non-decreasing functions T_1 and T_2 such that $T_1(x) \le T_2(x)$ concentrated on a set Γ not containing pairs (x', y'), (x, y^+) and (x, y^-) s.t. $y^- < y' < y^+$ and either $y' \le x' < x$ or $x < x' \le y'$. The proof is similar to the case of Spence-Mirrlees cost functions, but needs more combinatorial arguments.

2.2.1. *Martingale inequalities.* Let us briefly describe an application closely related to martingale OT, namely the subject of *martingale inequalities*. These are statements of the form

For all martingales M in a given class C, we have $\mathbb{E}[F(M)] \le 0$.

Of course the class C and the functional F have to be specified. Probably the most celebrated martingale inequality is Doob's L^2 inequality, which we state here for discrete time square integrable martingales $\{M_t : t = 1, \dots, n\}$:

$$\mathbb{E}\left[\left(\sup_{t=1,\dots,n} M_t\right)^2\right] \le 4\mathbb{E}[M_n^2]. \tag{2.4}$$

Notice that the constant 4 in the r.h.s. is independent of n, so unsurprisingly an analogue inequality holds in continuous time too.

We emphasize that (2.4) for $n = 2$ can be obtained directly from the identity $(M_1 \vee M_2)^2 \le M_1^2 + M_2^2$. However for $n > 2$ this naive approach would give an n-dependent constant, worsening as n increases. Instead, let us revisit the case $n = 2$ through the lens of martingale OT, and more precisely, via the duality result presented in Theorem 2.13. We take $c(x,y) = -(x \vee y)^2$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ with $\mu \le_c \nu$. Then

$$P_c^{mg}(\mu,\nu) = \sup\left\{\int \varphi d\mu + \int \psi d\nu : (\varphi, \psi, h)\ s.t.\ \varphi(x) + \psi(y) + h(x)[y-x] \le -(x \vee y)^2\right\}.$$

We now take as an educated guess $\varphi \equiv 0$, $\psi(y) = -4y^2$ and $h(x) = 4x$. We have

$$-4y^2 + 4x(y-x) = -4y^2 + 4xy - 4x^2,$$

which equals both $-(2x-y)^2-3y^2$ and $-(2y-x)^2-3x^2$, and so it is in any case smaller or equal than $-(x\vee y)^2$. Thus whenever $M=(M_0,M_1)$ is a martingale and $M_1\sim\mu, M_2\sim\nu$, we have

$$-\mathbb{E}[(M_1\vee M_2)^2]\geq P_c^{mg}(\mu,\nu)\geq -4\mathbb{E}[M_2^2].$$

Since μ,ν were otherwise arbitrary, we conclude (2.4) for $n=2$. Emboldened by this method, we first easily check, for $n\in\mathbb{N}$, the following weak duality:

$$\inf_{\substack{\{M_t\}_{t=1}^n \text{ martingale}\\ M_i\sim\mu_i,\, i=1,\dots,n}} \mathbb{E}\left[c(M_1,\dots,M_n)\right]\geq \sup_{\substack{u_i\in C_b(\mathbb{R}), i=1,\dots,n,\ h_t\in C_b(\mathbb{R}^t),\, t=1,\dots,n-1\\ \sum_{i=1}^n u_i(x_i)+\sum_{t=1}^{n-1} h_t(x_1,\dots,x_t)(x_{t+1}-x_t)\leq c(x_1,\dots,x_n)}} \sum_{i=1}^n \int u_i d\mu_i, \quad (2.5)$$

where the μ_i are in increasing convex order so that in the l.h.s. we are optimizing over a non-empty set. Similar to the case $n=2$, we consider $c(x_1,\dots,x_n)=-(\max_{s=1,\dots,n} x_s)^2$ and make an educated guess: $u_i\equiv 0$ if $i<n$, $u_n(y)=-4y^2$ and $h_t(x_1,\dots,x_t)=4\max_{s=1,\dots,t} x_s$. We leave it as an exercise (see Problem 36) that these functions do define a feasible element for the r.h.s. and then conclude (2.4).

Remark 2.39. In fact it is possible to obtain equality (and existence of optimizers for the l.h.s.) in (2.5). See Section 2.2.2 for some details. We also stress that there are many other martingale inequalities, and it is known that essentially for all these there is a duality argument as in the above Doob L^2 case.

We close by rephrasing the arguments above in terms of mathematical finance: We considered as financial derivative $C:=-(\max_{s=1,\dots,n} M_s)^2$ and showed that the highest model-free fair price for C, given data at time 1 and time n, is smaller than $4\mathbb{E}[M_n^2]$. In order to do this, we exhibited that the sum of the static derivatives $u_i\equiv 0$ (for $i<n$) and $u_n=-4M_n^2$ with the outcome of the trading strategy $h_t=4\max_{s=1,\dots,t} M_s$ sub-replicates the derivative C.

2.2.2. *Martingale OT in multiple periods.* Fix $n\in\mathbb{N}$ and $\{\mu_i\}_{i=1}^n\subseteq\mathcal{P}_1(\mathbb{R})$ in convex order. The n-period martingale optimal transport problem with cost function $c:\mathbb{R}^n\to\mathbb{R}$ is given, in probabilistic notation, by

$$\inf_{\substack{\{M_t\}_{t=1}^n \text{ martingale}\\ M_i\sim\mu_i,\, i=1,\dots,n}} \mathbb{E}\left[c(M_1,\dots,M_n)\right],$$

namely the l.h.s. of (2.5). In terms of martingale transport plans (measures), this reads equivalently

$$\inf_{\substack{Q\in\mathcal{P}(\mathbb{R}^n)\ s.t.\ p_i(Q)=\mu_i,\, i=1,\dots,n,\\ \int (x_{k+1}-x_k)q^{x_1,\dots,x_k}Q(dx_{k+1})=0,\,\forall k=1,\dots,n-1}} \int c(x_1,\dots,x_n)Q(dx_1,\dots,dx_n). \quad (2.6)$$

So if μ_n has finite first moment, then the minimization problem (2.6) runs over a W_1-compact set of probability measures in $\mathbb{R}^n$ (exercise) and hence it admits a minimizer as long as c is lower semicontinuous and lower bounded by a function with linear growth. Under these assumptions we obtain no-duality-gap in (2.5) by a minimax argument (exercise). But more can be said: for instance if ν has a finite second moment, then the infimum runs over a compact set w.r.t. the W_2-topology on $\mathcal{P}_2(\mathbb{R}^n)$, and so on.

Let us finally remark that another possible Martingale OT problem in multiple periods is given by

$$\inf_{\substack{\{M_t\}_{t=1}^n \text{ martingale}\\ M_1\sim\mu,\, M_n\sim\nu}} \mathbb{E}\left[c(M_1,\dots,M_n)\right],$$

that is, here only initial and terminal marginals are fixed, although the cost function sees the whole path. Here it is also possible to develop a theory, so that for instance we have existence of minimizers and no duality gap. One elegant idea (not the simplest, though) is

to begin with a martingale inequality! For instance Problem 36, related to the above Doob L^2 inequality, states that for all $s_1, \dots, s_n \in \mathbb{R}$ we have

$$\left[\max_{i=1,\dots,n} s_i\right]^2 \leq 4s_n^2 - 4\sum_{k=1}^{n-1}\left(\max_{i=1,\dots,k} s_i\right)[s_{k+1} - s_k].$$

This already shows that, denoting by $\|\cdot\|_\infty$ the sup norm in $\mathbb{R}^n$, we have

$$\text{Probability}(\|M\|_\infty \geq K) \leq \frac{4\int y^2 d\nu(y)}{K^2},$$

for all M participating in the above infimum. Hence if ν has a finite second moment we can obtain from here that the set of laws of all M participating in that infimum is a compact set w.r.t. the weak topology, and so on.

2.3. **Existence, duality, and geometry of optimizers: continuous time.** In this section we want to sketch how one can build a continuous time theory for MOT in dimension 1 (higher dimensional run along similar lines, but are not only notationally more involved, e.g. regarding irreducible components). Since there are no non-trivial Monge-type solutions to discrete MOT, continuous time MOT necessarily has to be of diffusive nature. Moreover, if (S_0, S_1) is a martingale under a measure q there is no obvious recipe to construct a continuous martingale interpolating between $\mathsf{Law}_q(S_0)$ and $\mathsf{Law}_q(S_1)$ even less if the 2-marginal restriction of the resulting martingale is supposed to give back the original coupling.

Instead, we directly develop a continuous time theory. To this end, we recall the Benamou-Brenier picture of the Kantorovich-Wasserstein distance W_2 from Theorem 1.70

$$W_2^2(\mu,\nu) = \inf_{(\rho,v)\in V(\mu,\nu)} \int_0^1 \int |v_t|^2 \rho_t(x) dx dt.$$

In probabilistic terms this can be reformulated as

$$W_2^2(\mu,\nu) = \inf\left\{\mathbb{E}\int_0^1 |\dot{X}_t|^2 dt : X_0 \sim \mu, X_1 \sim \nu\right\},$$

where the infimum runs over all random C^1 curves $(X_t)_{t\in[0,1]}$ with derivative $\dot{X}_t$ at time t. Since continuous martingales are typically only Hölder continuous there is no chance to make sense of a time derivative of a martingale M. However, if $M_1 \sim \nu$ has second moments then the quadratic variation of M, denoted by $\langle M\rangle$, is well defined (recall that the quadratic variation is the unique increasing adapted process starting in 0 such that $t \mapsto M_t^2 - \langle M\rangle_t$ is a martingale). In particular, $t \mapsto \langle M\rangle_t$ is non-decreasing and we can associate its Lebesgue-Stieltjes measure by setting $\int_s^t d\langle M\rangle_u := \langle M\rangle_t - \langle M\rangle_s$. If $d\langle M\rangle_t \ll dt$ we denote its density w.r.t. dt by $\langle \dot{M}\rangle_t$, and interpret it as the speed of the martingale. If $d\langle M\rangle_t$ is not absolutely continuous w.r.t. dt we put $\langle \dot{M}\rangle_t = \infty$.

With this interpretation in mind we introduce the following martingale variant of the Benamou-Brenier minimization problem:

Definition 2.40. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with $\mu \preceq_c \nu$. Then, we define*

$$\mathsf{C}_{BB}(\mu,\nu) := \inf\left\{\mathbb{E}\int_0^1 c(\langle \dot{M}\rangle_t) dt : M \text{ martingale }, M_0 \sim \mu, M_1 \sim \nu\right\}. \tag{MOTBB}$$

This is a variant of the Benamou-Brenier formula considering the stochastic side of the picture. Looking at the analytical part, we need to find a PDE of the evolution of the marginals that replaces the continuity equation (1.9). To this end, let M be a solution to the stochastic differential equation (SDE)

$$dM_t = \sigma(t, M_t) dB_t.$$

Let $a_t = \sigma_t^2$ and $f \in C_b^2(\mathbb{R})$. Then, it follows by Itô's formula that

$$f(M_t) - f(M_0) = \int_0^t f'(M_s) dB_s + \frac{1}{2}\int_0^t f''(M_s) a_s(M_s) ds.$$

Denoting the time t marginal of M by ρ_t we thus obtain by taking expectations (recalling that $\int \cdot dB_t$ is a martingale)

$$\int f d\rho_t - \int f d\rho_0 = \frac{1}{2}\int_0^t \int f''(x) a_s(x) d\rho_s(x) ds,$$

so that (ρ, a) satisfy the Fokker-Planck-equation (to be understood in a weak sense as for (1.9))

$$\partial_t \rho_t = \frac{1}{2}\Delta a_t \rho_t.$$

This leads us to the following second version of a martingale Benamou-Brenier formula, from an analytical PDE point of view:

Definition 2.41. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with $\mu \preceq_c \nu$. Then, we define*

$$\mathsf{C}_{FPE}(\mu,\nu) := \inf\left\{\int_0^1 \int c(a_t(x))\rho_t(dx)dt : \partial_t\rho_t = \frac{1}{2}\Delta a_t\rho_t, \rho_0 = \mu, \rho_1 = \nu\right\}. \qquad \text{(FPE)}$$

Remark 2.42. Both optimization problems (MOTBB) and (FPE) are closely linked via the martingale problem in stochastic analysis. Indeed, the martingale problem associated to the operator $a_t\Delta$ induces a curve of marginals $(\rho_t)_t$ solving the Fokker-Planck equation. Given a pair (ρ, a) solving the Fokker-Plank equation, there is, by the superposition principle [Tre16, Theorem 2.5], a continuous process $M = (M_t)_{t\in[0,1]}$ solving the martingale problem associated to $a\Delta$ such that $\rho_t = \mathsf{Law}(M_t)$ for every $t \in [0,1]$.

As an important consequence of the last remark, any candidate for (FPE) induces a candidate for the problem (MOTBB) with the same cost. Therefore, we obtain that $\mathsf{C}_{BB}(\mu,\nu) \le \mathsf{C}_{FPE}(\mu,\nu)$.

The other direction, i.e. $\mathsf{C}_{BB}(\mu,\nu) \ge \mathsf{C}_{FPE}(\mu,\nu)$, is in general not true. However, if the cost function c is strictly convex, then it is possible to construct from an admissible candidate for (MOTBB) an admissible candidate for (FPE) by conditioning (analytically projecting) on the current value. In general, the so constructed candidate for (FPE) will have lower cost than the candidate for (MOTBB) we started with. Putting these ideas into rigorous formulas leads to the following result, for a proof of which we refer to [HT19, Theorem 3.3].

Proposition 2.43. *Assume that c is strictly convex and there are positive constants λ, Λ such that $\lambda|a|^p \le c(a) \le \Lambda|a|^p$ for some $p \in (1,\infty)$. Then, $\mathsf{C}_{BB}(\mu,\nu) = \mathsf{C}_{FPE}(\mu,\nu)$.*

Combining Proposition 2.43 with Remark 2.42 shows that in the situation of the last Proposition it is sufficient to only solve either the stochastic or the analytical version of the martingale Benamou-Brenier formulation since either of them induces a solution to the other. We will first focus on the PDE/FPE version and show how optimizer are closely related to the porous medium equation. We will use this information to explicitly construct optimizer for a specific cost in the stochastic version of the problem.

Similar to discrete time the key to understand the optimizer is the dual problem. To write it down, recall the Legendre transform c^* of c is defined via $c^*(u) = \sup_a a \cdot u - c(a)$. We will also need to consider the following Hamilton-Jacobi-Bellman PDE

$$\partial_t\varphi(t,x) = -c^*\left(\frac{1}{2}\Delta\varphi(t,x)\right). \qquad \text{(HJB)}$$

We say that $\varphi(t,x) = \varphi_t(x) \in C^{1,2}([0,1]\times\mathbb{R})$ is a solution to (HJB) if there is equality in (HJB) and we say it is a supersolution if the inequalities $\le$ holds for every t, x.

Theorem 2.44 (Duality)**.** *Let c be as in Proposition 2.43. Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ such that $\mathsf{C}_{FPE}(\mu,\nu) < \infty$. Then,*

$$\mathsf{C}_{FPE}(\mu,\nu) = \sup\left\{\int \varphi_1(x)d\nu - \int \varphi_0(x)d\mu\right\},$$

where the supremum runs over all smooth supersolutions to (HJB)*. Moreover, the infimum in* (FPE) *is in fact a minimum.*

Proof. We only show the easy part of the duality relation. The full result, including the existence of an optimizer, follows by an application of the Fenchel-Rockafellar duality result, see [HT19, Theorem 4.3].

Let φ be any supersolution to (HJB) and (ρ, a) be an admissable candidate for (FPE). Then, we have the following chain of inequalities

$$\begin{aligned}\int_{\mathbb{R}} \varphi(1,x)d\rho_1(x) - \int_{\mathbb{R}} \varphi(0,x)d\rho_0(x) &= \int_0^1 \frac{d}{dt}\int \varphi_t d\rho_t dt \\ &= \int_0^1 \int_{\mathbb{R}} \left(\partial_t\varphi(t,x) + \frac{1}{2}a_t(x)\Delta\varphi(t,x)\right) d\rho_t(x)dt \\ &\leq \int_0^1 \int_{\mathbb{R}} \left(-c^*(\frac{1}{2}\Delta\varphi(t,x)) + \frac{1}{2}a_t(x)\Delta\varphi\right) d\rho_t(x)dt \\ &\leq \int_0^1 \int_{\mathbb{R}} c(a_t(x))d\rho_t(x)dt. \end{aligned} \tag{2.7}$$

Minimizing over all possible pairs (ρ, a) yields that

$$\int_{\mathbb{R}} \varphi(1,x)d\rho_1(x) - \int_{\mathbb{R}} \varphi(0,x)d\rho_0(x) \leq \mathsf{C}_{FPE}(\mu,\nu).$$

□

Remark 2.45. Observe that the set of solutions to the Fokker-Planck equation with fixed initial and terminal law is a convex set (w.r.t. the usual linear convex combinations). Therefore, the strict convexity of c implies directly the uniqueness of solutions to (FPE).

However, this does not imply that the solutions to (MOTBB) are unique as well, in the sense that the laws of the corresponding stochastic processes are unique. The reason is that there might be multiple solutions to the corresponding martingale problem. The question of uniqueness to the martingale problem is closely linked to the regularity of the diffusion coefficient a. In dimension one, Lipschitz continuity is sufficient to grant uniqueness in the martingale problem.

As a consequence of duality we obtain the following sufficient optimality criterion:

Corollary 2.46. *Let c be as above and assume that $\varphi \in C^{1,2}([0,1]\times\mathbb{R})$ solves* (HJB)*. Put*

$$a_t(x) := \nabla c^*(\frac{1}{2}\Delta\varphi(t,x)). \tag{2.8}$$

Let ρ_t be a solution to the Fokker-Planck equation w.r.t. a. Then, $(\rho_t)_t$ is a minimizer for (FPE)*, i.e.*

$$\mathsf{C}_{FPE}(\rho_0,\rho_1) = \int_0^1 \int c(a_s)d\rho_s ds.$$

Proof. Observe, that a satisfies for every t, x the optimality condition in the Legendre transform so that

$$c(a_t(x)) + c^*(\frac{1}{2}\Delta\varphi_t(x)) = a_t(x)\cdot\frac{1}{2}\Delta\varphi_t(x).$$

Going back to (2.7) we see that the first inequality is an equality since φ is assumed to be a solution. The second inequality is an equality as well since a and $\frac{1}{2}\Delta\varphi$ are dual variables for the Legendre transform. □

Example 2.47. Brownian motion is an optimizer to (FPE). Indeed, let $\mu \in \mathcal{P}(\mathbb{R})$, $X_0 \sim \mu$, B be standard Brownian motion, $\nu = \mathsf{Law}(X_0 + B_1)$, $M_t = X_0 + B_t$ s.t. M has diffusion coefficient $a(t,x) = 1$. Then, for any cost function c as in the last Corollary, there is $R \in \mathbb{R}$ such that $a \equiv 1 = \nabla c^*(R/2)$. In particular,

$$\varphi(t,x) := -t\cdot c^*(R/2) + \frac{1}{2}Rx^2$$

solves (HJB). Hence, M is optimal by Corollary 2.46.

Remark 2.48. Combining (2.8) with (HJB) it is possible to write down an equation for the optimal diffusion coefficient. Assume that $c^*(u) = 2(u_+)^q$, so that $c(a) = a^p/(p(2q)^{p/q})$ for p and q satisfying $1/p + 1/q = 1$. Then, one gets

$$\partial_t a(t,x) = -\frac{1}{2p}\frac{\Delta a^p(t,x)}{a^{p-2}(t,x)} = -\frac{1}{2}\left(a(t,x)\Delta a(t,x) + (p-1)(\partial_x a(t,x))^2\right). \tag{2.9}$$

Remark 2.49. To construct further examples using Corollary 2.42, resp. to advance this theory further, it seems to be useful to connect (HJB) to the porous medium equation

$$\partial_t u(t,x) = -\Delta\frac{1}{2}c^*(u(t,x)). \tag{PME}$$

Indeed, set $u = \frac{1}{2}\Delta\varphi$ and taking $\frac{1}{2}\Delta$ on both sides of (HJB) yields (PME). In fact, (2.9) is the so called pressure equation associated to the porous medium equation.

The advantage is that (PME) is very well studied and various results on existence, uniqueness, regularity are available (see [Váz07]). However, for a natural class of cost functions, namely $c(a) = a^p$ with $p > 1$, the known results to (PME) seem to be not powerful enough to make this passage even more fruitful by e.g. providing an existence and regularity theory for dual optimizer (leading to potential uniqueness statements for the stochastic variant via the martingale problem). The problem is that for these examples the Legendre transform is $c^*(u) = (u_+)^q$ with $q = \frac{p}{p-1}$ and $u_+ = u \vee 0$ which is not regular enough at zero for classical results to be applicable. One needs to develop new theory to resolve this issue, which is still open.

In the rest of this section we want to analyze optimizers for the stochastic variant of the Benamou-Brenier picture for MOT for a particular cost function.

In Example 2.47, we have seen that Brownian motion is an optimizer for (MOTBB) for any convex cost function c. This is of course only possible if μ and ν are related via $\nu = \mu * \gamma$, where $*$ denotes convolution and γ is a standard Gaussian measure. We now want to construct a martingale which is "as close as possible" to Brownian motion in a natural way. The prototypical example is the following:

Example 2.50 (Bass martingale). Let $\mu = \delta_0$ and ν be arbitrary with mean 0, so that $\mu \preceq_c \nu$. Recall that γ denotes a standard Gaussian. Let f be an increasing map such that $f(\gamma) = \nu$. This map exists, e.g. by Brenier's Theorem. Let $B \equiv (B_t)_{t\in[0,1]}$ be a 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 $t \in [0,1]$

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

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 $M_1 = f(B_1) \sim \nu$ and $M_0 = P_1 f(0) = \mathbb{E}[f(B_1)] = \int f(x)\nu(dx) = 0$ 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. To allow for more general initial measures, the only flexibility that we have is to allow for a general starting measure α of Brownian motion.

So let us assume that B is a Brownian motion with initial law α so that $B_1 \sim \alpha * \gamma$. Then, for any measure ν there is an increasing map f such that $f(\alpha * \gamma) = \nu$ (by Brenier's result). Then, we can define again

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

with the difference that $M_0 = P_1 f(B_0) \sim P_1 f(\alpha)$. Hence, for any α we get a potentially different (note that f depends on α) initial measure $P_1 f(\alpha)$.

We call the martingale M a *standard stretched Brownian motion* since in a certain sense M tries to behave like Brownian motion but in a deformed/stretched geometry.

The question we aim to answer in the following is, whether for any pair $\mu \preceq_c \nu$ there is an α such that the resulting martingale M connects μ to ν?

To answer this question we will identify the standard stretched Brownian motion as an optimizer to (MOTBB) for a particular cost function. By construction a standard stretched Brownian motion satisfies the SDE

$$dM_t = \sigma_t(M_t)dB_t$$

with $\sigma_t = \nabla f_t \circ f_t^{-1}$. Writing $a = \sigma^2$ and calculating various derivatives, one can see that a satisfies (2.9) with $p = 1/2$. This leads us outside the case of strictly convex functions (and $p > 1$). It suggest to consider the cost function $c(a) = \sqrt{a}$ together with a maximization problem. We introduce

$$\mathsf{C}_{MBB} := \mathsf{C}_{MBB}(\mu,\nu) := \sup\left\{\mathbb{E}\int_0^1 \sigma_t(M_t)dt : M_t = M_0 + \int_0^t \sigma_s(M_s)dB_s, M_0 \sim \mu, M_1 \sim \nu\right\}. \tag{MBB}$$

(Here the optimization runs over all sufficiently rich probability spaces supporting Brownian motion B and a progressively measurable σ.)

Definition 2.51. *Any optimizer to* (MBB) *is called stretched Brownian motion.*

We have the following result clarifying the precise relation between stretched and standard stretched Brownian motion:

Theorem 2.52. *A candidate martingale M^* is an optimizer to* (MBB) *iff it is a standard stretched Brownian motion on each irreducible component of (μ,ν). In particular, stretched Brownian motion is a standard stretched Brownian motion in each irreducible component.*

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*:

$$\mathsf{C}_{WOT} := \mathsf{C}_{WOT}(\mu,\nu) := \sup\left\{\int \mu(dx) \sup_{q\in \mathsf{Cpl}(\pi_x,\gamma)} \int q(dm,db)m\cdot b\right\}, \tag{WT}$$

where the (first) supremum runs over all family of kernels $(\pi_x)_{x\in\mathbb{R}}$ s.t. $\int y\pi_x(dy) = x$ for all x and $\int \mu(dx)\pi_x(dy) = \nu(dy)$. Note that the cost function is non-linear in the optimization variable. Completing the square in (WT) shows that

$$1 + \int y^2 d\nu(y) - 2\mathsf{C}_{WOT} = \inf\left\{\int \mu(dx) W_2^2(\pi_x,\gamma)\right\},$$

where the infimum runs over the same set of kernels $(\pi_x)_x$ as above.

Theorem 2.53. *Assume that μ,ν have second moments. The optimization problem* (WT) *and* (MBB) *are equivalent. More precisely,*

(1) $\mathsf{C}_{WOT} = \mathsf{C}_{MBB} < \infty$;
(2) (WT) *has a unique optimizer;*
(3) (MBB) *has a unique-in-law optimizer M^*;*
(4) $\pi^* = \mathsf{Law}(M_1^*, M_0^*)$ *and* $M^* = G(\pi^*)$ *for some function G, i.e. M^* can be explicitly costructed from π^*.*

Proof. Let M be feasible for (MBB). By Itô's formula and the martingale property of M we have

$$\mathbb{E}\left[\int_0^1 \sigma_t\, dt\right] = \mathbb{E}[M_1\cdot B_1 - M_0\cdot B_0] = \mathbb{E}[M_1\cdot(B_1-B_0)] = \mathbb{E}[\ \mathbb{E}[M_1\cdot(B_1-B_0)\,|\,M_0]\ ].$$

Letting $q_x = law\,(M_1, B_1 - B_0\,|\,M_0 = x)$ we find $q_x \in \mathsf{Cpl}(\pi_x\,,\,\gamma)$ for $\pi_x = law\,(M_1\,|\,M_0 = x)$ and

$$\mathbb{E}\left[\int_0^1 \sigma_t\, dt\right] = \int \mu(dx)\int q_x(dm,db)\ m\cdot b.$$

From this we easily conclude $\mathsf{C}_{WOT} \geq \mathsf{C}_{MBB}$.

Now let π be feasible for (WT). For each x we can find $F^x(\cdot)$ convex such that $\nabla F^x(\gamma) = \pi_x$. We now define $M_t^x := E[\nabla F^x(B_1)|\mathcal{F}_t^B]$ for a given standard Brownian motion on $\mathbb{R}$ with Brownian filtration $\mathcal{F}^B$. Potentially enlarging our probability space we can assume

the existence of a random variable X independent of the Brownian motion B with $X \sim \mu$. We denote the filtration (on the potentially bigger probability space) by $\mathcal{F}$. Since $M_0^x = \int y\pi_x(dy) = x$ and $\int \mu(dx)\pi_x(dy) = \nu(dy)$ we conclude that $\{M_t^X\}_{t\in[0,1]}$ is a continuous martingale from μ to ν. By construction

$$\int \mu(dx) \sup_{q\in\mathsf{Cpl}(\pi_x,\gamma)} \int q(dm,db)\ m\cdot b = \int \mu(dx) \int \gamma(db)\ b\cdot\nabla F^x(b) = \mathbb{E}\left[\mathbb{E}\left[B_1\cdot M_1^X|X\right]\right],$$

and the last term equals $\mathbb{E}[\int_0^1 \sigma_t dt]$ as before (σ can easily be computed from ∇F^x). This proves $\mathsf{C}_{WOT} \le \mathsf{C}_{MBB}$ and hence $\mathsf{C}_{WOT} = \mathsf{C}_{MBB}$. The finiteness $\infty > \mathsf{C}_{WOT}$ follows from $m\cdot b \le |m|^2 + |b|^2$ and ν and γ having finite second moment.

To show that (WT) is attained let us denote by $(\pi^n)_{n\in\mathbb{N}}$ (where $\pi^n(dx,dy) = \pi_x^n(dy)\mu(dy)$) an optimizing sequence. The set $\mathsf{MT}(\mu,\nu)$ by Lemma 2.10. By [Bal00, Theorem 3.7] we obtain the existence of a measurable kernel $x \mapsto \pi_x \in \mathcal{P}(\mathbb{R})$ and a subsequence, still denoted by $(\pi^n)_n$, such that on a μ-full set

$$\tfrac{1}{N}\textstyle\sum_{n\le N} \pi_x^n(dy) \to \pi_x(dy),$$

with respect to weak convergence in $\mathcal{P}(\mathbb{R})$. In particular $\frac{1}{N}\sum_{n\le N}\pi^n \to \pi$ in the weak topology in $\mathcal{P}(\mathbb{R}\times\mathbb{R})$, where $\pi(dx,dy) := \mu(dx)\pi_x(dy)$. Since $\mathsf{MT}(\mu,\nu)$ is closed, we have that $\pi \in \mathsf{MT}(\mu,\nu)$. Finally,

$$\begin{aligned}
\mathsf{C}_{WOT} &= \lim_n \int \mu(dx) \sup_{q\in\mathsf{Cpl}(\pi_x^n,\gamma)} \int q(dm,db)\ m\cdot b \\
&= \lim_N \int \mu(dx) \tfrac{1}{N}\textstyle\sum_{n\le N} \sup_{q\in\mathsf{Cpl}(\pi_x^n,\gamma)} \int q(dm,db)\ m\cdot b \\
&\le \lim_N \int \mu(dx) \sup_{q\in\mathsf{Cpl}(\frac{1}{N}\sum_{n\le N}\pi_x^n,\gamma)} \int q(dm,db)\ m\cdot b \\
&\le \int \mu(dx) \limsup_N \sup_{q\in\mathsf{Cpl}(\frac{1}{N}\sum_{n\le N}\pi_x^n,\gamma)} \int q(dm,db)\ m\cdot b \\
&\le \int \mu(dx) \sup_{q\in\mathsf{Cpl}(\pi_x,\gamma)} \int q(dm,db)\ m\cdot b \ \le \mathsf{C}_{WOT}\,.
\end{aligned}$$

The first inequality holds by concavity of $\eta \mapsto H(\eta) := \sup_{q\in\mathsf{Cpl}(\eta,\gamma)} \int q(dm,db)\, m\cdot b$ w.r.t. convex combinations of measures. The second inequality is Fatou's lemma, noticing that the integrand is bounded in $L^1(\mu)$ (the bound equals the sum of the second moments of μ and γ). The third inequality follows by weak convergence of the averaged kernel on a μ-full set and upper semicontinuity of $H(\cdot)$. For uniqueness it suffices to notice that $H(\cdot)$ is actually strictly concave, which is an easy consequence of Brenier's Theorem. Hence, (WT) is attained and we denote the unique optimizer by π^*.

Taking π^* we may build an optimizer M^* for (MBB) as in the first part of the proof (as the value of both problems agree).

We finally establish the uniqueness of optimizers for (MBB). Let $\tilde{M}$ be any such optimizer. From the previous considerations, we deduce that the law of $(\tilde{M}_0, \tilde{M}_1)$ is the unique optimizer π^* of (WT). Conditioning on $\{\tilde{M}_0 = x\}$ we thus have that $\tilde{M}$ connects δ_x to π_x^*. It follows that $\mu(dx)$-a.s. $\tilde{M}$ conditioned on $\{\tilde{M}_0 = x\}$ is optimal between these marginals. Indeed,

$$\sup_{N_t = x + \int_0^t \sigma_s\, dB_s,\, N_1\sim\pi_x^*} \mathbb{E}\left[\int_0^1 \sigma_t\, dt\right] = \sup_{q\in\mathsf{Cpl}(\pi_x^*,\gamma)} \int q(dm,db)\ m\cdot b\ , \tag{2.10}$$

by the results obtained so far, since if $\tilde{M}$ conditioned on $\{\tilde{M}_0 = x\}$ was not optimal for the l.h.s. it could not deliver the equality $\mathsf{C}_{WOT} = \mathsf{C}_{MBB}$. So it suffices to show that the l.h.s. of (2.10) is uniquely attained. But any candidate martingale N with volatility σ satisfies $\mathbb{E}[\int_0^1 \sigma_t dt] = \mathbb{E}[N_1 B_1]$ (since here we can assume $B_0 = 0$). Hence, Brenier's Theorem implies that $\tilde{M}_1 = \nabla F^x(B_1)$ on $\{\tilde{M}_0 = x\}$, for a convex function F^x. Since the optimal transport map ∇F^x is unique, and the martingale property determines uniquely the law of $\tilde{M}$, we finally get $\tilde{M} = M^*$ in law. □

Remark 2.54. The proof of Theorem 2.53 shows how to build the optimizer for (MBB) via the following procedure, making the statement $M^* = G(\pi^*)$ in Theorem 2.53 (4) precise:

(1) Find the unique optimizer π^* of (WT).

(2) Find convex functions F^x such that $\nabla F^x(\gamma) = \pi^*_x$.
(3) Define $M^x_t := \mathbb{E}[\nabla F^x(B_1)|B_t] = P_{1-t}\nabla F^x(B_t)$.
(4) Take $X \sim \mu$ independent of B and let $M_t := M^X_t$.

Remark 2.55. Observe also, that Theorem 2.53 implies that any standard stretched Brownian motion is a stretched Brownian motion.

To show that any stretched Brownian motion is a standard stretched Brownian motion on each irreducible component we need to argue that the (potentially) different maps F^x do in fact agree. One way to argue this is via combining the strong Markov property of Brownian motion with a monotonicity principle for the weak transport problem. This is slightly outside the scope of this lecture and we refer the interested reader to [BVBHK17]. There you can find also extensions/and limitations/challenges to higher dimensions.

2.4. **Exercises.**

Problem 27. (Convex order I)

For $\mu, \nu \in \mathcal{P}_1(\mathcal{R})$, show that the following are equivalent:

(1) $\mu \leq_c \nu$.
(2) μ and ν have the same mean and $\int f d\mu \leq \int f d\nu$ for all f non-negative, non-decreasing, convex, and Lipschitz.
(3) μ and ν have the same mean and $\int [x-t]_+ d\mu(x) \leq \int [x-t]_+ d\nu(x)$ for all t.

Problem 28. (Convex order II)

Let $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{a_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{b_i}$, where $\{a_i\}_i, \{b_i\}_i \subseteq \mathbb{R}$. Show that the following are equivalent:

(1) $\mu \leq_c \nu$.
(2) $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i$, and for each k, the sum of the k largest a's is less or equal than the sum of the k largest b's.

Hint: You may find (3) of Problem 27 useful.

Problem 29. (Convex order III)

Let θ be convex. Suppose $x_1 \leq x_2 \ldots_{\leq} x_n$ and $y_1 \leq y_2 \ldots_{\leq} y_n$. Using the characterization given in Problem 28 prove that

$$\sum_{i=1}^n \theta(x_i - y_i) = \inf_{\sigma \text{ permutation}} \sum_{i=1}^n \theta(x_i - y_{\sigma(i)}).$$

Problem 30. (Martingale Couplings)

Let $\mu, \nu \in \mathcal{P}_1(\mathcal{R})$ with $\mu \leq_c \nu$.

a) Prove that the product coupling $\mu \otimes \nu$ is a martingale coupling of μ and ν if and only if $\mu = \delta_x$ for some $x \in \mathbb{R}$.
b) Show that $\mathsf{MT}(\mu, \mu) = \{(id, id)_\# \mu\}$.

Hint: As a first step, show that for all convex $\varphi \in L^1(\mu)$ and for any version $(\pi_x)_{x\in\mathbb{R}}$ of the disintegration of $\pi \in \mathsf{MT}(\mu,\mu)$ w.r.t. the first component there holds $\int \varphi(y) d\pi_x(y) = \varphi(x)$ for μ-a.e. $x \in \mathbb{R}$.

Problem 31. (Quadratic Cost)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ with $\mu \leq_c \nu$. Show that any martingale coupling of μ and ν is a minimizer of the martingale optimal transport problem

$$\inf_{\pi \in \mathsf{MT}(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 d\pi(x,y).$$

Problem 32. (Potential Functions)

We define the functional $u : \mathcal{P}_1(\mathbb{R}) \to C(\mathbb{R}), \mu \mapsto u_\mu$ where the continuous function $u_\mu : \mathbb{R} \to \mathbb{R}$ is defined as

$$u_\mu(x) := \int_{\mathbb{R}} |y - x| d\mu(y).$$

a) Let $\mu \in \mathcal{P}_1(\mathbb{R})$ with mean $m = \int_{\mathbb{R}} y d\mu(y)$. Show that u_μ is a non-negative convex function that satisfies

$$\lim_{x \to \pm\infty} |u_\mu(x) - |x - m|| = 0.$$

b) Let $m \in \mathbb{R}$ and v a non-negative convex function that satisfies

$$\lim_{x \to \pm\infty} |v(x) - |x - m|| = 0.$$

Since v is convex, v is everywhere right-differentiable and the right-derivative v'_+ is monotonously increasing. For all $a < b$ in $\mathbb{R}$ we define

$$\mu((a, b]) := \frac{1}{2} \left(v'_+(b) - v'_+(a) \right).$$

Prove that $\mu \in \mathcal{P}_1(\mathbb{R})$ and that there holds $u_\mu = v$.

c) Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$. Show that $\mu \preceq_c \nu$ if and only if $u_\mu \leq u_\nu$.

Hint: You can use that $\mu \preceq_c \nu$ if the integral relation is satisfied for all convex functions $\varphi \in C^2(\mathbb{R})$ that satisfy $\varphi'' = 0$ outside of $[-n, n]$ for some $n \in \mathbb{N}$.

d) Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$. Show that $\mu = \nu$ if and only if $u_\mu = u_\nu$.

Problem 33. (Convex Envelope)

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function that is bounded from below by an affine function. The convex envelope of f is the function $\mathrm{conv}(f) : \mathbb{R} \to \mathbb{R}$ defined as

$$(\mathrm{conv}(f))(x) := \sup\{\varphi(x) : \varphi \text{ convex}, \varphi \leq f\}$$

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

a) Show that $\mathrm{conv}(f)$ is the greatest convex function below f.
b) Let g be a continuous function with $f \leq g$. Prove that $\mathrm{conv}(f) \leq \mathrm{conv}(g)$.
c) Let g be a convex function with $f \geq g$. Prove that $\mathrm{conv}(\mathrm{conv}(f) - g) = \mathrm{conv}(f - g)$.
d) Let $x_0 \in \mathbb{R}$ such that $f(x_0) > (\mathrm{conv}(f))(x_0)$. Show that $\mathrm{conv}(f)$ is locally affine at x_0, i.e. there exists $\varepsilon > 0$ and $a, b \in \mathbb{R}$ such that

$$\forall x \in [x_0 - \varepsilon, x_0 + \varepsilon] : (\mathrm{conv}(f))(x) = ax + b.$$

Problem 34. (Shadow)

Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$ with $\mu \preceq_c \nu$ and μ' a finite measure with $\mu' \leq_+ \mu$, i.e. $\mu - \mu'$ defines a finite measure on $\mathbb{R}$.

a) Let $v : \mathbb{R} \to \mathbb{R}$ be defined as

$$v := u_\nu - \mathrm{conv}(u_\nu - u_\mu).$$

Show that v is a non-negative and convex function with

$$\lim_{x \to \pm\infty} |v(x) - k|x - m|| = 0.$$

Hint: A function $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if for all $x_0 \in \mathbb{R}$ there exist $\varepsilon > 0$ and $a \in \mathbb{R}$ with $f(x) \geq ax + f(x_0)$ for all $x \in [x_0 - \varepsilon, x_0 + \varepsilon]$.

b) By a), there exists a finite measure η such that $u_\eta := u_\nu - \mathrm{conv}(u_\nu - u_{\mu'})$. Prove that the measure η satisfies
   (i) $\mu' \preceq_c \eta \leq_+ \nu$ and
   (ii) for all finite measures η' with $\mu' \preceq_c \eta' \leq_+ \nu$ there holds $\eta \preceq_c \eta'$.
   Hint: Recall Remark 2.18.

c) We denote the measure η in b) by $\mathcal{S}^\nu(\mu')$. Let μ'_1 and μ'_2 be two finite measures with $\mu'_1 + \mu'_2 = \mu'$. Show that

$$\mathcal{S}^\nu(\mu') = \mathcal{S}^\nu(\mu'_1) + \mathcal{S}^{\nu - \mathcal{S}^\nu(\mu'_1)}(\mu'_2).$$

Remark: The measure $\mathcal{S}^\nu(\mu')$ is uniquely determined by the properties (i) and (ii). It is called the shadow of μ in ν.

Problem 35. (Left-Monotone Coupling)

Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$ with $\mu \leq_c \nu$. For all $a \in \mathbb{R}$ and Borel sets $B \subseteq \mathbb{R}$ we set

$$\pi((-\infty, a] \times B) := (\mathcal{S}^\nu(\mu_{|(-\infty,a]}))(B) \tag{2.11}$$

where $\mu_{|(-\infty,a]}$ denotes the restriction of μ onto $(-\infty, a]$ and the shadow on the r.h.s. is defined as in Problem 34.

a) Show that (2.11) uniquely defines a probability measure on $\mathbb{R}^2$ and that π is a martingale coupling of μ and ν.

b) Prove that π is left-monotone, i.e. there exists a Borel set $\Gamma \subseteq \mathbb{R}^2$ with $\pi(\Gamma) = 1$ such that for all pairs $(x, y^-), (x, y^+), (x', y') \in \Gamma$ there holds

$$x < x', y^- < y^+ \Rightarrow y' \notin (y^-, y^+).$$

Hint: Let φ be the density of the standard normal distribution. The function c defined by $c(x, y) := -\int_{\mathbb{R}} 1_{(-\infty,x]}(u)\varphi(u)\sqrt{1+y^2}du$ satisfies $c_{xyy} < 0$ and $|c(x, y)| \leq 2 + |y|$.

Problem 36. (Superhedging of the supremum squared)

Show that for all $s_1, \ldots, s_n \in \mathbb{R}$ we have

$$\left[\max_{i=1,\ldots,n} s_i\right]^2 \leq 4s_n^2 + \sum_{k=1}^{n-1} h_k[s_{k+1} - s_k],$$

with $h_k = -4 \max_{i=1,\ldots,k} s_i$.

Problem 37. (Lipschitz martingale kernels)

Let $q \in \mathsf{MT}(\mu, \nu)$ and recall that q_x stands for the conditional distribution of the second variable given that the first is equal to x. Prove the equivalence of the following statements:

(1) $\forall x, \bar{x} :\ W_1(q_x, q_{\bar{x}}) \leq |x - \bar{x}|$;
(2) $\forall x, \bar{x} :\ W_1(q_x, q_{\bar{x}}) = |x - \bar{x}|$;
(3) $\forall x \leq \bar{x}, \forall a :\ q_x((-\infty, a]) \geq q_{\bar{x}}((-\infty, a])$.

We say that q has Lipschitz martingale kernels if these conditions hold.

Problem 38. (Standatrd stretched Brownian motion)

Let M be a standard stretched Brownian motion and define $q = Law(M_0, M_1)$. Prove that q has Lipschitz martingale kernels. (The same is valid for $q = Law(M_s, M_t)$ if $s \leq t$.)

3. The Skorokhod embedding problem

In Subsection 2.1 we saw that the martingale optimal transport problem is closely connected to finding model independent bounds on option prices. In Subsection 2.3 we have looked at continuous time versions of the MOT problem. In this section, we consider another possibility to obtain worst case bounds for option prices which however only works for time-invariant derivatives, derivatives whose payoff is invariant under time-changes. The key observation by David Hobson is that for these options the questions of worst case bounds naturally relates to the Skorokhod embedding problem.

3.1. **Motivation.** Let us assume that we are in the following situation:

- stock prices evolve continuously in continuous time given by a continuous stochastic process $(S_t)_{t\geq 0}$
- European call options with maturity at time 1 are liquidly traded with all possible strikes so that the time-1-marginal is known to be $\mu_1 =: \nu$, i.e. $S_1 \sim \nu$.

Example 3.1 (One-touch digital option, due to David Hobson [Hob98]). Denote the running maximum of the stock price process by

$$\bar{S}_t := \max_{0\leq s\leq t} S_s$$

and the first hitting time of the level L by

$$H_L := \inf\{t \geq 0 : S_t \geq L\}.$$

Consider an option with payoff

$$G(S) := \mathbb{1}_{\{S_t \geq L \text{ for some } t\in[0,1]\}} = \mathbb{1}_{\{\bar{S}_1 \geq L\}} = \mathbb{1}_{\{H_L \leq 1\}}.$$

We are interested in price bounds for G, i.e.

$$\sup_{\mathbb{Q} \text{ mg. meas. : } S_1 \sim_{\mathbb{Q}} \nu} \mathbb{E}_{\mathbb{Q}}[G(S)].$$

Observe that for any $K < L$, there holds

$$\mathbb{1}_{\{H_L\leq 1\}} \leq \frac{(S_1 - K)_+}{L - K} + \frac{(S_{H_L} - S_1)}{L - K} \mathbb{1}_{\{H_L\leq 1\}}. \tag{3.1}$$

The second term on the RHS of (3.1) are gains from trading so that $\mathsf{price}\left(\frac{(S_{H_L}-S_1)}{L-K}\mathbb{1}_{\{H_L\leq 1\}}\right) = 0$. In particular, we get for any admissable martingale measure $\mathbb{Q}$

$$\Rightarrow \mathbb{E}_{\mathbb{Q}}[G(S)] \leq \frac{\mathbb{E}_{\mathbb{Q}}[(S_1 - K)_+]}{L - K}.$$

Since, the LHS does not depend on $K < L$ we immediately obtain for any admissable martingale measure $\mathbb{Q}$ (recalling the notation $C_{t,K}$ from Subsection 2.1)

$$\mathbb{E}_{\mathbb{Q}}[G(S)] \leq \inf_{K<L} \frac{\mathbb{E}_{\mathbb{Q}}[(S_1 - K)_+]}{L - K} = \inf_{K<L} \frac{\mathsf{price}(C_{1,K})}{L - K}. \tag{3.2}$$

Question: Is (3.2) best possible? Otherwise said, is (3.1) the optimal superhedging strategy?

To answer this question in the positive we need to do accomplish the following
Exercise: Find a martingale measure $\mathbb{Q}$, such that $S_1 \sim_{\mathbb{Q}} \nu$ and

$$\mathbb{E}_{\mathbb{Q}}[G(S)] = \inf_{K<L} \frac{\mathsf{price}(C_{1,K})}{L - K}.$$

Key observation: If $(S_t)_{t\in[0,1]}$ is a continuous martingale, by the Theorem of Dambins-Dubins-Schwarz, there is a time change $(\tau_t)_{t\in[0,1]}$ of a Brownian Motion B such that $S_t = B_{\tau_t}$.

In particular, τ_1 is a stopping time of Brownian Motion, such that $B_{\tau_1} \sim \nu$. This means that τ_1 is a solution to the Skorokhod embedding problem (SEP):

> Given $\nu \in \mathcal{P}(\mathbb{R})$, $\int |x| d\nu < \infty$, $\int x d\nu = 0$ find a stopping time τ of Brownian motion B such that $B_\tau \sim \nu$ and $(B_{t\wedge\tau})_{t\geq 0}$ is uniformly integrable.

If τ is a solution to (SEP) we write τ solves SEP(ν). Here we implicitly assume that $B_0 = 0$, one could also consider SEP(μ, ν) for the variant where the Brownian motion starts in a random variable $B_0 \sim \mu \preceq_c \nu$.

Remark 3.2.
- The u.i condition is equivalent to requiring that τ is minimal, i.e. if ρ is another stopping time such that $B_\rho \sim \nu$ and $\rho \leq \tau$ then $\rho = \tau$.
- The zero-mean assumption of ν is there so that we can work with Brownian motion starting in 0. The problem is of course invariant under shifts by a constant m to account for measures ν with mean m.

Let us continue with Example 3.1:

Given a solution τ to SEP(ν). Then $S_t := B_{\frac{t}{1-t}\wedge\tau}$ is a continuous martingale with $S_1 \sim \nu$. Moreover, $\{H_L \leq 1\} = \{\bar{S}_1 \geq L\} = \{\bar{B}_\tau \geq L\}$ is invariant under time changes! Thus, finding a continuous martingale S with $S_1 \sim \nu$ attaining the bound (3.2) is equivalent to finding a solution to SEP(ν) attaining this bound. Let us try to do the latter so that we consider

$$\sup_{\tau \text{ solves SEP}(\nu)} \mathbb{E}[G(B_{\cdot\wedge\tau})]. \tag{3.3}$$

To this end, we need to find structural constraints satisfied by any solution to (SEP) maximizing the left-hand side of (3.3):

For any stopping time τ such that $(B_{t\wedge\tau})_{t\geq 0}$ is u.i. we have using the strong Markov property in the first step

$$\begin{aligned} 0 = \mathbb{E}[(B_\tau - L)\mathbb{1}_{\bar{B}_\tau \geq L}] &= \mathbb{E}(B_\tau - K)\mathbb{1}_{\bar{B}_\tau \geq L} + (K-L)\mathbb{P}(\bar{B}_\tau \geq L) \\ &\leq \mathbb{E}[(B_\tau - K)\mathbb{1}_{B_\tau \geq K}] + (K-L)\mathbb{P}(\bar{B}_\tau \geq L). \end{aligned}$$

Consequently, we have denoting $C_\nu(K) = \int (x-K)_+ \nu(dx)$

$$\mathbb{P}(\bar{B}_\tau \geq L) \leq \inf_{K<L} \frac{\mathbb{E}[(B_\tau - K)_+]}{L-K} \overset{\text{if } B_\tau \sim \nu}{=} \inf_{K<L} \frac{C_\nu(K)}{L-K}$$

with equality if and only if

$$\{B_\tau > K\} \subseteq \{\bar{B}_\tau \geq L\} \subseteq \{B_\tau \geq K\} \tag{3.4}$$

Observe that (3.4) poses additional constraints on solutions to (SEP). These constraints are satisfied by a special solution to (SEP) found by Azéma and Yor: Define the barycentre function

$$\psi_\nu(x) = \begin{cases} \frac{1}{\nu([x,\infty))} \int_{[x,\infty)} y d\nu(dy) & , \text{ if } x < \inf\{y : \nu[y,\infty) = 0\} \\ x & , \text{ if } \nu([x,\infty)) = 0 \end{cases}$$

Then $\psi_\nu(x) \geq x$, ψ_ν is non-decreasing and left continuous and it can be shown that

$$\tau_{AY} := \inf\{t \geq 0 : B_t \leq \psi_\nu^{-1}(\bar{B}_t)\}$$

is a solution to SEP(ν). In particular, (3.4) is satisfied, we cannot improve on (3.4) and we have:

Proposition 3.3. *The Azéma-Yor solution to (SEP) maximizes $\mathbb{P}(\bar{B}_\tau \geq L)$ simultanuously for all L among all (minimal) solutions to (SEP). Hence, it also maximizes $\mathbb{E}[f(\bar{B}_\tau)]$ for all increasing $f : \mathbb{R}_+ \to \mathbb{R}$.*

Remark 3.4. There are 20+ further solutions to SEP(ν), most have some kind of optimality property, many are related to robust finance via the "time-change" method of David Hobson, e.g. the Root and Rost embedding which will be prominent in the next section. Notably, most of these solutions use different techniques and methods. The goal of the

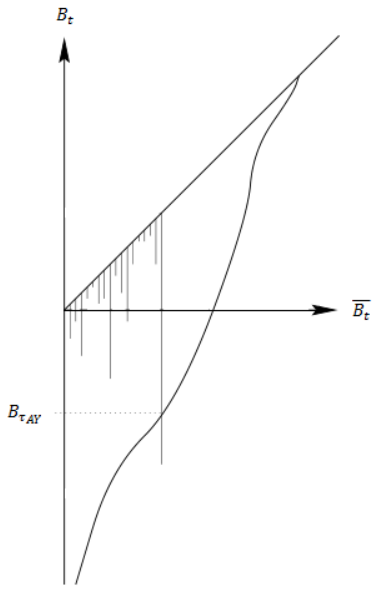


FIGURE 2. The Azema-Yor solution as a first hitting time of a "barrier" in the $\bar{B} - B$ phase space.

following section is to understand the structure of these "extremal" solutions and to find a unifying construction which also allows us to come up with new solutions with pre-specified optimality properties (which then directly relate back to worst case bounds in model-independent finance).

3.2. **Mass transport approach to Skorokhod embedding.** Throughout this section we fix a measure $\nu \in \mathcal{P}(\mathbb{R})$ satisfying $V := \int x^2 d\nu < \infty$ and $\int x d\nu = 0$. The second moment assumption on ν is not necessary but reduces technicalities. The goal of this section is to introduce and analyse an *optimal SEP* which will allow us to construct solutions to SEP(ν) with prescribed optimality properties in a systematic fashion.

In this section we will consider two probability spaces:

- The Wiener space $(C_0(\mathbb{R}_+), \mathbb{W})$ of functions starting in 0 equipped with the natural filtration $(\mathcal{F}_t)_t$ as well as the augmented filtration $\mathcal{F}^a$ (all null sets of $\mathbb{W}$ are included in $\mathcal{F}_0^a$). The topology is given by uniform convergence on compact sets.
- Its extension $\overline{C_0(\mathbb{R}_+)} = C_0(\mathbb{R}_+) \times [0,1]$, $\overline{\mathbb{W}} = \mathbb{W} \otimes Leb|_{[0,1]}$ equipped with $\overline{\mathcal{F}}$, the augmentation of $(\mathcal{F}_t \otimes \mathcal{B}([0,1]))_{t\geq 0}$.

We consider the set of stopped path

$$S := \{(f,s) : f \in C_0[0,s], f(0) = 0\}$$

together with a "cost funtion"

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

We note that S is a Polish space. One choice for a metric is given by (wlog $s < t$)

$$d((f,s),(g,t)) = \max\{\sup_{u\leq s} |f(u) - g(u)|, |s-t|, |f(s) - g(t)|\}.$$

We also define the projection or restriction map

$$r : C_0(\mathbb{R}_+) \times \mathbb{R}_+ \to S \qquad (\omega, t) \mapsto (\omega_{|[0,t]}, t).$$

Let us introduce a preliminary version of the optimal SEP:

$$\inf\{\mathbb{E}[\gamma \circ r((B_s)_{s\leq\tau}, \tau) : \tau \text{ solves SEP}(\nu)\} \qquad (\mathsf{OptSEP}^{\mathsf{M}})$$

Key intuition: SEP/($\mathsf{OptSEP}^{\mathsf{M}}$) is similar to a transport problem. More precisely the ($\mathsf{OptSEP}^{\mathsf{M}}$) is a Monge-type problem (cf. (MP)). Indeed, $\omega \mapsto \tau(\omega)$ denotes a single time to stop. Hence, τ indicates where in time and, hence, in space-time the path ω places its proportion of mass of the Wiener measure $\mathbb{W}$ (for the Wiener measure this should not be taken literally; however in case of a random walk it should).

To stress this point of view, we associate to a stopping time τ on $(C_0(\mathbb{R}_+), \mathbb{W})$ the measure

$$\bar{\tau}(d\omega, dt) := \delta_{\tau(\omega)}(dt)\mathbb{W}(d\omega).$$

Following our intuition we should think about this measure as a kind of coupling between the Wiener measure $\mathbb{W}$ and the measure ν we want to embed (of course it is not a true coupling). Then, just as in optimal transport, we should relax this problem to allow for "general couplings", that is

$$\bar{\tau}(d\omega, dt) = \bar{\tau}_\omega(dt)\mathbb{W}(d\omega) \qquad \bar{\tau}_\omega \in \mathcal{P}(\mathbb{R}_+).$$

To keep the stopping time nature of these measure we need to encode the optionality (in case of stopping times, the property that $\{\tau \leq t\} \in \mathcal{F}_t\}$ for all $t \geq 0$).

Definition 3.5. *Set* $\mathsf{M} := \{\xi \in \mathcal{P}(C_0(\mathbb{R}_+) \times \mathbb{R}_+) : \text{proj}_{C_0(\mathbb{R}_+)}(\xi) = \mathbb{W}\}$. *A measure* $\xi \in \mathsf{M}$ *is called randomized stopping time, short* $\xi \in \mathsf{RST}$ *iff on* $\overline{C_0(\mathbb{R}_+)}$ *the random time*

$$\rho(\omega, u) := \inf\{t \geq 0 : \xi_\omega([0, t]) \geq u\}$$

defines an $\overline{\mathcal{F}}$*-stopping time.*

The following lemma shows that there is no loss in generality when restricting to a special probability space in Definition 3.5.

Lemma 3.6. *Let B be a Brownian Motion on some stochastic base* $(\Omega, \mathcal{G}, (\mathcal{G}_t)_{t\geq 0}, \mathbb{P})$. *Let* τ *be a* $\mathcal{G}$*-stopping time and consider*

$$\varphi : \Omega \to C_0(\mathbb{R}_+ \times \mathbb{R}_+) \qquad , \qquad \omega \mapsto ((B_t(\bar{\omega}))_{t\geq 0}, \tau(\bar{\omega})).$$

Then $\xi := \varphi(\mathbb{P}) \in RST$ *and for every* $\gamma : S \to \mathbb{R}$ *we have*

$$\int \gamma(f, s) r(\xi)(d(f, s)) = \mathbb{E}_{\mathbb{P}}[\gamma((B_t)_{t\geq 0}, \tau)]. \tag{3.5}$$

If Ω *is sufficiently rich such that it supports a uniformly distributed random variable which is* $\mathcal{G}_0$*-measurable, then for any* $\xi \in RST$ *we can find a* $\mathcal{G}$*-stopping time* τ *on* Ω *such that* $\xi = \varphi(\mathbb{P})$ *and (3.5) holds.*

Proof. Clearly $\xi := \varphi(\mathbb{P}) \in \mathsf{M}$. Write $(\xi_\omega)_{\omega \in C_0(\mathbb{R}_+)}$ for a disintegration wrt Wiener measure. We need to show that $\xi_\omega([0, t])$ is $\mathcal{F}_t^a$-measurable. Let $g : C(\mathbb{R}_+) \to \mathbb{R}$ be a measurable function. If $h = \mathbb{E}_{\mathbb{W}}[g|\mathcal{F}_t^a]$, writing $\mathcal{G}_t^a$ for the usual augmentation $\mathcal{G}$, and nothing that $(B_t)_{t\geq 0}$ is also a $\mathcal{G}^a$-Brownian motion, we have

$$\mathbb{E}_{\mathbb{P}}[g((B_s)_{s\geq 0})|\mathcal{G}_t^a] = h((B_r)_{r\geq 0}), \qquad \mathbb{P} - a.s.$$

It then follows that

$$\begin{aligned}
\int g(\omega)\xi_\omega([0, t])\mathbb{W}(d\omega) &= \mathbb{E}_{\mathbb{P}}[g((B_r)_{r\geq 0})\mathbb{1}_{\{\tau\leq t\}}] \\
&= \mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathbb{P}}[g((B_r)_{r\geq 0})|\mathcal{G}_t^a]\mathbb{1}_{\{\tau\leq t\}}] \\
&= \mathbb{E}_{\mathbb{P}}[h((B_r)_{r\geq 0})\mathbb{1}_{\{\tau\leq t\}}] \\
&= \int h(\omega)\xi_\omega([0, t])\mathbb{W}(d\omega).
\end{aligned}$$

Hence $\xi_\omega([0,t])$ is $\mathcal{F}_t^a$-measurable as required.
To prove the second part, we observe that by Definition 3.5, there exists an $\bar{\mathcal{F}}$-stopping time ρ' representing ξ. Since ρ' is $\bar{\mathcal{F}}$-predictable, it follows that there exists an almost surely equal $(\mathcal{F}_t^0 \times \mathcal{B}([0,1]))_{t\geq 0}$-stopping time ρ. Then we can define a random time on Ω by $\rho((B_s)_{s\geq 0}, U)$, where B is the Brownian motion, and U the independent $\mathcal{G}_0$-measurable, uniform random variable. Consider the map

$$\bar{\varphi} : \Omega \to \overline{C}_0(\mathbb{R}_+)\ \bar{\omega} \mapsto ((B_t(\bar{\omega})), Y(\bar{\omega})).$$

Since ρ is a $(\mathcal{F}_t^0 \times \mathcal{B}([0,1]))_{t\geq 0}$-stopping time and $\bar{\varphi}$ is measurable from $(\Omega, \mathcal{G}_t)$ to $(\overline{C}_0(\mathbb{R}_+), \mathcal{F}_t^0 \times \mathcal{B}([0,1])), \rho \circ (B,Y)$ is a $\mathcal{G}$-stopping time. □

We aim to show that the set of randomized stopping times embedding a given measure ν is compact. We first show that RST is closed. We will use that a random variable Z is $\mathcal{F}_t^a$-measurable iff for any measurable and bounded g there holds that $\mathbb{E}[Z(g - \mathbb{E}[g|\mathcal{F}_t^a]] = 0$.

Lemma 3.7. *$\xi \in \mathsf{M}$ is a randomized stopping time iff for all $f \in C_b(\mathbb{R}_+)$ supported on $[0,t]$ for some $t \geq 0$ and all $g \in C_b(C_0(\mathbb{R}_+))$*

$$\int f(s)(g - \mathbb{E}[g|\mathcal{F}_t^a])(\omega)\ \xi(d\omega, ds) = 0. \tag{3.6}$$

Proof. Let $\xi \in \mathsf{RST}$ with representation ρ as in Definition 3.5. Then the left-hand side of (3.6) equals

$$\int \int f(\rho(\omega,u))du(g - \mathbb{E}[g|\mathcal{F}_t^a])(\omega)\mathbb{W}(d\omega),$$

which equals zero since $\omega \mapsto \int f(\rho(\omega,u))du$ is $\mathcal{F}_t^a$-measurable by the optionality of ρ.

Conversely, we need to show that $\{\rho(\omega,u) \leq t\} \in \overline{\mathcal{F}}_t$ for all $t \geq 0$ which holds iff $\omega \mapsto A_t(\omega) := \int \mathbb{1}_{[0,t]}(\rho(\omega,u))du$ is $\mathcal{F}_t^a$-measurable (since $\{(\omega,u) : \rho(\omega,u) \leq t\} = \{(\omega,u) : u \leq A_t(\omega)\}$) which follows from (3.6) by a monotone class argument (so that we can test with $f = \mathbb{1}_{[0,t]}$). □

Corollary 3.8. RST *is closed w.r.t. the weak topology induced by $C_b(C_0(\mathbb{R}_+) \times \mathbb{R}_+)$.*

Proof. This follows from (3.6) upon choosing a version of $\mathbb{E}[g|\mathcal{F}_t^a]$ which is continuous and bounded. A possible choice is (continuity then follows by dominated convergence):

$$\int g(\omega_{|[0,t]} \otimes \tilde{\omega})\mathbb{W}(d\tilde{\omega}),$$

where $\otimes$ denotes concatentation of paths defined via

$$\omega_{|[0,t]} \otimes \tilde{\omega}(s) = \begin{cases} \omega(s) & s \leq t \\ \omega(t) + \tilde{\omega}(s-t) & s > t \end{cases}.$$

□

For $\xi \in RST$ and optional $Y = Y(\omega,t) = Y_t(\omega)$ ($(\omega,t) \mapsto Y_t(\omega)$ is measurable w.r.t. the optional σ-algebra which is generated by the right-continuous adapted processes) which is bounded or positive we define Y_ξ as $Y(\xi)$. Let ρ be the representation of ξ as in Definition 3.5 and write $\bar{Y}_t(\omega,u) = Y_t(\omega)$ for the extended process. Then we have $Y_\xi \sim \bar{Y}_\rho$. Specializing to $Y_t = t =: T(\omega,t)$ we obtain that

$$\mathbb{E}_{\bar{\mathbb{W}}}[\rho] = \int T(\omega,t)d\xi(\omega,t),$$

where $T : C_0(\mathbb{R}_+) \times \mathbb{R}_+ \to \mathbb{R}_+$ denotes the projection $(\omega,t) \mapsto t$. Recall that we assumed $\int x^2 d\nu =: V < \infty$.

Lemma 3.9. *Let $\xi \in \mathsf{RST}$ with representation ρ and projection map T as above. Assume that $B_\xi = \nu$, i.e. $\bar{B}_\rho \sim \nu$. Then, the following are equivalent:*

(1) $\int T d\xi = \mathbb{E}_{\bar{\mathbb{W}}}[\rho] < \infty$

(2) $\int T d\xi = \mathbb{E}_{\overline{\mathbb{W}}}[\rho] = V$
(3) $(\bar{B}_{\rho\wedge t})_{t\geq 0}$ *is u.i.*

Proof. This follows from optional stopping applied to $(B_t^2 - t)_{t\geq 0}$. □

Definition 3.10. *We define the set of randomized stopping times embedding a given measure* ν *by* $\mathsf{RST}(\nu) := \{\xi \in \mathsf{RST} : B_\xi = \nu, \int T d\xi = V\}$.

Proposition 3.11. $\mathsf{RST}(\nu) \neq \emptyset$ *and compact w.r.t. the weak topology induced by* $C_b(C_0(\mathbb{R}_+)\times \mathbb{R}_+)$.

Proof. $\mathsf{RST}(\nu) \neq \emptyset$ follows since the Azema-Yor embedding is in the set $\mathsf{RST}(\nu)$. Since

$$B_\xi = \nu \quad \Leftrightarrow \quad \int f(\omega(t))\xi(d\omega, dt) = \int f(x)d\nu(x)$$

for all $f \in C_b(\mathbb{R})$ the set $\mathsf{RST}(\nu)$ is a closed subset of the closed set RST and it therefore suffices to show tightness. Since $\mathrm{proj}_{C_0(\mathbb{R}_+)}(\xi) = \mathbb{W}$ for each $\xi \in \mathsf{RST}(\nu)$ tightness follows from Markov's inequality and Lemma 3.9 via

$$\xi(t \geq L) \leq \frac{1}{L}\int T d\xi = \frac{V}{L}$$

which can be made arbitrarily small by choosing L big enough. □

Finally we can state our relaxed optimization problem:

$$P_\gamma^{\mathsf{SEP}} := \inf_{\xi\in\mathsf{RST}(\nu)} \int \gamma \circ r \, d\xi \tag{OptSEP}$$

We remark that – just as (KP) and (MOT) – this is a linear optimization over a compact and convex set. The following corollary is immediate.

Corollary 3.12. *If* $\gamma : S \to \mathbb{R}$ *is lower semicontinuous and bounded from below there exists an optimizer.*

Remark 3.13.
- Both conditions on γ in Corollary 3.12 can be relaxed, e.g. it is sufficient to require γ to be bounded from below in the sense that
$$\gamma(f, s) \geq -a - b \cdot s$$
for some constants $a, b > 0$ (since γ and $\tilde{\gamma} := \gamma + a + b \cdot s \geq 0$ have the same optimizers by Lemma 3.9).
- Using min-max arguments one can prove a duality theory for (OptSEP). Since there is no direct route to a monotonicity principle nor a direct way to interprete it as a subhedging result (however both is possible) we omit the statement.

Let us now turn to the geometry of optimizers. To this end, we start with the following definition which you should compare with cyclical monotonicity for optimal transport.

Definition 3.14. *Let* $\gamma : S \to \mathbb{R}$ *be Borel-measurable.*

(1) $((f, s), (g, t)) \in S \times S$ *is called* stop-go-pair*, short* $(f, g) \in \mathsf{SG}$*, if* $f(s) = g(t)$ *and for all stopping times* σ *of Brownian Motion* B *with* $0 < \mathbb{E}[\sigma] < \infty$ *(cf. Figure 3)*

$$\mathbb{E}[\gamma(f \otimes (B_u)_{u\leq\sigma}, s + \sigma)] + \gamma(g, t) > \gamma(f, s) + \mathbb{E}[\gamma(g \otimes (B_u)_{u\leq\sigma}, t + \sigma)], \tag{3.7}$$

where $\otimes$ *denotes concatenation of paths as before.*

(2) A set $\Gamma \subseteq S$ *is called* $\gamma-$monotone *if*

$$\mathsf{SG} \cap (\Gamma^< \times \Gamma) = \emptyset,$$

where $\Gamma^< := \{(f, s) \in S : \exists (g, t) \in \Gamma, t > s, g \equiv f \text{ on } [0, s]\}$ *is the set of not-yet stopped paths.*

Theorem 3.15 (monotonicity principle for SEP)**.** *Let* $\gamma : S \to \mathbb{R}$ *be Borel and* ξ *an optimizer to* (OptSEP) *with representation* ρ *as before. Then, there exists a* $\gamma-$*monotone set* $\Gamma \subseteq S$ *such that*

$$\bar{\mathbb{W}}[((\bar{B}_s)_{s\leq\rho}, \rho) \in \Gamma] = 1. \tag{3.8}$$

A sloppy way to read this theorem is that one cannot improve an optimizer by *pathwise* modifications as given by (3.7), i.e. for any optimizer there is a set Γ on which it is concentrated in the sense of (3.8) s.t. $\mathsf{SG} \cap (\Gamma^< \times \Gamma) = \emptyset$.

A proof of this result is outside the scope of this lecture. It relies on the dual theory to (OptSEP) in combination with Choquet's capacibility theorem together with a variant of Lemma 1.18 adapted to the current setup (which requires some care due to the optionality requirement for randomized stopping times).

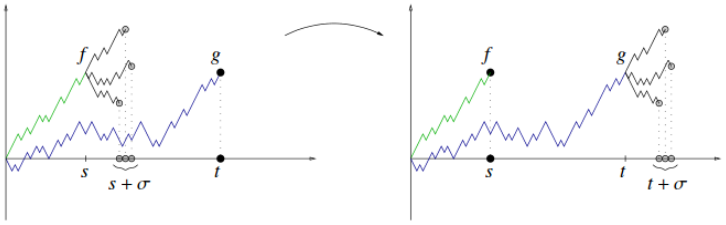


FIGURE 3. The left hand side of (3.7) corresponds to averaging the function γ over the stopped paths on the left picture; the right hand side to averaging the function γ over the stopped paths on the right picture.

As an application of this monotonicity principle we want to derive the Root embedding. Similarly to the solution of Azéma and Yor the solution of Root is connected to model independent finance. It is the extremal model for options on variance where the payoff depends on the realized quadratic variation of the logarithm of the price process (e.g. see [CW13]).

Theorem 3.16. *Let $h : \mathbb{R}_+ \to \mathbb{R}$ be strictly convex, $\gamma(\omega, t) = h(t)$ and assume that $P_\gamma^{\mathsf{SEP}} < \infty$. Then a minimizer of* (OptSEP) *exists and, moreover, for any minimizer $\hat{\tau}$, there exists a barrier R such that*

$$\hat{\tau} = \inf\{t \geq 0 : (t, B_t) \in R\}.$$

In particular, the minimizer $\hat{\tau}$ is unique and a non-randomized stopping time.

Proof. **Step 1.** We first pick- by Corollary 3.12- a stopping time $\hat{\tau}$ which attains P_γ^{SEP}. By Theorem 3.15 there exists a set $\Gamma \subseteq S$ such that $((B_s)_{s \leq \hat{\tau}}, \hat{\tau}) \in \Gamma$ almost surely, and such that $(\Gamma^< \times \Gamma) \cap SG = \emptyset$.

Step 2. Next, consider paths $(f, s), (g, t) \in S$ such that $f(s) = g(t)$. We want to understand when $((f, s), (g, t)) \in \mathrm{SG}$, i.e., under which conditions (f, s) should be stopped and Brownian motion should continue to go after (g, t). In the present case (3.7) amounts to

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

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

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

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

We claim that the hitting times of $\mathcal{R}_{CL}$ and $\mathcal{R}_{OP}$ are ordered and sandwich $\hat{\tau}$.

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

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

Step 3. Now consider $(\omega, u) \in C_0(\mathbb{R}_+)\times[0,1]$ such that $(g,t) = ((\bar{B}_s(\omega,u))_{s\leq\hat{\tau}(\omega,u)}, \hat{\tau}(\omega,u)) \in \Gamma$. Then it follows immediately that:

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

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

Step 4. Uniqueness follows as in the previous section since we have showed that any minimizer is a true stopping time and the hitting time of a barrier (cf. Remark 1.39). □

Remark 3.17. (1) The barrier in Theorem 3.16 is unique in the following sense. Call the optimizer from Theorem 3.16 τ_{Root}. If S is another barrier such that

$$\tau' = \inf\{t \geq 0 : (t, B_t) \in S\}$$

solves SEP(ν) then a.s.

$$\tau' = \tau_{Root}.$$

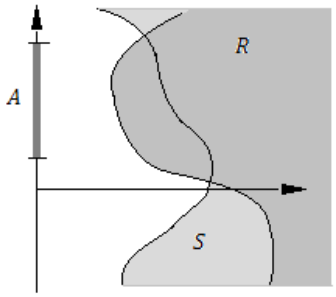
The argument due to Loynes goes as follows:
Suppose R and S are barriers both inducing (minimal) solutions to SEP(ν). W.l.o.g. we can assume R and S are closed. Consider the barrier $R \cup S$ with hitting time $\tau_{R\cup S}$. Let

$$A \subseteq \Omega_R := \{x : (t,x) \in S \Rightarrow (t,x) \in R\}$$

Then

$$\mathbb{P}(B_{\tau_{R\cup S}} \in A) \leq \mathbb{P}(B_{\tau_R} \in A) = \nu(A).$$

Similarly, for



$$A' \subseteq \Omega_S := \{x : (t,x) \in R \Rightarrow (t,x) \in S\}$$

we have

$$\mathbb{P}(B_{\tau_{R\cup S}} \in A') \leq \mathbb{P}(B_{\tau_S} \in A') = \nu(A').$$

Since $\nu(\Omega_R \cup \Omega_S) = 1$, $\tau_{R\cup S}$ embeds ν. By minimality (cf. Remark 3.2(i)) of τ_R and τ_S this implies $\tau_R = \tau_S$ a.s.

(2) Theorem 3.16 is not constructive in the sense that there is no closed expression for the barrier. However, there are PDE characterizations for the barrier, e.g. [CW13].

(3) Requiring the function h in Theorem 3.16 to be concave we can run (essentially) the same argument to show Rost's result, the existence of an inverse barrier $R \subseteq \mathbb{R}_+ \times \mathbb{R}$ ($(s,x) \in R, t \leq s \Rightarrow (t,x) \in R$) such that the first hitting time of R solves SEP(ν). However, here one needs to be careful at time zero. Again the barrier is unique by Loynes's argument.

(4) All known solutions (and many more) to (OptSEP) can be derived by essentially the same argument up to running a secondary optimization problem among all optimizers (which are usually a compact set) to (OptSEP) (see below).

(5) Considering $\gamma(f,s) = c(f(0), f(s))$ for a Spence-Mirrlees cost function c allows to recover the left-monotone solution to (MOT). In fact, all known one-dimensional solutions to (MOT) can be derived in this way, see [HS18]
(6) There is a variant of the monotonicity principle Theorem 3.15 which is necessary and sufficient. However, its precise relation to Theorem 3.15 is not clear, see [BNS19].

3.2.1. *Secondary optimization and the Azéma-Yor embedding.* In this section we want to shortly indicate what we mean with secondary optimization and how to use this to derive the Azéma-Yor embedding.

For $\gamma : S \to \mathbb{R}$ write Opt_γ for the set of optimizers to (OptSEP). Consider another function $\tilde{\gamma} : S \to \mathbb{R}$. We call $\hat{\tau} \in \mathsf{Opt}_\gamma$ a secondary optimizer/ minimizer if it solves

$$P_{\tilde{\gamma}|\gamma} = \inf_{\tau \in \mathsf{Opt}_\gamma} \mathbb{E}[\tilde{\gamma}((B_s)_{s\le\tau}, \tau)]. \tag{OptSEP$_2$}$$

Theorem 3.18. *Let $\gamma, \tilde{\gamma} : S \to \mathbb{R}$ be lower semi continuous and bounded from below then* (OptSEP$_2$) *admits a minimizer.*

Definition 3.19. (i) $((f,s),(g,t)) \in S \times S$ *is a* secondary stop-go-pair*, written* $(f,g) \in \mathsf{SG}_2$*, if and only if* $f(s) = g(t)$ *and for any stopping time* σ *of Brownian Motion* B *satisfying* $0 < \mathbb{E}[\sigma] < \infty$

$$\gamma(f,s) + \mathbb{E}[\gamma(g \oplus (B_u)_{u\le\sigma}, t+\sigma)] \le \mathbb{E}[\gamma(f \oplus (B_u)_{u\le\sigma}, s+\sigma)] + \gamma(g,t), \tag{3.9}$$

and if "=" holds in (3.9) then

$$\tilde{\gamma}(f,s) + \mathbb{E}[\tilde{\gamma}(g \oplus (B_u)_{u\le\sigma}, t+\sigma)] < \mathbb{E}[\tilde{\gamma}(f \oplus (B_u)_{u\le\sigma}, s+\sigma)] + \tilde{\gamma}(g,t). \tag{3.10}$$

(ii) $\Gamma \subseteq S$ *is called* $\tilde{\gamma}|\gamma$-monotone *if* $\mathsf{SG}_2 \cap (\Gamma^< \times \Gamma) = \emptyset$.

Theorem 3.20. *Let $\gamma, \tilde{\gamma} : S \to \mathbb{R}$ be Borel. Suppose $P_{\tilde{\gamma}|\gamma} < \infty$ and that ξ is an optimizer with representation ρ on $\bar{C}_0(\mathbb{R}_+)$. Then there exists a $\tilde{\gamma}|\gamma$- monotone set $\Gamma \subseteq S$ such that*

$$\bar{\mathbb{W}}[((B_s)_{s\le\rho}, \rho) \in \Gamma] = 1.$$

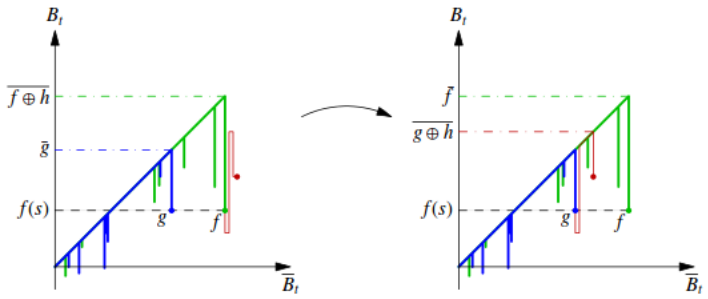


FIGURE 4. The stop-go pairs for the Azema-Yor embedding. On the left,the blue path (g,t) is stopped, and the green path, (f,s), is allowed to continue; a possible continuation, h, being shown in red. On the right hand side we see the effect of allowing g to go and stopping f: the maximum of g is increased, but the maximum of f stays the same.

Example 3.21. Azema-Yor embedding:
Pick

$$\gamma(f,s) = -\bar{f} = -\max_{0\le u\le s} f(u)$$

and

$$\tilde{\gamma}(f,s) = \varphi(\bar{f}),$$

for some $\varphi : \mathbb{R}_+ \to \mathbb{R}$ bounded, strongly increasing and continuous. Then, it is possible to show that (Exercise!)

$$\{((f,s),(g,t)) \in S \times S, f(s) = g(t), \bar{g} < \bar{f}\} \subseteq SG_2.$$

From here the argument follows closely the proof of Theorem 3.16.We omit the details (cf. Exercises).

3.3. **Exercises.**

Problem 39. (The Second Constraint)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ with $\mu \preceq_c \nu$. Moreover, let B be a Brownian Motion and τ a $\mathcal{F}^B$-stopping time such that $B_\tau \sim \nu$.

a) Prove that $\mathbb{E}[\tau] < +\infty$ if and only if $(B_{t\wedge\tau})_{t\geq 0}$ is uniformly integrable, i.e. for all $\varepsilon > 0$ it exists $K \in \mathbb{N}$ such that $\sup_{t\geq 0} \mathbb{E}[|B_{\tau\wedge t}|\mathbb{1}_{|B_t|\geq K}] < \varepsilon$.
b) Show that if $(B_{t\wedge\tau})_{t\geq 0}$ is uniformly integrable, τ is a minimal solution, i.e. for all stopping times τ' with $\tau' \leq \tau$ a.s. and $B_{\tau'} \sim \nu$ there holds $\tau' = \tau$ a.s.

Problem 40. (Embedding of Two Diracs)

Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$ with $\mu \preceq_c \nu$ such there exists $a < b$ in $\mathbb{R}$ such that $\text{supp}(\nu) = \{a, b\}$.

a) Show that $\text{supp}(\mu) \subseteq [a, b]$.
b) Find a solution to SEP(μ, ν).
c) Show that there exists only one solution to SEP(μ, ν).

Problem 41. (Azéma-Yor Solution I)

Let $n \in \mathbb{N}$, $x_1 < ... < x_n$ in $\mathbb{R}$ and $\alpha_1, ..., \alpha_n \in [0, 1]$ such that $\nu_n = \sum_{i=1}^n \alpha_i x_i$ is a centred probability measure. Moreover, let X be a ν_n-distributed random variable. We define the disrete time process $(Y_i)_{i=0,...,n-1}$ as

$$Y_0 = 0 \qquad \text{and} \qquad Y_i = \mathbb{E}\left[X|\mathbb{1}_{X=x_1}, ..., \mathbb{1}_{X=x_i}\right]$$

for all $1 \leq i \leq n-1$.

a) Show that $(Y_i)_{i=0,...,n-1}$ is a martingale with $Y_{n-1} \sim \nu_n$.
b) For all $1 \leq i \leq n-1$ we set

$$b_i := \mathbb{E}\left[X|X \neq x_1, ..., X \neq x_i\right]$$

and $b_0 = 0$. Prove that for all $1 \leq i \leq n-1$ there holds

$$Y_i \in \begin{cases} \{Y_{i-1}\} & Y_{i-1} < b_{i-1} \\ \{x_i, b_i\} & Y_{i-1} = b_{i-1} \end{cases} \qquad \text{and} \qquad x_i \leq b_{i-1} \leq b_i.$$

c) Let B be a standard Brownian Motion. Find uniformly integrable stopping times $\tau_1, ..., \tau_{n-1}$ such that

$$B_{\tau_1+...+\tau_i} \sim Y_i$$

for all $1 \leq i \leq n-1$.

Hint: Recall Problem 40.

d) Show that $\tau_{\mathsf{AY}} := \tau_1 + ... + \tau_{n-1}$ is a solution of SEP(ν_n) that satisfies

$$\tau_{\mathsf{AY}} = \inf\left\{t \geq 0 : B_t \leq \psi_{\nu_n}^{-1}\left(\sup_{s\leq t} B_s\right)\right\}$$

where the barycenter function ψ_{ν_n} is defined as in the lecture.

Problem 42. ((Randomized) Stopping Times)

Let B be a standard Brownian Motion.

a) Let $\nu_1 := \mathcal{N}(0, 2)$ be the normal distribution with mean 0 and variance 2. Find a solution of SEP(ν_1).

b) Let $\alpha := \mathbb{P}[\sup_{t\leq 1}|B_t| \geq 1]$ and $\nu_2 := \alpha(\delta_{-1}+\delta_1)+(1-\alpha)(\delta_{-2}+\delta_2)$. Find a stopping time that solves SEP(ν_2).
c) Let ν_2 be as in b). Find a randomized stopping time (different from your solution in b)) that solves SEP(ν_2).

Problem 43. (Left-Monotone Solution)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ with $\mu \preceq_c \nu$ and $\mu(\{x\}) = 0$ for all $x \in \mathbb{R}$. Moreover, let π_{lm} be the left-monotone coupling of μ and ν (cf. Example 2.33)

a) We know from these lecture notes that there exists two Borel measurable maps T_1, T_2 from $\mathbb{R}$ to $\mathbb{R}$ with

$$\begin{aligned} &T_1(x) \leq x \leq T_2(x) \qquad \text{for all } x \in \mathbb{R} \text{ and} \\ &\pi_{\mathsf{lm}}(\{(x, T_i(x)) : x \in \mathbb{R}, i \in \{1,2\}\}) = 1. \end{aligned} \tag{3.11}$$

Show that for μ-a.e. $x, x' \in \mathbb{R}$ there holds

$$\forall y \in \mathbb{R}: \ x \geq x', \ y \notin (T_1(x), T_2(x)) \ \Rightarrow \ y \notin (T_1(x'), T_2(x')). \tag{3.12}$$

Hint: Recall that π_{lm} is concentrated on a left-monotone set.

b) We can modify the maps T_1 and T_2 such that (3.11) is still satisfied and (3.12) holds for all $x, x' \in \mathbb{R}$. Hence, the set $\mathcal{R} := \{(x,y) \in \mathbb{R}^2 : y \notin (T_1(x), T_2(x))\}$ is a right-barrier in $\mathbb{R}$. We define the stopping time

$$\tau_{\mathsf{lm}} := \inf\{t \geq 0 : (-B_0, B_t) \in \mathcal{R}\}.$$

Prove that τ_{lm} satisfies $\text{Law}(B_0, B_{\tau_{\mathsf{lm}}}) = \pi_{\mathsf{lm}}$.

Problem 44. (Azéma-Yor Solution II)

Let B be a standard Brownian motion and $\nu \in \mathcal{P}_2(\mathbb{R})$ a centred probability measure. We want to show that there exists a solution τ to SEP(ν) that maximizes $\mathbb{E}\left[\sup_{r\leq\tau} B_r\right]$ among all solutions to SEP(ν) and that is of the form

$$\tau = \inf\left\{t \geq 0 : B_t \leq \psi\left(\sup_{r\leq t} B_t\right)\right\} \tag{3.13}$$

for some increasing function ψ.

a) Let $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a strictly increasing continuous function and define the cost functions

$$\begin{aligned} &\gamma((f,s)) := -\sup_{r\leq s} f(r) \qquad \text{and} \\ &\tilde{\gamma}((f,s)) = \varphi\left(\sup_{r\leq s} f(r)\right) f(s)^2. \end{aligned}$$

Show that there exists a randomized stopping time τ that minimizes the secondary optimization w.r.t. $\tilde{\gamma}|\gamma$ (cf. OptSEP_2).

b) Prove that the set of secondary stop-go pairs satisfies

$$\mathsf{SG}_2 \supseteq \left\{(f,s),(g,t) \in S \times S : \ f(s) = g(t), \ \sup_{r\leq s} f(r) > \sup_{r\leq t} g(r)\right\}.$$

c) Let Γ be a $\tilde{\gamma}|\gamma$-monotone set. We define

$$\begin{aligned} \mathcal{R}_{cl} &:= \left\{(m,x) : \exists (g,t) \in \Gamma, \ \sup_{r\leq t} g(r) \leq m, \ g(t) = x\right\} \qquad \text{and} \\ \mathcal{R}_{op} &:= \left\{(m,x) : \exists (g,t) \in \Gamma, \ \sup_{r\leq t} g(r) < m, \ g(t) = x\right\}. \end{aligned}$$

Let τ_{cl} and τ_{op} be the corresponding first hitting times. Prove that $\tau_{cl} \leq \tau \leq \tau_{op}$.

d) Show that there exists an increasing function such that

$$\tau_{cl} = \tau_{op} = \inf\left\{t \geq 0 : B_t \leq \psi\left(\sup_{r\leq t} B_t\right)\right\} \qquad a.s.$$

Hint: You can use that the distribution of $\sup_{r\leq t} B_r$ *is absolutely continuous w.r.t. the Lebesgue measure*

e) Combine a)-d) with statements from the lecture to deduce that there exists a solution to SEP(ν) that maximizes $\mathbb{E}\left[\sup_{r\leq\tau} B_r\right]$ among all solutions to SEP(ν) and that is of the form (3.13).

4. Causal Optimal Transport and Adapted Weak Topologies

Considering for $\varepsilon \geq 0$ the probability measures on $\mathbb{R}^2$:

$$\mathbb{P}^\varepsilon := 1/2\, \delta_{(\varepsilon,1)} + 1/2\, \delta_{(-\varepsilon,-1)}, \tag{4.1}$$

it is immediate that in the sense of weak convergence or Wasserstein distance

$$\mathbb{P}^\varepsilon \to \mathbb{P}^0 = 1/2\delta_{(0,1)} + 1/2\delta_{(0,-1)} \text{ as } \varepsilon \to 0.$$

This can be seen graphically in Figure 5, where also a transport map is depicted:

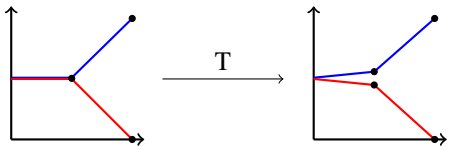


Figure 5. Depiction of stochastic processes with laws $\mathbb{P}^0$ and $\mathbb{P}^\varepsilon$, from left to right. Map T sends the blue path on the left, to the blue path on the right, and similarly for the red paths.

But now let us think of $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ as the path-space of all $\mathbb{R}$-valued processes indexed by the *time-index* set $\{1, 2\}$. This means that for $(x, y) \in \mathbb{R}^2$ we think of x as a position at time 1 and y as a position at time 2. In this interpretation for $\varepsilon > 0$ the measure $\mathbb{P}^\varepsilon$ denotes the law of a stochastic process which is random at time 1 (being either $\pm\varepsilon$), and deterministic at time 2 (being equal to the sign of what happened at time 1). Contrast this to μ^0, the law of a stochastic process deterministic at time 1 (equal to zero) and random at time 2 (being either ± 1). From this perspective, the processes described for $\varepsilon > 0$ couldn't be more different than the process described for $\varepsilon = 0$!

The reason for this dissonance between convergence of measures and convergence of stochastic processes, is that the former largely ignores the time- or information-structure which on the other hand is inherent to stochastic processes. Whereas for stochastic processes we can talk about the *arrow of time*, mathematically encoded by *filtrations*, for measures there is in principle no preferred direction where time (the coordinates) is evolving. To further illustrate the point, let us consider some examples coming from mathematical finance/ stochastic optimization:

Example 4.1 (Optimal Stopping/Pricing of an american option). We revisit $\mathbb{P}^\varepsilon, \mathbb{P}^0$ as in (4.1), and consider

$$v(\mathbb{P}) := \sup\{\mathbb{E}_\mathbb{P}[S_\tau] : \tau \in \{1, 2\} \text{ is } S_1\text{-measurable}\}.$$

This is a so-called optimal stopping problem (τ is a *stopping time*). Alternatively, if S is the value of a stock (perhaps compared to a benchmark, so that negative values make sense), then $v(\mathbb{P})$ is about finding the optimal time to sell it given the available information. Clearly $v(\mathbb{P}^0) = 0$, since e.g. S is a $\mathbb{P}^0$-martingale. On the other hand, for $\mathbb{P}^\varepsilon$ with $\varepsilon > 0$ we may choose $\tau = 1 + \mathbf{1}_{S_1>0}$, so that $v(\mathbb{P}^\varepsilon) \geq 1/2[1 - \varepsilon] \sim 1/2$. We deduce that

$$v(\mathbb{P}^\varepsilon) \not\to v(\mathbb{P}^0).$$

Example 4.2 (Utility Maximization). Say $U : \mathbb{R} \to \mathbb{R}$ is strictly increasing and concave, and consider

$$u(\mathbb{P}) := \max\{\mathbb{E}_\mathbb{P}[U(\pi(S_2 - S_1))] : \pi \in [-1, 1] \text{ is } S_1\text{-measurable}\}.$$

Let for $\varepsilon \geq 0$

$$\mathbb{P}^\varepsilon := \frac{1}{4}[\delta_{(\varepsilon,1)} + \delta_{(\varepsilon,0)} + \delta_{(-\varepsilon,0)} + \delta_{(-\varepsilon,-1)}],$$

so again $\mathbb{P}^\varepsilon \to \mathbb{P}^0$ as $\varepsilon \to 0$, in weak convergence or Wasserstein distance (see Figure 6). By Jensen's inequality we have $u(\mathbb{P}^0) = U(0)$ since $\mathbb{P}^0$ is a martingale measure. On the

other hand, taking for $\varepsilon > 0$ the *strategy* $\pi^* = \text{sign}(S_1)$, we derive by concavity of U

$$u(\mathbb{P}^\varepsilon) \geq \frac{1}{2}[U(1-\varepsilon) + U(-\varepsilon)] \geq \frac{1}{2}[(1-\varepsilon)U(1) + \varepsilon U(0) + U(-\varepsilon)]$$
$$\to u(\mathbb{P}^0) + \frac{1}{2}(U(1) - U(0)) > u(\mathbb{P}^0)$$

as U is strictly increasing. Hence $u(\mathbb{P}^\varepsilon)$ does not converge to $u(\mathbb{P}^0)$. Observe that the strategy π^* precisely exploits the information at time 1 that the process S is going to experience a large movement with $1/2$ probability in a foreseeable (at time 1) upwards/downwards direction, at least for $\varepsilon > 0$. For $\varepsilon = 0$ this peculiarity of π^* is lost since the same large movement of the process X can now occur in any (unforeseeable) direction.

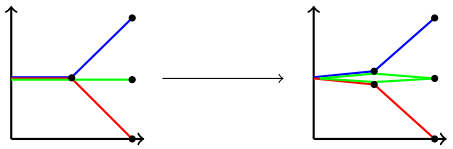


FIGURE 6. Illustration for Example 4.2. Here external randomization is needed for the green paths.

Example 4.3 (Super Hedging). If $C(S_1, S_2)$ is an option, then super-replicating it at zero cost under model $\mathbb{P}$ means finding $\pi = \pi(S_1)$ such that $C \leq \pi[S_2 - S_1]$ ($\mathbb{P}$-a.s.). A smoothed-out version of this is to determine e.g.

$$w(\mathbb{P}) := \inf\{\, \mathbb{E}_{\mathbb{P}}[(C(S_1, S_2) - \pi[S_2 - S_1])_+)] : \pi \in [-2, 2] \text{ is } S_1\text{-measureable } \}.$$

For simplicity we take $C(S_1, S_2) = (S_2)_+$, and again consider Example (4.1). We see that $w(\mathbb{P}^\varepsilon) = 0$ for all $0 < \varepsilon < 1/2$, by taking $\pi = (1-\varepsilon)^{-1}\mathbf{1}_{S_1 > 0}$. On the other hand

$$w(\mathbb{P}^0) = \inf_{\pi \in [-2,2]} 1/2\{(1-\pi)_+ + (\pi)_+\} = 1,$$

and so $w(\mathbb{P}^\varepsilon) \not\to w(\mathbb{P}^0)$.

The above three examples illustrate that the failure of the conventional notions of convergence of probability measures has the following serious negative consequence: most problems in mathematical finance and in stochastic optimization will not behave in a *stable* way. This is troubling, since it means that in practical terms we should not trust a specific model $\mathbb{P}$, as small deviations from it can produce very different conclusions/optimizers.

Overview: The purpose of this part of the course is to introduce a topology which is best suited to remedy the aforementioned instability problem. It turns out that this topology will be metrized by a transport-like distance. This will resemble very closely the definition of Wasserstein distances, with one important difference: the set of couplings used in determining this transport distances will be required to fulfil a certain adaptability, or *causality*, property. This leads us to study the more general problem of *causal optimal transport*.

4.1. **Causal Transport in Discrete Time.** $(\mathbb{X}_t)_t$ (resp. $(\mathbb{Y}_t)_t$) are Polish spaces where our stochastic processes take their values. Also

$$\mathcal{X} = \Pi_{t=1}^N \mathbb{X}_t \text{ (resp. } \mathcal{Y} = \Pi_{t=1}^N \mathbb{Y}_t)$$

are the associated path-spaces of N-step $\mathbb{X}_t$-valued (resp. $\mathbb{Y}_t$-valued) stochastic processes. We denote by X the canonical (identity) process on $\mathcal{X}$, and analogously for Y on $\mathcal{Y}$. Similarly, we denote by (X, Y) the canonical process on $\mathcal{X} \times \mathcal{Y}$. In the latter space we consider the natural filtrations $\mathcal{F}^X$ and $\mathcal{F}^Y$ for X and respectively Y. If $q \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ then we write q_x and q_y for the conditional probability given respectively $X = x$ and $Y = y$ respectively. We convene that $q_{x_1,\dots,x_t,y_1,\dots,y_t}$ is the conditional distribution under q of (x_{t+1}, y_{t+1}) given $(x_1, \dots, x_t, y_1, \dots, y_t)$, unless explicitly written otherwise. Similar conventions apply to measures on $\mathcal{X}$ or $\mathcal{Y}$.

Definition 4.4. *Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ be given. A coupling $q \in \mathsf{Cpl}(\mu,\nu)$ is called causal (between μ and ν) if for any $t \in \{1,\dots,N\}$ and $B \in \mathcal{F}_t^{\mathcal{Y}}$, the mapping $x \in \mathcal{X} \to q_x(B)$ is $\mathcal{F}_t^{\mathcal{X}}$-measurable. The set of all such plans will be denoted*

$$\mathsf{Cpl}_c(\mu,\nu).$$

Remark 4.5. Another way to state the causality property is that, under q, for any t, the random variable Y_t is independent of $X_{t+1},\dots,X_N$ given $X_1,\dots,X_t$. Hence, for a causal coupling, Y is $\mathcal{F}^{\mathcal{X}}$-adapted modulo some randomness which is independent of X.

Along these lines, the following remark is illuminating:

Remark 4.6. If $T : \mathcal{X} \to \mathcal{Y}$ is an adapted measurable map then $q^T(dx,dy) := \mu(dx)\delta_{T(x)}(dy)$ is causal (from μ to $T(\mu)$). Adapted here should be understood as

$$T(x_1,\dots,x_N) = (\,T_1(x_1), T_2(x_1,x_2),\dots,T_n(x_1,\dots,x_N)\,),$$

with $T_t : \Pi_{i=1}^t \mathbb{X}_i \to \mathbb{Y}_t$. It can be formalized, that causal couplings are mixtures/limits of such q^T.

Analogously, we will be interested in transport plans that are "causal in both directions", or bicausal in our terminology. Put $e(x,y) = (y,x)$.

Definition 4.7. *The set of all bicausal plans is explicitly given by*

$$\mathsf{Cpl}_{bc}(\mu,\nu) = \{q \in \mathsf{Cpl}_c(\mu,\nu) \text{ s.t. } e(q) \in \mathsf{Cpl}_c(\nu,\mu)\}\,.$$

The following two optimization problems will play a major role in the rest of this Section.

Definition 4.8. *Given some Borel cost function c defined on $\mathcal{X}\times\mathcal{Y}$ and probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, the* causal optimal transport problem *is to find the minimal cost at which they can be coupled in a causal way, i.e.*

$$\inf_{q\in\mathsf{Cpl}_c(\mu,\nu)} \int c\,dq. \tag{Pc}$$

Minimizing over the set $\mathsf{Cpl}_{bc}(\mu,\nu)$ defines the bicausal optimal transport problem

$$\inf_{q\in\mathsf{Cpl}_{bc}(\mu,\nu)} \int c\,dq. \tag{Pbc}$$

The following proposition, especially part (3), shows that causal optimal transport corresponds to an optimal transport problem under additional linear constraints, just as martingale optimal transport. In particular the set of causal couplings is convex. Observe also that it is non-empty, as the independent coupling $\mu\otimes\nu$ is causal (in fact bicausal).

Proposition 4.9. *The following statements are equivalent:*

(1) q is a causal coupling between the measures μ and ν.

(2) Decomposing q in terms of successive regular kernels

$$q(dx_1,\dots,dx_N,dy_1,\dots,dy_N) = \bar{q}(dx_1,dy_1)q_{x_1,y_1}(dx_2,dy_2)\dots q_{x_1,\dots,x_{N-1},y_1,\dots,y_{N-1}}(dx_N,dy_N), \tag{4.2}$$

then $\bar{q} \in \mathsf{Cpl}(\mathrm{proj}_{\mathbb{X}_1}(\mu), \mathrm{proj}_{\mathbb{Y}_1}(\nu))$ and for $t<N$ and q-almost all $x_1,\dots,x_t,y_1,\dots,y_t$

$$\mathrm{proj}_{\Pi_{j=1}^t \mathbb{X}_j}(q_{x_1,\dots,x_t,y_1,\dots,y_t}) \quad = \quad \mu_{x_1,\dots,x_t}, \tag{4.3}$$

and for ν-almost all $y_1,\dots,y_t$

$$q_{y_1,\dots,y_t}(dy_{t+1}) = \nu_{y_1,\dots,y_t}(dy_{t+1}). \tag{4.4}$$

(3) $q \in \mathsf{Cpl}(\mu,\nu)$ and for all $t \in \{1,\dots,N\}$, $h_t \in C_b(\Pi_{i=1}^t \mathbb{Y}_i)$ and $g_t \in C_b(\mathcal{X})$ we have

$$\int h_t(y_1,\dots,y_t)\left\{g_t(x_1,\dots,x_N) - \int g_t(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\mu_{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)\right\}dq = 0.$$

(4) $q \in \mathsf{Cpl}(\mu,\nu)$ and for every bounded continuous $\mathcal{F}^{\mathcal{Y}}$-adapted process H and each bounded $(\mu,\mathcal{F}^{\mathcal{X}})$-martingale M we have

$$\int \textstyle\sum_{t<N} H_t(y_1,\dots,y_t)\,[M_{t+1}(x_1,\dots,x_{t+1}) - M_t(x_1,\dots,x_t)]\,dq = 0.$$

Proof. STEP 1: Equivalence between Points 1 and 3:

Denote $f^h(x_1,\dots,x_N) := \int h_t(y_1,\dots,y_t)\gamma^{x_1,\dots,x_N}(dy_1,\dots,dy_t)$ with $h_t \in C(\mathbb{R}^t)$. By definition $\gamma \in \Pi_c(\mu,\nu)$ if and only if for all $t \le N$ and all such f^h we have

$$f^h(x_1,\dots,x_N) = \int f^h(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\mu^{x_1,\dots,x_t}(dx_{t+1},\dots,dx_N),$$

which is equivalent to the following:

$$\int g(x_1,\dots,x_N)\Big[f^h(x_1,\dots,x_N) - \int f^h(x_1,\dots,x_t,x_{t+1},\dots,x_N)\mu^{x_1,\dots,x_t}(dx_{t+1},\dots,dx_N)\Big]d\mu = 0,$$

for every function $g \in C_b(\mathbb{R}^N)$ and for all $t \le N$. The fact we can take the g's continuous and not merely Borel bounded comes from the fact that μ is a Borel finite measure on a Polish space. It is easy to see that the previous equation is equivalent to

$$\int f^h(x_1,\dots,x_N)\Big[g(x_1,\dots,x_N) - \int g(x_1,\dots,x_t,x_{t+1},\dots,x_N)\mu^{x_1,\dots,x_t}(dx_{t+1},\dots,dx_N)\Big]d\mu = 0.$$

Finally, by the tower property of conditional expectations the latter is equivalent to:

$$\int h_t(y_1,\dots,y_t)\Big[g_t(x_1,\dots,x_N) - \int g_t(x_1,\dots,x_t,x_{t+1},\dots,x_N)\mu^{x_1,\dots,x_t}(dx_{t+1},\dots,dx_N)\Big]d\gamma = 0.$$

STEP 2: Equivalence between Points 1 and 2:

A $\gamma \in \Pi(\mu,\nu)$, decomposed as in (4.2), is causal if and only of for any time $t \le N$, $\gamma^{x_1,\dots,x_t,y_1,\dots,y_t}(dx_1,\dots,dx_N) = \gamma^{x_1,\dots,x_t}(dx_1,\dots,dx_N)$. Since the x-marginal of γ is μ, these facts imply (4.3). On the other hand, the y-marginal of γ is ν, so (4.4) directly follows.

For the converse direction, it is enough to verify Point 3. for any $t = 1,\dots,N-1$. Since the functions h_t therein depend only on $y_1,\dots,y_t$, the latter can be computed as

$$\begin{aligned}&\int h_t(y_1,\dots,y_t)\Big[g_t(x_1,\dots,x_N) - \int g_t(x_1,\dots,x_t,x_{t+1},\dots,x_N)\mu^{x_1,\dots,x_t}(dx_{t+1},\dots,dx_N)\Big]\\ &\mu^{x_1,\dots,x_{N-1}}(dx_N)\dots\mu^{x_1,\dots,x_t}(dx_{t+1})\gamma^{x_1,\dots,x_{t-1},y_1,\dots,y_{t-1}}(dx_t,dy_t)\dots\gamma^{x_1,y_1}(dx_2,dy_2)\bar{\gamma}(dx_1,dy_1),\end{aligned}$$

which is zero as desired because of $\mu^{x_1,\dots,x_{N-1}}(dx_N)\dots\mu^{x_1,\dots,x_t}(dx_{t+1}) = \mu^{x_1,\dots,x_t}(dx_{t+1},\dots,dx_N)$ (disintegration property).

STEP 3: Equivalence between Points 3 and 4:

Evidently in Point 3 we could have taken h_t and g_t Borel bounded, as STEP 1 suggests. Choosing then $g_t = M_{t+1}$ and $h_t = H_t$ for each $t < N$, and summing up, proves Point 4 from Point 3. Conversely, given t, h_t and g_t we build $H_s = h_t\mathbf{1}_{s\ge t}$ and $M_s = \int g_t(x_1,\dots,x_s,x_{s+1},\dots,x_N)\mu^{x_1,\dots,x_s}(dx_{s+1},\dots,dx_N)$ and conclude by telescopic sum on s. □

Following the strategy from Section 2, the above Proposition 4.2 (3) allows us to derive the existence of optimizers for Problem (Pc) as well as a dual problem (in fact two versions of the dual problem). To this end, we introduce the following sets of functions

$$\mathbb{F} := \left\{ \begin{array}{c} F : \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \text{ s.t. } F(x_1,\dots,x_N,y_1,\dots,y_N) = \\ \sum\limits_{t<N} h_t(y_1,\dots,y_t)\Big[g_t(x_1,\dots,x_N) - \int g_t(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\mu_{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)\Big], \\ \text{with } h_t \in C_b(\Pi_{i=1}^t \mathbb{Y}_i), g_t \in C_b(\mathcal{X}) \text{ for all } t < N \end{array} \right\}, \quad (4.5)$$

$$\mathbb{S} := \left\{ \begin{array}{c} S : \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \text{ s.t. } S(x_1,\dots,x_N,y_1,\dots,y_N) = \\ \sum\limits_{t<N} H_t(y_1,\dots,y_t)\left[M_{t+1}(x_1,\dots,x_{t+1}) - M_t(x_1,\dots,x_t)\right], \\ \text{with } H_t \in C_b(\Pi_{i=1}^t \mathbb{Y}_i), M_t \in C_b(\Pi_{i=1}^t \mathbb{X}_i) \text{ for all } t < N, \text{ and with } \{M_t\}_t \text{ a } \mu\text{-martingale} \end{array} \right\}. \quad (4.6)$$

Then we have the following result on existence and duality:

Theorem 4.10. *Suppose that $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{+\infty\}$ is lower semicontinuous and bounded from below. Then there is no duality gap*

$$\begin{aligned}\inf_{q\in\mathsf{Cpl}_c(\mu,\nu)} \int c dq &= \sup_{\substack{\Phi\in C_b(\mathcal{X}),\Psi\in C_b(\mathcal{Y}),F\in\mathbb{F}\\ \Phi\oplus\Psi\le c+F}} \left[\int \Phi d\mu + \int \Psi d\nu\right]\\ &= \sup_{\substack{\Phi\in C_b(\mathcal{X}),\Psi\in C_b(\mathcal{Y}),S\in\mathbb{S}\\ \Phi\oplus\Psi\le c+S}} \left[\int \Phi d\mu + \int \Psi d\nu\right]\\ &= \sup_{\substack{\Psi\in C_b(\mathcal{Y}),F\in\mathbb{F}\\ \Psi\le c+F}} \left[\int \Psi d\nu\right],\end{aligned}$$

and the infimum on the l.h.s. (i.e. (Pc)*) is attained.*

The key to proving this result is the following lemma proving compactness of $\mathsf{Cpl}_c(\mu,\nu)$.

Lemma 4.11. *Let $B \subseteq \mathcal{P}(\mathcal{Y})$ be a weakly compact set of measures, and $\mu \in \mathcal{P}(\mathcal{X})$ be given. Then the set $\mathsf{Cpl}_c(\mu, B) := \cup_{\nu\in B}\mathsf{Cpl}_c(\mu,\nu)$ is weakly compact. In particular the set $\mathsf{Cpl}_c(\mu,\nu)$ is weakly compact.*

Proof. Call τ and σ the Polish topologies of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Consider

$$\begin{aligned}\mathbb{X}_1 \ni x_1 &\mapsto \mu_{x_1}(dx_2,\dots,dx_N) \in \mathcal{P}(\Pi_{i=2}^N \mathbb{X}_i)\\ \mathbb{X}_1\times\mathbb{X}_2 \ni (x_1,x_2) &\mapsto \mu_{x_1,x_2}(dx_3,\dots,dx_N) \in \mathcal{P}(\Pi_{i=3}^N \mathbb{X}_i)\\ &\vdots\\ \Pi_{i=1}^{N-1}\mathbb{X}_i \ni (x_1,\dots,x_{N-1}) &\mapsto \mu_{x_1,\dots,x_{N-1}}(dx_N) \in \mathcal{P}(\mathbb{X}_N),\end{aligned}$$

for the regular conditional distributions of μ. We can view the collection of these $N-1$ measurable mappings as a measurable function from $\mathcal{X}$ into a Polish space. By [Kec95, Theorem 13.11], there is a stronger (finer)[3] Polish topology on $\mathcal{X}$, which we call $\hat\tau$, whose Borel sets are the same as for τ, and such that the above mapping is continuous when the domain space $\mathcal{X}$ is given the $\hat\tau$ topology. Let us denote by Σ_1 the topology on $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ generated by convergence w.r.t. $\tau\times\sigma$-continuous bounded functions, and Σ_2 the topology generated by convergence w.r.t. $\hat\tau\times\sigma$-continuous bounded functions. By Proposition 4.9 we know that causality can be tested by integration against functions of the form

$$h(y_1,\dots,y_t)\left[g(x_1,\dots,x_N) - \int g(x_1,\dots,x_t,\bar x_{t+1},\dots,\bar x_N)\mu_{x_1,\dots,x_t}(d\bar x_{t+1},\dots,d\bar x_N)\right],$$

for each t, h bounded σ-continuous and g bounded τ-continuous. Notice that the function in brackets is then by definition also $\hat\tau$-continuous, so the overall expression is $\hat\tau\times\sigma$-continuous. It follows that $\mathsf{Cpl}_c(\mu,B)$ is Σ_2-closed. On the other hand, $\mathsf{Cpl}_c(\mu,B)$ is also Σ_2-tight, since as a Borel measure μ is still tight w.r.t. the stronger topology induced by $\hat\tau$-continuous bounded functions ($\hat\tau$ is still Polish). Thus $\mathsf{Cpl}_c(\mu,B)$ is Σ_2-compact and in particular also Σ_1-compact. □

Proof of Theorem 4.10 . Existence follows from Lemma 4.11. By Proposition 4.9, (Pc) is equal to

$$\inf_{q\in\mathsf{Cpl}(\mu,\nu)}\sup_{F\in\mathbb{F}}\int [c+F]dq.$$

Going back to the proof of Lemma 4.11, there is a stronger topology under which $\mathsf{Cpl}(\mu,\nu)$ is compact, $q\mapsto\int F dq$ is continuous for all $F\in\mathbb{F}$, and $q\mapsto\int c dq$ is lower semicontinuous. As in the proof of Theorem 2.13 we may apply the minmax Theorem 2.15 obtaining the equality between (Pc) and

$$\sup_{F\in\mathbb{F}}\inf_{q\in\mathsf{Cpl}(\mu,\nu)}\int [c+F]dq.$$

[3]If τ and τ' are two topologies on $\mathcal{X}$ s.t. $\tau\subseteq\tau'$ we say that τ' is stronger/finer than τ.

Applying the Kantorovich duality Theorem 1.21 to the inner infimum above we conclude

$$\inf_{q\in \mathsf{Cpl}_c(\mu,\nu)} \int c dq \;=\; \sup_{\substack{\Phi\in C_b(\mathcal{X}),\Psi\in C_b(\mathcal{Y}),F\in\mathbb{F}\\ \Phi\oplus\Psi\leq c+F}} \Big[\int \Phi d\mu + \int \Psi d\nu\Big].$$

Observe that $\Phi - \int \Phi d\mu \in \mathbb{F}$, from which it is possible to eliminate Φ from the above supremum. We leave the statement involving $\mathbb{S}$ as an exercise. □

4.2. **Bicausal Transport in Discrete Time.** Recall the set of bicausal transports

$$\mathsf{Cpl}_{bc}(\mu,\nu) = \{q \in \mathsf{Cpl}_c(\mu,\nu) \text{ s.t. } e(q) \in \mathsf{Cpl}_c(\nu,\mu)\},$$

where $e(x,y) = (y,x)$. We leave it as an exercise to the reader to find the analogue of Theorem 4.10 in the bicausal case. We are interested here in *finding* the optimizer of the bicausal transport problem (Pbc). Under certain conditions, this also leads to the optimizer of (Pc). Recall (Pbc)

$$\inf_{q\in \mathsf{Cpl}_{bc}(\mu,\nu)} \int c dq.$$

It will useful to first consider the following "nested" problem:

$$\inf_{q^1\in\mathsf{Cpl}(\mathrm{proj}_{\mathbb{X}_1}(\mu),\mathrm{proj}_{\mathbb{Y}_1}(\nu))} \int q^1(dx_1,dy_1) \inf_{q^2\in\mathsf{Cpl}(\mu_{x_1},\nu_{y_1})} \int q^2(dx_2,dy_2)\dots$$
$$\dots \inf_{q^N\in\mathsf{Cpl}(\mu_{x_1,\dots,x_{N-1}},\nu_{y_1,\dots,y_{N-1}})} \int q^N(dx_N,dy_N) c(x_1,\dots,x_N,y_1,\dots,y_N). \quad \text{(Dyn-Pbc)}$$

The previous recursive problem is motivated by the following structure result, the proof of which is analogous to the causal case, Proposition 4.9.

Proposition 4.12. *Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$.*
If $q \in \mathsf{Cpl}_{bc}(\mu,\nu)$ is decomposed as in (4.2)*, then the following conditions on the kernels hold:*

(i) $\bar{q} \in \mathsf{Cpl}(\mathrm{proj}_{\mathbb{X}_1}(\mu), \mathrm{proj}_{\mathbb{Y}_1}(\nu))$*, and*
(ii) *successively for $t < N$ and for q-almost every $x_1,\dots,x_t,y_1,\dots,y_t$ there holds*

$$q_{x_1,\dots,x_t,y_1,\dots,y_t}(dx_{t+1},dy_{t+1}) \in \mathsf{Cpl}(\mu_{x_1,\dots,x_t}(dx_{t+1}), \nu_{y_1,\dots,y_t}(dy_{t+1})).$$

Conversely, given regular kernels

$$\bar{q}(dx_1,dy_1), q_{x_1,y_1}(dx_2,dy_2),\dots,q_{x_1,\dots,x_{N-1},y_1,\dots,y_{N-1}}(dx_N,dy_N),$$

satisfying the properties (i)–(ii), the measure q constructed as in (4.2) *belongs to $\mathsf{Cpl}_{bc}(\mu,\nu)$.*

The recursion corresponding to (Dyn-Pbc) (starting from $V_N^c := c$) is:

$$V_t^c(x_1,\dots,x_t,y_1,\dots,y_t) =$$
$$\inf_{q^{t+1}\in\mathsf{Cpl}(\mu_{x_1,\dots,x_t},\nu_{y_1,\dots,y_t})} \int q^{t+1}(dx_{t+1},dy_{t+1}) V_{t+1}^c(x_1,\dots,x_{t+1},y_1,\dots,y_{t+1}), \quad (4.7)$$

and so we want to compare the values of (Dyn-Pbc), (Pbc) and

$$V_0^c := \inf_{q^1\in\mathsf{Cpl}(\mathrm{proj}_{\mathbb{X}_1}(\mu),\mathrm{proj}_{\mathbb{Y}_1}(\nu))} \int V_1^c(x_1,y_1) q^1(dx_1,dy_1).$$

Proposition 4.13. *Given a Borel bounded from below cost function c, we have that the nested problem* (Dyn-Pbc) *is well-defined, namely the successive integrals in* (4.7) *are well-defined, and the values of* (Dyn-Pbc)*,* (Pbc) *and V_0^c coincide and hence the optimization problems are equivalent.*

We omit the proof, as it needs some aspects of the theory of measurable selections.

Remark 4.14. The equivalence between (Dyn-Pbc), (Pbc) and V_0^c is called *Dynamic Programming Principle* or DPP in short. The power of the DPP is that it reduces the computation of (Pbc) to solving a sequence of classical transport problems (see (4.7)). In fact, this also suggests how to build an optimizer for (Pbc).

We now show how sometimes Problems (Pbc) and (Pc) are equivalent:

Proposition 4.15. *Assume that c has a separable structure*

$$c(x_1, \dots, x_N, y_1, \dots, y_N) = \sum_{t \le N} c_t(x_t, y_t). \tag{4.8}$$

Further suppose that the starting measure μ is the product of its marginals, i.e.

$$\mu(dx_1, \dots, dx_N) = \mu_1(dx_1) \dots \mu_N(dx_N). \tag{4.9}$$

Then the values of (Pc) *and* (Pbc) *coincide.*

Proof. We only need to show that the value of the bicausal problem is less or equal than that of the causal problem. Start with $q \in \mathsf{Cpl}_c(\mu, \nu)$ and decompose it as in (4.2). From Proposition 4.9, we know that the following conditions ($t < N$) are satisfied by the kernels $\bar{q}(dx_1, dy_1), q_{x_1,y_1}(dx_2, dy_2)$, up to $q_{x_1,\dots,x_{N-1},y_1,\dots,y_{N-1}}(dx_N, dy_N)$:

$$q_{x_1,\dots,x_t,y_1,\dots,y_t}(dx_{t+1}, \mathbb{Y}_{t+1}) = \mu_{t+1}(dx_{t+1}) \tag{4.10}$$

and

$$\begin{aligned} \int_{x_1,\dots,x_t} q_{x_1,\dots,x_t,y_1,\dots,y_t}(\mathbb{X}_{t+1}, dy_{t+1}) q_{y_1,\dots,y_t}(dx_1, \dots, dx_t) &= \nu_{y_1,\dots,y_t}(dy_{t+1}) \\ &= q_{y_1,\dots,y_t}(\mathbb{X}_{t+1}, dy_{t+1}) \end{aligned} \tag{4.11}$$

We can rewrite (4.10) in the following way:

$$\begin{aligned} \int_{x_1,\dots,x_t} q_{x_1,\dots,x_t,y_1,\dots,y_t}(dx_{t+1}, \mathbb{Y}_{t+1}) q_{y_1,\dots,y_t}(dx_1, \dots, dx_t) &= \mu_{t+1}(dx_{t+1}) \\ &= q_{y_1,\dots,y_t}(dx_{t+1}, \mathbb{Y}_{t+1}) \end{aligned} \tag{4.12}$$

Therefore, we can construct a new plan $\tilde{q}$ as follows

$$\tilde{q}(dx_1, \dots, dx_N, dy_1, \dots, dy_N) = \bar{q}(dx_1, dy_1) \tilde{q}_{y_1}(dx_2, dy_2) \dots \tilde{q}_{y_1,\dots,y_{N-1}}(dx_N, dy_N),$$

with

$$\tilde{q}_{y_1,\dots,y_t}(dx_{t+1}, dy_{t+1}) = \int_{x_1,\dots,x_t} q_{x_1,\dots,x_t,y_1,\dots,y_t}(dx_{t+1}, dy_{t+1}) q_{y_1,\dots,y_t}(dx_1, \dots, dx_t). \tag{4.13}$$

Due to (4.12) and (4.11), one can see that for any $t < N$ each kernel

$$q_{y_1,\dots,y_t} \in \mathsf{Cpl}(\mu_{t+1}(dx_{t+1}), \nu_{y_1,\dots,y_t}(dy_{t+1})).$$

Then, from Proposition (4.12), we know that $\tilde{q} \in \mathsf{Cpl}_{bc}(\mu, \nu)$. By (4.13) and separability of c we obtain $\int c dq = \int c d\tilde{q}$ so that $\int_{q \in \mathsf{Cpl}_{bc}(\mu,\nu} \int c dq \le \int_{q \in \mathsf{Cpl}_c(\mu,\nu} \int c dq$ proving the result. □

Remark 4.16. One can give examples showing that if either μ is not the product of its marginals, or the cost function is not separable, then the causal-bicausal equality may fail (Exercise).

4.2.1. *The case of $\mathcal{X} = \mathcal{Y} = \mathbb{R}^N$.* To finalize this part, we specialize the discussion to

$$\mathbb{X}_t = \mathbb{Y}_t = \mathbb{R} \text{ so that } \mathcal{X} = \mathcal{Y} = \mathbb{R}^N.$$

We denote by

$$F_\eta(\cdot) := \eta((-\infty, \cdot]),$$

the usual cumulative distribution function of a probability measure η on the line, and by $F_\eta^{-1}(u)$ its left-continuous generalized inverse, i.e.

$$F_\eta^{-1}(u) = \inf \left\{ y : F_\eta(y) \ge u \right\}.$$

Let us illustrate the DPP for $N = 2$:

Example 4.17. Take $N = 2$ and $c = [x_1 - y_1]^2 + [x_2 - y_2]^2$. Using the optimality of the monotone coupling on the line we get:

$$\begin{aligned} V_1^c(x_1, y_1) &:= \inf_{q_2 \in \mathsf{Cpl}(\mu_{x_1}, \nu_{y_1})} \int q_2(dx_2, dy_2)[x_2 - y_2]^2 \\ &= \int_0^1 \left[F_{\nu_{y_1}}^{-1}(u) - F_{\mu_{x_1}}^{-1}(u) \right]^2 du, \end{aligned}$$

so that V_0^c is equal to

$$\inf_{q_1 \in \mathsf{Cpl}(\mathrm{proj}_{\mathbb{X}_1}(\mu), \mathrm{proj}_{\mathbb{Y}_1}(\nu))} \Big\{ \int_{x_1, y_1} q_1(dx_1, dy_1)[x_1 - y_1]^2 + \int_0^1 \Big[F^{-1}_{\nu_{y_1}}(u) - F^{-1}_{\mu_{x_1}}(u) \Big]^2 du \Big\}.$$

To simplify notation, we denote F_{η_1} the cdf of $\mathrm{proj}_{\mathbb{X}_1}(\eta)$ whenever η is a measure $\mathcal{X}_N$.

Definition 4.18. *The increasing N-dimensional* Knothe-Rosenblatt *rearrangement*[4] *of μ and ν is defined as the law of the random vector $(X_1^*, \dots, X_N^*, Y_1^*, \dots, Y_N^*)$ where*

$$\begin{aligned} X_1^* &= F^{-1}_{\mu_1}(U_1), & Y_1^* &= F^{-1}_{\nu_1}(U_1), & &\text{and inductively} \\ X_t^* &= F^{-1}_{\mu_{X_1^*, \dots, X_{t-1}^*}}(U_t), & Y_t^* &= F^{-1}_{\nu_{Y_1^*, \dots, Y_{t-1}^*}}(U_t), & &\text{for } t = 2, \dots, N, \end{aligned} \tag{4.14}$$

for $U_1, \dots, U_N$ independent and uniformly distributed random variables on $[0,1]$. Additionally, if μ-a.s. all the conditional distributions of μ are atomless (e.g. if μ has a density), then this rearrangement is induced by a (Monge) map

$$(x_1, \dots, x_N) \mapsto T(x_1, \dots, x_N) := (T^1(x_1), T^2(x_2; x_1), \dots, T^N(x_N; x_1, \dots . x_{N-1})),$$

where

$$\begin{aligned} T^1(x_1) &:= F^{-1}_{\nu_1} \circ F_{\mu_1}(x_1), \\ T^t(x_t; x_1, \dots, x_{t-1}) &:= F^{-1}_{\nu_{T^1(x_1), \dots, T^{t-1}(x_{t-1}; x_1, \dots, x_{t-2})}} \circ F_{\mu_{x_1, \dots, x_{t-1}}}(x_t), \quad t \geq 2. \end{aligned} \tag{4.15}$$

The intuition of the Knothe-Rosenblatt rearrangement is as follows: one first couples the first marginals of μ and ν increasingly, then given this coupling one couples the conditional distributions of μ and ν at "time 2 given time 1" likewise in increasing fashion, and so forth. Observe that if the map T above was differentiable, then its Jacobian matrix would be lower-triangular.

It is straightforward, but slightly tedious, to observe that the Knothe-Rosenblatt rearrangement is a bicausal coupling:

Remark 4.19. The Knothe-Rosenblatt rearrangement (4.14) (coinciding with (4.15) in the atomless case) is always bicausal: for each bounded Borel $g(\cdot, \cdot)$ we may define

$$y_1 \mapsto G(y_1) := \int_0^1 g(y_1, F^{-1}_{\nu_{y_1}}(v)) dv,$$

so that denoting $X_1^* := F^{-1}_{\mu_1}(U_1)$, $Y_1^* := F^{-1}_{\nu_1}(U_1)$ and $Y_2^* := F^{-1}_{\nu_{F^{-1}_{\nu_1}(U_1)}}(U_2)$, we get

$$\begin{aligned} E[f(X_1^*) g(Y_1^*, Y_2^*)] &= \int_0^1 \int_0^1 f(F^{-1}_{\mu_1}(u_1)) g\Big(F^{-1}_{\nu_1}(u_1), F^{-1}_{\nu_{F^{-1}_{\nu_1}(u_1)}}(u_2)\Big) du_2 du_1 \\ &= \int_0^1 f(F^{-1}_{\mu_1}(u_1)) G\left(F^{-1}_{\nu_1}(u_1)\right) du_1 \\ &= E[\, E[f(X_1^*)|Y_1^*]\, G(Y_1^*)] \\ &= E[\, E[f(X_1^*)|Y_1^*]\, g(Y_1^*, Y_2^*)]. \end{aligned}$$

Thus the law of X_1^* given (Y_1^*, Y_2^*), equals the law of X_1^* given Y_1^*. The same holds inverting the roles of μ and ν and a similar argument applies to greater time indices.

It was observed in [CGS10] that in given situations the Knothe-Rosenblatt rearrangement is a limit of Brenier maps:

Remark 4.20. Assume that both μ and ν are absolutely continuous, and consider the classical optimal transport problem between these measures under the cost function

$$c^\varepsilon := \sum_{i=1}^N \varepsilon^i |x_i - y_i|^2.$$

As in Brenier's theorem the aforementioned transport problem admits a unique optimizer q_ε^* for every $\varepsilon > 0$. Then as $\varepsilon \to 0$ the couplings q_ε^* converge to the Knothe-Rosenblatt rearrangement between μ and ν. This is natural, since the ε^i weighs make it relatively expensive to violate the bicausality property.

[4]The reader might find it in the literature by the name *quantile transform* or *Knothe-Rosenblatt coupling*.

As may be suspected from its structure, or the previous remark, the Knothe-Rosenblatt rearrangement plays a similar role as Brenier maps in the world of (bi)causal transport.

Definition 4.21. *For two real functions f and g, we say that they are co-monotone if either they are both increasing, or they are both decreasing, or one of them is constant (the other being arbitrary).*

Theorem 4.22. *For each $t = 1, \dots, N-2$, all $(x_1, \dots, x_{t-1}), (y_1, \dots, y_{t-1})$, and $u \in \mathbb{R}$, suppose:*

(a) the functions $x_t \mapsto F_{\mu_{x_1,\dots,x_{t-1},x_t}}(u)$ and $y_t \mapsto F_{\nu_{y_1,\dots,y_{t-1},y_t}}(u)$ are co-monotone.

Assume further that $c(x_1, \dots, x_N, y_1, \dots, y_N) := \sum_{t \le N} c_t(x_t - y_t)$, where each c_t is convex and finite. Then the Knothe-Rosenblatt rearrangement (4.14) *is optimal for* (Pbc)*. Additionally, if μ-a.s. all the conditional distributions of μ are atomless (e.g. if μ has a density), then this rearrangement is induced by the Monge map determined by* (4.15)*.*

Corollary 4.23. *Suppose μ and ν are Markovian (i.e. $\mu^{x_1,\dots,x_t}(dx_{t+1}) = \mu^{x_t}(dx_{t+1})$ for all t, and similarly for ν) and*

$$\forall t \in \{1, \dots, N_1\}, \forall a \in \mathbb{R} : \; x_t \mapsto F_{\mu^{x_t}}(a) \text{ and } y_t \mapsto F_{\nu^{y_t}}(a) \text{ are decreasing .}$$

(We say that the transition kernels of μ and ν are increasing in first order stochastic dominance.) Then Condition (a) from Theorem 4.22 holds.

Corollary 4.24. *Suppose $c(x_1, \dots, x_N, y_1, \dots, y_N) := \sum_{t \le N} c_t(x_t - y_t)$, where each c_t is convex and finite, and that μ is equal to the product of its one-dimensional marginals. Then the Knothe-Rosenblatt rearrangement* (4.14) *is optimal for* (Pc)*. Additionally, if all the one-dimensional marginals of μ are atomless (e.g. if they have a density), then this rearrangement is induced by the Monge map determined by* (4.15)*.*

Proof of Corollary 4.24. It follows immediately from Proposition 4.15, Theorem 4.22, and the fact that Condition (a) therein is automatically fulfilled when μ has a product form. □

We will need the following technical lemma in order to prove Theorem 4.22:

Lemma 4.25. *For any convex function $c : \mathbb{R} \to \mathbb{R}$ and any functions $f, g : \mathbb{R} \to \mathbb{R}$ which are co-monotone, and for all pairs $x \le \bar{x}$ and $y \le \bar{y}$, the following holds*

$$c(f(x) - g(y)) + c(f(\bar{x}) - g(\bar{y})) - c(f(\bar{x}) - g(y)) - c(f(x) - g(\bar{y})) \le 0. \tag{4.16}$$

Proof. It is known that any convex function has a monotone slope, meaning that for any $h \ge 0$ the function

$$\frac{c(z) - c(z-h)}{h},$$

is increasing in z. This implies, that for any $z \le \bar{z}$

$$c(z) - c(z-h) \le c(\bar{z}) - c(\bar{z} - h).$$

If both of the functions f and g are non decreasing, we take $z = f(\bar{x}) - g(\bar{y})$, $\bar{z} = f(\bar{x}) - g(y)$, and $h = f(\bar{x}) - f(x) \ge 0$ so that $z \le \bar{z}$. For non increasing functions f and g, we write

$$c(f(x) - g(y)) = c((-g(y)) - (-f(x)),$$

and conclude as before. If either f or g is constant there is nothing to prove. □

Proof of Theorem 4.22 . We start with $N = 2$. From classical optimal transport it is known that

$$\inf_{q_2 \in \mathsf{Cpl}(\mu_{x_1}, \nu_{y_1})} \int c_2(x_2 - y_2) q_2(dx_2, dy_2) = \int c_2(F^{-1}_{\mu_{x_1}}(u_2) - F^{-1}_{\nu_{y_1}}(u_2)) du_2.$$

Consider the function

$$V^{u_2}(x_1, y_1) := c_1(x_1 - y_1) + c_2(F^{-1}_{\mu_{x_1}}(u_2) - F^{-1}_{\nu_{y_1}}(u_2)). \tag{4.17}$$

For every fixed value u_2, the function $V^{u_2}(x_1, y_1)$ verifies the following *L-superadditivity* inequality

$$V^{u_2}(x_1, y_1) + V^{u_2}(\bar{x}_1, \bar{y}_1) - V^{u_2}(\bar{x}_1, y_1) - V^{u_2}(x_1, \bar{y}_1) \le 0, \tag{4.18}$$

for any $x_1 \leq \bar{x}_1$ and $y_1 \leq \bar{y}_1$. Equivalently

$$\begin{aligned} &c_1(x_1 - y_1) + c_1(\bar{x}_1 - \bar{y}_1) - c_1(\bar{x}_1 - y_1) - c_1(x_1 - \bar{y}_1) + \\ &c_2(F^{-1}_{\mu_{x_1}}(u_2) - F^{-1}_{\nu_{y_1}}(u_2)) + c_2(F^{-1}_{\mu_{\bar{x}_1}}(u_2) - F^{-1}_{\nu_{\bar{y}_1}}(u_2)) - \\ &c_2(F^{-1}_{\mu_{\bar{x}_1}}(u_2) - F^{-1}_{\nu_{y_1}}(u_2)) - c_2(F^{-1}_{\mu_{x_1}}(u_2) - F^{-1}_{\nu_{\bar{y}_1}}(u_2)) \leq 0. \end{aligned} \tag{4.19}$$

Indeed, from Lemma 4.25 together with the stated Condition (a), one obtains the above inequality. We now integrate (4.18) obtaining for the value function that

$$V^c_1(x_1, y_1) + V^c_1(\bar{x}_1, \bar{y}_1) - V^c_1(\bar{x}_1, y_1) - V^c_1(x_1, \bar{y}_1) \leq 0. \tag{4.20}$$

It can be seen (cf. Problem 4 on Exercise Sheet 2 [CSS76]) that Property (4.20) is enough to guarantee the optimality of the monotone coupling when V^c_1 is the cost function. This and the DPP proves the optimality of the Knothe-Rosenblatt rearrangement for $N = 2$. Iterating our arguments for general N, we conclude. □

4.2.2. *An application of the Knothe-Rosenblatt rearrangement.* The Euclidean isoperimetric inequality states that among all subsets of $\mathbb{R}^N$ with a given perimeter, the Euclidean Ball has the largest volume. This can be proved, modulo technicalities, with the help of the Knothe-Rosenblatt rearrangement. We follow [Vil09, Ch. 2]

Let μ and ν denote respect. the uniform probability meausures on a nice set Ω and on the ball B, with $|\partial\Omega| = |\partial B| = N|B|$. Let T be the Knothe-Rosenblatt map from μ to ν, which we assume smooth enough. By the change of variables formula we have

$$\frac{1}{|\Omega|} = \det(\nabla T(x))\frac{1}{|B|}.$$

Since ∇T is a triangular matrix (here we used the structure of T), we have $\det(\nabla T(x)) = \Pi_i \lambda_i(x)$ and $\mathrm{div}(T)(x) = \sum_i \lambda_i(x)$, where $\{\lambda_i(x)\}_{i=1}^N$ are the eigenvalues of $\nabla T(x)$. By the arithmetic-geometric mean inequality, we have

$$[\det(\nabla T(x))]^{1/N} \leq \frac{\mathrm{div}(T)(x)}{N}$$

Hence we conclude $\frac{1}{|\Omega|^{1/N}} \leq \frac{\mathrm{div}(T)(x)}{N|B|^{1/N}}$. Integrating this w.r.t Lebesgue measure on Ω, and using the divergence theorem, we have

$$|\Omega|^{1-1/N} \leq \frac{1}{N|B|^{1/N}} \int_\Omega \mathrm{div}(T)(x)dx = \frac{1}{N|B|^{1/N}} \int_{\delta\Omega} (T \cdot \hat{n})(x)\mathcal{H}^{n-1}(dx),$$

where $\hat{n}$ is the unit outer normal vector to $\partial\Omega$ and $\mathcal{H}^{n-1}$ denotes the surface measure on $\partial\Omega$. As $T(x) \in B$ has norm bounded by 1 if follows that $|T \cdot \hat{n}(x)| \leq 1$ so that we can conclude $|\Omega|^{1-1/N} \leq \frac{|\partial\Omega|}{N|B|^{1/N}} = \frac{|\partial B|}{N|B|^{1/N}} = |B|^{1-1/N}$ so that $|\Omega| \leq |B|$, as desired.

4.3. **The weak adapted topology.** Let us go back to the family of measures described in (4.1), or for that matter, in any of the examples presented thereafter. We want to define a topology on probability measures which

- takes explicit into account that information increases with time,
- guarantees the stability (ie. continuity) of the optimization problems described in the examples.

Looking at the sequence in (4.1), we want this topology to preclude its convergence as $\varepsilon \to 0$. Intuitively, this topology should "remember" that the value $S_2 = 1$ came from $S_1 = \varepsilon$ and not from $S_1 = -\varepsilon$, and vice-versa. It is hence intuitive to identify for $\varepsilon > 0$

$$\mathbb{P}^\varepsilon := 1/2\,\delta_{(\varepsilon,1)} + 1/2\,\delta_{(-\varepsilon,-1)} \;\text{ with }\; \tilde{\mathbb{P}}^\varepsilon := 1/2\,\delta_{(\varepsilon,\delta_1)} + 1/2\,\delta_{(-\varepsilon,\delta_{-1})}.$$

Observe that $\tilde{\mathbb{P}}^\varepsilon$ is a probability measure on $\mathbb{R} \times \mathcal{P}(\mathbb{R})$, ie. $\tilde{\mathbb{P}}^\varepsilon \in \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))$. Then clearly

$$\tilde{\mathbb{P}}^\varepsilon \to \tilde{\nu} := 1/2\delta_{(0,\delta_1)} + 1/2\delta_{(0,\delta_{-1})} \;\text{ as } \varepsilon \to 0.$$

However the natural element in $\mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))$ which can be identified with $\mathbb{P}^0$ is

$$\tilde{\mathbb{P}}^0 := \delta_{(0,1/2\delta_1+1/2\delta_{-1})},$$

which is different from $\tilde{\nu}$. Intuitively, the element $\tilde{\nu}$ did remember in the limit the relationship between the observations at time 1 and 2, whereas $\tilde{\mathbb{P}}^0$ is oblivious of time 1. This experiment suggest that we should endow $\mathcal{P}(\mathbb{R}\times\mathbb{R})$ with the topology inherited when we identify it as a subset of $\mathcal{P}(\mathbb{R}\times\mathcal{P}(\mathbb{R}))$!

Throughout this section we will only consider, for the sake of familiarity, the case of stochastic processes on two time steps, corresponding to

$$N = 2,\ \mathbb{X}_t = \mathbb{Y}_t, \text{ and } \mathsf{X} = \mathsf{Y} = \mathbb{X}_1\times\mathbb{X}_2.$$

We will also write $p_1 := \mathrm{proj}_{\mathbb{X}_1}$. Most arguments still work in the general case. (It is a good exercise to convince yourself that in the case of $N = 3$ the right space to consider is $\mathcal{P}(\mathbb{R}\times\mathcal{P}(\mathbb{R}\times\mathcal{P}(\mathbb{R})))$ reflecting a tree-like structure that takes the information flow into account, see Remark 4.40.)

Definition 4.26. *Let*

$$J : \mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2) \mapsto \mathcal{P}(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2))$$

be defined by $J(\mu) = \mu\circ(x_1 \mapsto (x_1, \mu_{x_1}))^{-1}$. *This is the* embedding *operator from* $\mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2)$ *to* $\mathcal{P}(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2))$.

Intuitively, one considers the first coordinate x_1 as an $\mathbb{X}_1$-valued random variable and μ_{x_1} as a $\mathcal{P}(\mathbb{X}_2)$-valued random variable. Then $J(\mu)$ is just the joint law of these two random variables under μ.

Definition 4.27. *Let*

$$I : \mathcal{P}(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2)) \mapsto \mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2)$$

be defined by the following condition: $\mu = I(P)$ *iff*

$$\forall f\in C_b(\mathbb{X}_1\times\mathbb{X}_2):\ \int_{\mathbb{X}_1\times\mathbb{X}_2} f(x_1,x_2)d\mu(x_1,x_2) = \int_{\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2)}\left(\int_{\mathbb{X}_2} f(x_1,z)p(dz)\right)P(dx_1,dp).$$

This is the intensity *operator.*

Intuitively, I takes a measure $P\in\mathcal{P}(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2))$ and averages its second coordinate in order to obtain $I(P)\in P(\mathbb{X}_1\times\mathbb{X}_2)$. Remark that

$$I\circ J(\mu) = \mu,$$

so I is just the left inverse of J. On the other hand, if $P\in\mathrm{Range}(J)$, then $J\circ I(P) = P$.

Definition 4.28. *The weak adapted topology on* $\mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2)$ *is the relativization of the weak topology on* $\mathcal{P}(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2))$ *when we identify* $\mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2)$ *with* $J(\mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2))\subseteq\mathcal{P}(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2))$. *Equivalently, the weak adapted topology on* $\mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2)$ *is the coarsest topology (ie. the weakest one) on the domain of* J *which makes* J *continuous when we equip the target space* $\mathcal{P}(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2))$ *with the weak topology.*

Remark 4.29. A sequence $\{\mu_n\}_n\subseteq\mathcal{P}(\mathbb{X}_1\times\mathbb{X}_2)$ converges to μ in the weak adapted topology, if and only if for all $F\in C_b(\mathbb{X}_1\times\mathcal{P}(\mathbb{X}_2))$ it holds

$$\int F(x_1,(\mu_n)_{x_1})\mu_n(dx) \to \int F(x_1,\mu_{x_1})\mu(dx).$$

Example 4.30. The discussion at the beginning of this subsection shows that, for the sequence in (4.1), we have $\mathbb{P}^\varepsilon \not\to \mathbb{P}^0$ in the weak adapted topology. Let us revisit Example 4.2 (see Figure 6 too). We have

$$I(\mathbb{P}^\varepsilon) = \frac{1}{2}\left[\delta_{(\varepsilon,1/2\delta_1+1/2\delta_0)} + \delta_{(-\varepsilon,1/2\delta_{-1}+1/2\delta_0)}\right],$$

which converges to $\tilde{\nu} := \frac{1}{2}\left[\delta_{(0,1/2\delta_1+1/2\delta_0)} + \delta_{(0,1/2\delta_{-1}+1/2\delta_0)}\right]$. On the other hand

$$I(\mathbb{P}^0) = \delta_{(0,1/4\delta_1+1/2\delta_0+1/4\delta_{-1})} \neq \tilde{\nu}.$$

Since $I(\mathbb{P}^\varepsilon)\not\to I(\mathbb{P}^0)$ we conclude that $\mathbb{P}^\varepsilon\not\to\mathbb{P}^0$ in the weak adapted topology.

The following result can be initially skipped. It will be useful in the applications, but we refer to the appendix for its proof.

Lemma 4.31. *Let $\sigma : \mathbb{R} \to [0,1]$ any continuous and increasing function such that $\lim_{z\to-\infty}\sigma(z) = 0$ and $\lim_{z\to\infty}\sigma(z) = 1$. Then $\{\mu_n\}_n \subseteq P(\mathbb{X}_1 \times \mathbb{X}_2)$ converges to μ in the weak adapted topology if and only if for all $h \in C_b(\mathbb{X}_1), g \in C_b(\mathbb{X}_2)$ it holds*

$$\int h(x_1)\sigma\left(\int g(z)(\mu_n)_{x_1}(dz)\right)\mu_n(dx) \to \int h(x_1)\sigma\left(\int g(z)\mu_{x_1}(dz)\right)\mu(dx).$$

Second, we establish a crucial property of the embedding operator J. The message is very important, as it provides a first evidence that the weak adapted topology is *natural* (and not just some arbitrary strengthening of the usual weak topology). The result is non-trivial, since in general J is not weakly-continuous.

Proposition 4.32. *We have*

(i) *$K \subseteq \mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2)$ is tight if and only if $J(K)$ is tight (ie. relatively compact wrt. the weak topology on $\mathcal{P}(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))$).*

(ii) *$K \subseteq \mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2)$ is relatively compact in the weak adapted topology if and only if K is tight and $\overline{J(K)} \subseteq$ Range(J).*

Proof. In Lemma A.3 we show that K tight implies $J(K)$ tight. Noting that $I(J(K)) = K$ and that I is continuous, we have that $J(K)$ tight implies K tight.

If K is relatively compact in the weak adapted topology, then it must be tight, and $\overline{J(K)} \subseteq$ Range(J) follows from definition of the relative topology on Range(J). For the converse direction, denote $\tilde{K}$ the closure of K in the weak adapted topology. Notice that $\overline{J(K)}$ is compact, since K tight implies $J(K)$ tight. Since J is a homeomorphism into its range (in the weak adapted topology), and $J(\tilde{K}) \subseteq \overline{J(K)}$, then $\tilde{K}$ must be compact in the weak adapted topology. □

The study of compact sets is out of the scope of this lecture notes, needing the notion of modulus of continuity by Eder. However, as an illustration, we can give the following simple statement:

Lemma 4.33. *Suppose $\mathbb{X}_1$ is sigma-compact and let $K \subseteq \mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2)$ be weakly compact (i.e. weakly closed and tight) and such that*

$\{x_1 \mapsto \mathbb{P}_{x_1} : \mathbb{P} \in K\} \subseteq C(\mathbb{X}_1; \mathcal{P}(\mathbb{X}_2))$ *is equicontinuous, and $\{\mathbb{P}_{x_1} : \mathbb{P} \in K\}$ is tight for each x_1.*

Then K is compact in the adapted weak topology.

Proof. If $(\mathbb{P}_n)_n \subseteq K$, then there is a subsequence (which we relabel) such that $\mathbb{P}_n \to \mathbb{Q}$ weakly and $\mathbb{Q} \in K$. By tightness and sigma-compactness, we find an increasing sequence $(K^1_m)_m$ of compact sets in $\mathbb{X}_1$ such that $p_1(\mathbb{P}_n)(K^1_m) \geq 1 - 1/m$ for all n, and $\cup_m K^1_m = \mathbb{X}_1$. By Arzela-Ascoli Theorem, and a diagonalization argument, there is a subsequence (which we relabel) and a continuous map $x_1 \mapsto f(x_1) \in \mathcal{P}(\mathbb{X}_1)$ such that $x_1 \mapsto (\mathbb{P}_n)_{x_1}$ converges to f uniformly on each K^1_m. Exercise: Check that $f(x_1) = \mathbb{Q}_{x_1}$ ($p^1(\mathbb{Q})$-a.s. x_1), and taking $F \in C_b(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))$ check that

$$\int F(x_1, (\mathbb{P}_n)_{x_1})\mathbb{P}_n(dx) \to \int F(x_1, \mathbb{Q}_{x_1})\mathbb{Q}(dx).$$

By Remark 4.29 we conclude $\mathbb{P}_n \to \mathbb{Q}$ in weak adapted topology. □

4.3.1. *Metrizing the weak adapted topology.* We now turn to the question of how to metrize the weak adapted topology.

Let D_i be any *bounded metric* which is *compatible* (i.e. which induces) with the topology on $\mathbb{X}_i$, and is complete. Associated to D_2 we may define a 1-Wasserstein distance on $\mathcal{P}(\mathbb{X}_2)$ via

$$\mathcal{P}(\mathbb{X}_2)^2 \ni (m,p) \mapsto \mathcal{W}_1(m,p) := \inf_{q\in\mathsf{Cpl}(p,m)} \int_{\mathbb{X}_2^2} D_2(z_1,z_2)q(dz_1,dz_2).$$

Remark 4.34. It can be seen that this metric induces the weak topology on $\mathcal{P}(\mathbb{X}_2)$, since D_2 is bounded. This can be obtained as a neat application of the Kantorovic-Rubinstein theorem, see also Theorem 1.53.

Accordingly, we can define a metric on $\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2)$ via

$$(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))^2 \ni ((x_1, m), (y_1, p)) \mapsto \tilde{D}((x_1, m), (y_1, p)) := D_1(x_1, y_1) + \mathcal{W}_1(m, p).$$

This metric induces the product topology on $\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2)$.

Finally, we can define a metric (actually a Wasserstein distance) on $\mathcal{P}(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))$: endowing $\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2)$ with the bounded compatible metric $\tilde{D}$ we consider the 1-Wasserstein distance

$$\begin{aligned}\mathcal{P}(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))^2 \ni (P, Q) \mapsto \tilde{\mathcal{W}}_1(P, Q) &:= \inf_{q \in \mathsf{Cpl}(P,Q)} \int_{(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))^2} \tilde{D}((x_1, m), (y_1, p)) q(d(x_1, m), d(y_1, p)) \\ &= \inf_{q \in \mathsf{Cpl}(P,Q)} \int_{(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))^2} \{D_1(x_1, y_1) + \mathcal{W}_1(m, p)\}\, q(d(x_1, m), d(y_1, p)).\end{aligned}$$

The following observation is crucial:

Remark 4.35. Applying the above formula to $P = J(\mu)$ and $Q = J(\nu)$ we obtain

$$\tilde{\mathcal{W}}_1(J(\mu), J(\nu)) = \inf_{q \in \mathsf{Cpl}(p_1(\mu), p_1(\nu))} \int_{(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))^2} \left\{D_1(x_1, y_1) + \mathcal{W}_1(\mu_{x_1}, \nu_{y_1})\right\} q(dx_1, dy_1).$$

The reason is that, if $q \in \mathsf{Cpl}(J(\mu), J(\nu))$, then $q(d(x_1, m), d(y_1, p)) = \tilde{q}(dx_1, dy_1)\delta_{\mu_{x_1}, \nu_{y_1}}(dm, dp)$, where $\tilde{q} \in \mathsf{Cpl}(p_1(\mu), p_1(\nu))$ is the marginal of q in the (x_1, y_1) component. Indeed, the measure $J(\mu)$ is supported on the graph of the function $x_1 \mapsto \mu_{x_1}$ (resp. $J(\nu)$ on $y_1 \mapsto \mu_{y_1}$), which implies that q must be supported on the set $\{(x_1, \mu_{x_1}, y_1, \nu_{y_1}) : x_1, y_1 \in \mathbb{X}_1\}$. All in all we deduce that $\tilde{\mathcal{W}}(J(\mu), J(\nu))$ is equal to the value of the bicausal transport problem (Pbc) between μ and ν for the cost function $c((x_1, x_2), (y_1, y_2)) = D_1(x_1, y_1) + D_2(x_2, y_2)$, thanks to the dynamic programming principle and Proposition 4.13.

Definition 4.36. *The function* $\mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2) \times \mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2) \to \mathbb{R}$ *given by*

$$(\mu, \nu) \mapsto \mathcal{AW}(\mu, \nu) := \inf_{q \in \mathsf{Cpl}_{bc}(\mu, \nu)} \int_{(\mathbb{X}_1 \times \mathbb{X}_2)^2} \{D_1(x_1, y_1) + D_2(x_2, y_2)\}\, q(dx_1, dx_2, dy_1, dy_2)$$

is called the adapted Wasserstein distance.

Of course we first should check that the term *distance* is well-deserved:

Lemma 4.37. *The adapted Wasserstein distance* $\mathcal{AW}$ *defines a metric on* $\mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2)$.

Proof. Since the actual Wasserstein distance is bounded from above by $\mathcal{AW}$, we have $\mathcal{AW}(\mu, \nu) = 0 \iff \mu = \nu$. Since bicausality is a symmetric condition and the D_i are symmetric, $\mathcal{AW}(\mu, \nu) = \mathcal{AW}(\nu, \mu)$. To finish the proof we just need to justify a "glueing argument"(cf. Lemma 1.48) which then would conduce to the triangle inequality just like for Wasserstein distances. Suppose $q \in \mathsf{Cpl}_{bc}(\mu, \nu)$ and $\tilde{q} \in \mathsf{Cpl}_{bc}(\nu, \eta)$. Exercise: find a suitable element $\hat{q} \in \mathsf{Cpl}_{bc}(\mu, \eta)$ built from q and $\tilde{q}$, and use it to complete the proof of the triangle inequality. □

Proposition 4.38. *The adapted Wasserstein distance* $\mathcal{AW}$ *metrizes (ie. induces) the weak adapted topology on* $\mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2)$.

Proof. If follows immediately from the definition of the weak adapted topology, the operator J, Remark 4.35 and Lemma 4.37. □

Remark 4.39. Unlike for Wasserstein spaces, it may happen that the metric space $(\mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2), \mathcal{AW})$ is incomplete. (Exercise: Provide an example of such situation for $\mathbb{X}_t = \mathbb{R}$.) Nevertheless, the weak adapted topology is Polish, ie. it is separable and there exist a compatible complete metric inducing this topology. (Exercise: Prove that it is separable.) It turns out that the complete metric space $(\mathcal{P}(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2)), \tilde{\mathcal{W}}_1)$ is the completion of $(\mathcal{P}(\mathbb{X}_1 \times \mathbb{X}_2), \mathcal{AW})$.

Remark 4.40. The extension of the adapted Wasserstein distance to $\mathbb{N} > 2$ and to unbounded metrics, is immediate:

$$\mathcal{P}(\mathsf{X}) \times \mathcal{P}(\mathsf{X}) \mapsto (\mu, \nu) \mapsto \mathcal{AW}_p(\mu, \nu)^p := \inf_{q \in \mathsf{Cpl}_{bc}(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{X}} \left\{ \sum_{i=1}^{N} D_i(x_i, y_i)^p \right\} q(dx, dy).$$

As in the classical case, convergence in $\mathcal{AW}_p$ is the same as convergence in $\mathcal{AW}$ (equiv. in weak adapted topology) plus convergence of moments of order p. The definition of the weak adapted topology is more cumbersone. In probabilistic notation, if $X \sim \mu \in \mathcal{P}_p(\mathsf{X})$, then the embedding J is defined as follows: Let $R_N^\mu := X_N$ and recursively $R_{t-1}^\mu := \mathrm{Law}(X_{t-1}, \mathrm{Law}_{X_1,\dots,X_{t-1}}(R_t^\mu))$. Then $J(\mu) := R_1^\mu$. The weak adapted topology is then the initial topology of J, where the range space is given with the suitable Wasserstein (distance) topology.

4.4. **Applications.** We illustrate here a few applications of the ideas developed so far. First we recall another kind of optimal transport problem, the *weak optimal transport* that already featured in our discussion of stretched Brownian motion in (WT), and show how the weak adapted topology provides the correct framework to study such problems. Second, we go back to our motivating examples and show that the adapted Wasserstein distances provide the correct notion of *closeness* that can guarantee the stability of the various optimization problem wrt. the reference model. It turns out that the language of weak optimal transport is also useful here. We finally show a very different application where we use the DPP of bicausal optimal transport to derive a celebrated functional inequality by Talagrand.

4.4.1. *Weak optimal transport.* In the classical optimal transport problem between the marginals $\mathbb{P} \in \mathcal{P}(\mathbb{X})$ and $\mathbb{Q} \in \mathcal{P}(\mathbb{Y})$, a *linear* cost criterion is minimized:

$$\mathsf{Cpl}(\mathbb{P}, \mathbb{Q}) \ni q \mapsto \int c(x, y) q(dx, dy).$$

Weak optimal transport is a generalization where certain kind of non-linear cost functions are considered, namely

$$\mathsf{Cpl}(\mathbb{P}, \mathbb{Q}) \ni q \mapsto \int C(x, q_x) p_1(q)(dx),$$

where

$$C : \mathbb{X} \times \mathcal{P}(\mathbb{Y}) \to \mathbb{R} \cup \{+\infty\}.$$

Evidently, setting $C(x, m) := \int_{\mathbb{Y}} c(x, y) m(dy)$ allows to consider classical optimal transport as a special case of weak optimal transport (so the terminology is misleading: weak optimal transport is the *stronger* one).

Using the tools from the previous parts we can easily obtain existence and duality for weak optimal transport:

Theorem 4.41. *Suppose that $C : \mathbb{X} \times \mathcal{P}(\mathbb{Y}) \to \mathbb{R} \cup \{+\infty\}$ is lower bounded, measurable, and for each $x \in \mathbb{X}$ the function $m \mapsto C(x, m)$ is convex and lower semicontinuous. Then*

(1) the weak optimal transport problem

$$\inf \left\{ \int C(x, q_x) \mathbb{P}(dx) : \ q \in \mathsf{Cpl}(\mathbb{P}, \mathbb{Q}) \right\} \tag{WOT}$$

admits at least one minimizer;

(2) there is no duality gap:

$$\mathit{value}(\mathrm{WOT}) = \sup_{\psi \in C_b(\mathbb{Y})} \int \psi(y) \mathbb{Q}(dy) - \int \varphi[\psi](x) \mathbb{P}(dx),$$

where $\varphi[\psi](x) := \sup_{m \in \mathcal{P}(\mathbb{Y})} \{ \int \psi(y) m(dy) - C(x, m) \}$.

Proof. Let Γ be the subset of $\mathcal{P}(\mathbb{X}\times\mathcal{P}(\mathbb{Y}))$ such that $M \in \Gamma$ iff $I(M) \in \mathsf{Cpl}(\mathbb{P},\mathbb{Q})$, where I is the intensity operator of Definition 4.27. We first extend Problem (WOT) by considering

$$\inf\left\{\int C(x,m)M(dx,dm) :\ M \in \Gamma\right\}. \tag{WOText}$$

Clearly value(WOText) $\leq$ value(WOT), since $J(q) \in \Gamma$ if $q \in \mathsf{Cpl}(\mathbb{P},\mathbb{Q})$. By Lemma A.3, since $I(\Gamma)$ is tight, then so is Γ. Since Γ is also easily seen to be closed, then it must be weakly compact too. By the lower semicontinuity statement in Lemma 4.42 below, which we may use since the first marginal of the elements in Γ is fixed and equals to $\mathbb{P}$, we conclude that Problem (WOText) admits an optimizer M. Let now $q := I(M) \in \mathsf{Cpl}(\mathbb{P},\mathbb{Q})$. Then by Jensen inequality

$$\begin{aligned}\int_{\mathbb{X}} C(x,q_x)\mathbb{P}(dx) &= \int_{\mathbb{X}} C\left(x, \int_{\mathcal{P}(\mathbb{Y})} m\, M_x(dm)\right)\mathbb{P}(dx)\\ &\leq \int C dM,\end{aligned}$$

and we conclude that value(WOText) = value(WOT) and q is optimal for (WOT).

For duality, we assume for simplicity that $\mathbb{Y}$ is compact. We first rewrite (WOText) as

$$\inf_{M:\, p_1(M)=\mathbb{P}} \sup_{\psi\in C_b(\mathbb{Y})} \int \psi d\mathbb{Q} + \int\left(C(x,m) - \int \psi(y)m(dy)\right)M(dx,dm).$$

The functional in the inf sup is linear and continuous in ψ and convex and lower semicontinuous in M, thanks to Lemma 4.42 below. The set $\{M :\ p_1(M) = \mathbb{P}\} \subseteq \mathcal{P}(\mathbb{X}\times\mathcal{P}(\mathbb{Y}))$ is compact, since it is tight and closed. By the minimax theorem we can exchange inf and sup, which readily gives the desired result. □

In the previous proof we used the following result. Its proof can be skipped.

Lemma 4.42. *Let $\mathbb{A},\mathbb{B}$ be Polish spaces. Suppose that $F : \mathbb{A}\times\mathbb{B} \to \mathbb{R}\cup\{+\infty\}$ is lower bounded, measurable, and for each $a \in \mathbb{A}$ the function $b \mapsto F(a,b)$ is lower semicontinuous. If $(M_n)_n \subseteq \mathcal{P}(\mathbb{A}\times\mathbb{B})$ weakly converges to M, and $p_1(M_n) = \mu$ for all n, then*

$$\liminf_n \int F dM_n \geq \int F dM.$$

Proof. By monotone convergence arguments, we may assume that F is bounded. By [Str11, Lemma 9.1.4(ii)-(iii)] we may take a compatible metric D on $\mathbb{B}$, such that $U(\mathbb{B})$, the space of bounded uniformly continuous functions, is separable. Observe then that $L_k(a,b) = \inf_{\bar b}\{F(a,\bar b) + kD(b,\bar b)\}$ increases pointwise to F as $k \to \infty$, and since L_k is D-Lipschitz in b, we have $L_k(a,\cdot) \in U(\mathbb{B})$. Using this, we can derive

$$\int F dR = \sup_{L\in\mathcal{L}} \int L dR,$$

with R any Borel probability measure, and $\mathcal{L}$ the set of bounded functions L such that $L(\cdot,b)$ is (analytically) measurable, $L(a,\cdot)$ is continuous, and $L \leq F$ [5]. All in all, this shows that we may further assume that F is (analytically) measurable in the first variable and uniformly continuous in the second one.

Consider the map $\mathbb{A} \ni a \mapsto F(a,\cdot) \in U(\mathbb{B})$. The $U(\mathbb{B})$ space is second countable (as a separable metric space) and the previous map is (analytically) measurable. Moreover, there exists $\mathbb{A} \ni a \mapsto \tilde F(a,\cdot) \in U(\mathbb{B})$ Borel measurable, such that $\mu(\{a :\ F(a,\cdot) = \tilde F(a,\cdot)\}) = 1$. We may apply [Kec95, Theorem 13.11], according to which there is a stronger Polish topology on $\mathbb{A}$ which preserves the Borel sets (so also measurable functions) and such that the map $\mathbb{A} \ni a \mapsto \tilde F(a,\cdot)$ is continuous. As a consequence $\tilde F$ is easily seen jointly continuous on $\mathbb{A}\times\mathbb{B}$, with this stronger topology on the first variable and the original one

[5] $L_k(\cdot,b)$ is not necessarily Borel measurable, but only analytically measurable in general, by a measurable selection argument. This technical fact poses no difficulty, and the reader may simply drop the term *analytically* from this proof at no risk.

on the second variable. We now prove that $\int F dM_n \to \int F dM$. By definition, if $f : \mathbb{A} \to \mathbb{R}$ is continuous in the new topology, it is also Borel in the original one. Assuming further that such f is bounded, and taking $g \in C_b(\mathbb{B})$, we deduce

$$\int fg\, dM_n = \int f(x)\mu(dx) \int g(m)p^2(M_n)(dm) \to \int f(x)\mu(dx) \int g(m)p^2(M)(dm) = \int fg\, dM.$$

As a consequence of [EK09, Proposition 4.6(b) (p.115)] we obtain that $\int \tilde{F} dM_n \to \int \tilde{F} dM$, and conclude since $\int \tilde{F} dR = \int F dR$ whenever $p_1(R) = \mu$. □

We remark that this is only the beginning of the story of weak optimal transport. For instance, studying necessary and sufficient optimality conditions (in the same spirit of the monotonicity principle), stability w.r.t. $(\mathbb{P}, \mathbb{Q})$, existence of dual optimizers, etc. are all challenging subjects which are only partially understood.

We now collect a few applications of weak optimal, which hopefully provide the reader with the impression of how useful this theory is. Equipped with a monotonicity principle, the applications become even more remarkable.

Strassen's Theorem on martingales. Let $S = \{S_i : i = 0, 2\}$ be a two-step martingale. By Jensen's inequality, we know that $\text{Law}(S_1) \leq_c \text{Law}(S_2)$ in the convex order. Strassen's Theorem provides us with the reverse implication. So let $\mu, \nu \in \mathbb{R}^N$ be such that $\mu \leq_c \nu$. For simplicity we further assume that ν (hence also μ) has a compact support, and denote by $\mathbb{X}$ its compact convex hull. We will show that there is a martingale S with first marginal μ and second marginal ν.

Let $\mathbb{X} \times \mathcal{P}(\mathbb{X}) \ni (x, m) \mapsto C(x, m)$ be defined by $C(x, m) = 0$ if $\int ym(dy) = x$ and $C(x, m) = +\infty$ otherwise. Consider the weak optimal transport problem with marginals $\mathbb{Q} := \mu$, $\mathbb{P} := \nu$ and cost functional C. If we can show that the value of this problem is zero, and that the problem is attained, this immediately provides the desired martingale. Indeed, we would get the existence of $q \in \mathsf{Cpl}(\mu, \nu)$ with $x = \int yq_x(dy)$ for μ-a.e. x.

First we observe that C fulfils the requirements of Theorem 4.41. By Point (1) therein, we do obtain the existence of an optimal q. Towards applying Point (2) therein, we observe

$$\varphi[\psi](x) = \sup\left\{ \int \psi dm : \ x = \int y dm \right\}.$$

Clearly $\psi \leq \varphi[\psi]$ and $\varphi[\psi](\cdot)$ is concave. Thus

$$\int \psi d\nu - \int \varphi[\psi] d\mu \leq \int \varphi[\psi] d\nu - \int \varphi[\psi] d\mu \leq 0,$$

by the convex order assumption. Hence the dual problem has value zero, and so does the primal problem too.

The convex Kantorovic-Rubinstein formula. Now we let $C(x, m) := |x - \int ym(dy)|$, and so consider the weak optimal transport problem

$$\inf\left\{ \int \left| x - \int yq_x(dy) \right| \mathbb{P}(dx) : \ q \in \mathsf{Cpl}(\mathbb{P}, \mathbb{Q}) \right\}. \tag{4.21}$$

First observe that if $\tilde{\varphi}$ is convex and 1-Lipschitz, and $q \in \mathsf{Cpl}(\mathbb{P}, \mathbb{Q})$, then by Jensen's inequality:

$$\begin{aligned} \int \left| x - \int yq_x(dy) \right| \mathbb{P}(dx) &\geq \int \left[\tilde{\varphi}(x) - \tilde{\varphi}\left(\int yq_x(dy) \right) \right] \mathbb{P}(dx) \\ &\geq \int \left[\tilde{\varphi}(x) - \int \tilde{\varphi}(y) q_x(dy) \right] \mathbb{P}(dx) \\ &= \int \tilde{\varphi} d(\mathbb{P} - \mathbb{Q}). \end{aligned} \tag{4.22}$$

Now let us assume for simplicity that both $\mathbb{P}, \mathbb{Q}$ are supported in a convex compact subset $\mathbb{X}$ of $\mathbb{R}^N$. We then readily get that C satisfies the hypotheses of Theorem 4.41. Towards

applying duality (Point (2) therein), we compute

$$\begin{aligned}\varphi[\psi](x) &= \sup_m \left\{ \int \psi dm - \left| x - \int y m(dy) \right| \right\} \\ &= \sup_z \left\{ \sup_{m: \int y m(dy) = z} \int \psi dm - |x - z| \right\} \\ &= \sup_z \left\{ \tilde{\psi}(z) - |x - z| \right\},\end{aligned}$$

with $\tilde{\psi}(z) := \sup_{m: \int y m(dy) = z} \int \psi dm$. Clearly $\tilde{\psi}$ is a concave majorant of ψ, and so $\varphi[\psi]$ is also a majorant of ψ. But then $(x, z) \mapsto \tilde{\psi}(z) - |x - z|$ is jointly concave, and so $\varphi[\psi]$ is concave. Finally, it is clear that $\varphi[\psi]$ is 1-Lipschitz. This all shows that the dual problem to (4.24) is bounded from above by

$$\sup_{f \text{ concave, 1-Lip}} \int f d(\mathbb{Q} - \mathbb{P}) = \sup_{\tilde{\varphi} \text{ convex, 1-Lip}} \int \tilde{\varphi} d(\mathbb{P} - \mathbb{Q}).$$

Since there is no duality gap, the above and (4.22) show

$$\inf \left\{ \int \left| x - \int y q_x(dy) \right| \mathbb{P}(dx) : \ q \in \mathsf{Cpl}(\mathbb{P}, \mathbb{Q}) \right\} = \sup_{\tilde{\varphi} \text{ convex, 1-Lip}} \int \tilde{\varphi} d(\mathbb{P} - \mathbb{Q}). \tag{4.23}$$

This is the so-called *convex Kantorovic-Rubinstein formula*. We shall make use of this identity in Section 4.4.2 below. Incidentally, one can derive Strassen's theorem on martingales from this identity too.

Brenier-Strassen Theorem. The existence of a Brenier map from μ to ν requires μ to be a somewhat regular measure. On the other hand, for the existence of a martingale, Strassen's Theorem requires convex order but otherwise no regularity of the marginal measures. Hence, one could expect that *if both 'maps' and 'martingales' are allowed*, then no conditions at all on μ or ν should be imposed. This is made precise by the Brenier-Strassen theorem.

We consider here $C(x, m) := |x - \int y m(dy)|^2$, leading to the weak optimal transport problem

$$\begin{aligned}&\inf \left\{ \int \left| x - \int y q_x(dy) \right|^2 \mu(dx) : \ q \in \mathsf{Cpl}(\mu, \nu) \right\} \\ &= \inf_{\eta \leq_c \nu} W_2^2(\mu, \eta).\end{aligned} \tag{4.24}$$

The equality holds, since for $q \in \mathsf{Cpl}(\mu, \nu)$ given, define $T(x) := \int y q_x(dy)$, and remark that $T(\mu) \leq_c \nu$, since the coupling $\tilde{q}$ defined in duality by $\int f(x, y) \tilde{q}(dx, dy) = \int f(T(x), y) q_x(dy) \mu(dx) \in \mathsf{Cpl}(T(\mu), \nu)$ is by construction a martingale coupling (alternatively use Jensen's inequality to observe that $\int \varphi dT(\mu) \leq \int \varphi d\nu$ for all convex φ). Moreover,

$$\int \left| x - \int y q_x(dy) \right|^2 \mu(dx) = \int |x - T(x)|^2 \mu(dx).$$

Conversely, given any $\eta \leq_c \nu$ there is a martingale π connecting η and ν so that the optimal coupling for W_2 can be extended using this martingale to a coupling between μ and ν. Moreover, if T is optimal between μ and η it follows that

$$W_2^2(\mu, \eta) = \int |x - T(x)|^2 \mu(dx) = \int \left| x - \int y \pi_{T(x)}(dy) \right|^2 \mu(dx).$$

Therefore, solving this optimization problem is in fact asking to find the optimal measure η minimizing $W_2(\mu, \eta)$ among all measures $\eta \leq_c \nu$.

Armed with a monotonicity principle for weak optimal transport it is possible to show the following characterisation without any assumption on μ.

Theorem 4.43. *η^* is optimal for* (4.24) *iff there exists a convex function φ with 1-Lipschitz gradient $\nabla\varphi$ such that $\nabla\varphi(\mu) = \eta^*$.*

4.4.2. *Stability in stochastic optimization and mathematical finance.* We revisit the motivating examples of optimal stopping, utility maximization and superhedging, in the light of what we have learnt about adapted weak topologies / adapted Wasserstein distances. For the sake of concreteness we take $N = 2$ throughout, though the general case is not much more difficult.

Optimal Stopping. Recall from Example 4.1 that weak convergence alone is not enough for the continuity of the optimal value of an optimal stopping problem.

Write $AC(\mathbb{X}_1 \times \mathbb{X}_2)$ for the set of all two-step processes $(L_t)_{t=1}^2$ which are adapted[6], bounded and satisfy that $x \mapsto L_t(x)$ is continuous for each $t \in \{1, 2\}$. Write $v^L(\mathbb{P})$ for the corresponding value function, given that the process S follows the law $\mathbb{P}$, i.e.

$$v^L(\mathbb{P}) := \inf\{\mathbb{E}_{\mathbb{P}}[L_\tau(S)] : \tau \in \{1, 2\} \text{ is } S_1\text{-measurable}\}.$$

Definition 4.44. *The* optimal stopping topology *is the coarsest topology which makes the functions*

$$\mathbb{P} \mapsto v^L(\mathbb{P})$$

continuous for all $(L_t)_{t=1}^2 \in AC(\mathbb{X}_1 \times \mathbb{X}_2)$.

Theorem 4.45. *The optimal stopping topology is equal to the weak adapted topology. In other words, $\mathbb{P}_n \to \mathbb{P}$ in the weak adapted topology iff $v^L(\mathbb{P}_n) \to v^L(\mathbb{P})$ for all $L \in AC(\mathbb{X}_1 \times \mathbb{X}_2)$. Further, we have*

$$|v^L(\mathbb{P}) - v^L(\mathbb{Q})| \leq \inf\left\{\int \max\{|L_1(x_1) - L_1(y_1)|, |L_2(x_1, x_2) - L_2(y_1, y_2)|\} q(dx_1, dx_2, dy_1, dy_2)\right\}, \tag{4.25}$$

where the infimum runs over $q \in \mathsf{Cpl}_{bc}(\mathbb{P}, \mathbb{Q})$. In particular, if L is further K-Lipschitz with respect to a metric $D((x_1, x_2), (y_1, y_2)) = D_1(x_1, y_1) + D_2(x_2, y_2)$, then

$$|v^L(\mathbb{P}) - v^L(\mathbb{Q})| \leq 2K\mathcal{AW}_1(\mathbb{P}, \mathbb{Q}), \tag{4.26}$$

with $\mathcal{AW}_1$ as in Definition 4.36.

The interest of the previous result is twofold. First of all, it gives another characterization of the weak adapted topology, and stresses its natural character, by showing its equivalence with the more pedestrian optimal stopping topology. In particular, the weak adapted topology cannot be really improved upon, as far as continuity/stability of optimal stopping is concerned. And second, it provides quantitative upper bounds for the misspecification of the reference model in optimal stopping problems.

Proof of Theorem 4.45. Without loss of generality $v^L(\mathbb{P}) \geq v^L(\mathbb{Q})$. Fixing ε take τ which is ε-optimal for $v^L(\mathbb{Q})$. Further let $q \in \mathsf{Cpl}_{bc}(\mathbb{P}, \mathbb{Q})$ arbitrary.

For $u \in [0, 1]$ define

$$\begin{aligned}\sigma(x_1, u) &:= \inf\{t \in \{1, 2\} : q(\tau(y_1) \leq t | x_1) \geq u\} \\ &= \inf\{t \in \{1, 2\} : q(\tau(y_1) \leq t | x_1, x_2) \geq u\},\end{aligned}$$

[6]Namely $L_1 : \mathbb{X}_1 \to \mathbb{R}$ and $L_2 : \mathbb{X}_1 \times \mathbb{X}_2 \to \mathbb{R}$.

where equality holds since q is causal. We then have that

$$\begin{aligned}\int_{[0,1]}\int L_{\sigma(x_1,u)}(x_1,x_2)dqdu &= \int_{[0,1]}\int [L_1(x_1)1_{q(\tau(y_1)=1|x_1,x_2)\geq u} + L_2(x_1,x_2)1_{q(\tau(y_1)=1|x_1,x_2)<u}]dqdu\\ &= \int [L_1(x_1)q(\tau(y_1)=1|x_1,x_2) + L_2(x_1,x_2)q(\tau(y_1)=2|x_1,x_2)]dq\\ &= \int [L_1(x_1)1_{\tau(y_1)=1} + L_2(x_1,x_2)1_{\tau(y_1)=2}]dq\\ &= \int [L_{\tau(y_1)}(x_1,x_2)]dq.\end{aligned}$$

As further $\sigma(\cdot,u)$ is x_1-measurable for every fixed $u\in[0,1]$ we have

$$v^L(\mathbb{P}) \leq \int_{[0,1]}\int L_{\sigma(x_1,u)}(x_1,x_2)dqdu,$$

and therefore

$$\begin{aligned}v^L(\mathbb{P}) - v^L(\mathbb{Q}) &\leq \int [L_{\tau(y_1)}(x_1,x_2) - L_{\tau(y_1)}(y_1,y_2)]dq + \varepsilon\\ &\leq \int \max\{|L_1(x_1)-L_1(y_1)|\,,\,|L_2(x_1,x_2)-L_2(y_1,y_2)|\}dq + \varepsilon,\end{aligned}$$

establishing Inequality (4.25). Then (4.26) follows easily, if L_1, L_2 are furthermore K-Lipschitz.

We now prove that the weak adapted topology is finer than the optimal stopping topology. We can metrize the latter by $\mathcal{AW}$ by taking D_t to be compatible bounded metrics. Now assume that $\mathcal{AW}(\mathbb{P}_n,\mathbb{P})\to 0$ and that $q_n\in\mathsf{Cpl}(\mathbb{P}_n,\mathbb{P})$ is less than $1/n$ away from attaining the infimum for $\mathcal{AW}(\mathbb{P}_n,\mathbb{P})$. Then $q_n\to q$ weakly, where $q\in\mathsf{Cpl}(\mathbb{P},\mathbb{P})$ is the identity coupling of $\mathbb{P}$ with itself. (A Monge coupling between q_n and q is given by $q_n\circ(((x_1,x_2),(y_1,y_2))\mapsto((x_1,x_2),(y_1,y_2),(y_1,y_2),(y_1,y_2)))^{-1}$.) Because

$$(x_1,x_2,y_1,y_2)\mapsto \max\{|L_1(x_1)-L_1(y_1)|\,,\,|L_2(x_1,x_2)-L_2(y_1,y_2)|\}$$

is a continuous bounded function, we get that

$$\int \max\{|L_1(x_1)-L_1(y_1)|\,,\,|L_2(x_1,x_2)-L_2(y_1,y_2)|\}dq_n$$

converges to

$$\int \max\{|L_1(x_1)-L_1(y_1)|\,,\,|L_2(x_1,x_2)-L_2(y_1,y_2)|\}dq = 0.$$

By (4.25), this finishes the argument.

We now prove that the optimal stopping topology is finer than the weak adapted topology. We first remark that, denoting $\bar{L}=-L$, we have

$$-v^L(\mathbb{P}) = \int \max\left\{\bar{L}_1(x_1)\,,\,\int \bar{L}_2(x_1,\bar{x}_2)\mathbb{P}_{x_1}(d\bar{x}_2)\right\}p_1(\mathbb{P})(dx_1),$$

by a measurable selection argument (known as dynamic programming principle in this context). We make now educated guesses for $\bar{L}$. If $f\in C_b(\mathbb{X}_1), h\in C_b(\mathbb{X}_2)$, with f non-negative, we let $\bar{L}^a_1=0$, $\bar{L}^a_2(x_1,x_2)=f(x_1)h(x_2)$, and $L^b_1=0$, $\bar{L}^b_2(x_1,x_2)=f(x_1)[h(x_2)-1]$. Then

$$\begin{aligned}-v^{L^a}(\mathbb{P}) &= \int f(x_1)\max\left\{0\,,\,\int h(\bar{x}_2)\mathbb{P}_{x_1}(d\bar{x}_2)\right\}p_1(\mathbb{P})(dx_1),\\ -v^{L^b}(\mathbb{P}) &= \int f(x_1)\max\left\{0\,,\,\int [h(\bar{x}_2)-1]\mathbb{P}_{x_1}(d\bar{x}_2)\right\}p_1(\mathbb{P})(dx_1).\end{aligned}$$

If $\sigma(z):=\max\{0,\min\{1,z\}\}$ then $\sigma(z)=\max\{0,z\}-\max\{0,z-1\}$, and so

$$v^{L^b}(\mathbb{P}) - v^{L^a}(\mathbb{P}) = \int f(x_1)\sigma\left(\int h(\bar{x}_2)\mathbb{P}_{x_1}(d\bar{x}_2)\right)p_1(\mathbb{P})(dx_1).$$

Thus if $\mathbb{P}_n \to \mathbb{P}$ in the optimal stopping topology, we also have

$$\int f(x_1)\sigma\left(\int h(\bar{x}_2)(\mathbb{P}_{x_1})_n(d\bar{x}_2)\right)p_1(\mathbb{P}_n)(dx_1) \to \int f(x_1)\sigma\left(\int h(\bar{x}_2)\mathbb{P}_{x_1}(d\bar{x}_2)\right)p_1(\mathbb{P})(dx_1).$$

As σ is a continuous and increasing function such that $\lim_{z\to-\infty}\sigma(z) = 0$ and $\lim_{z\to\infty}\sigma(z) = 1$, we conclude by Lemma 4.31 that $\mathbb{P}_n \to \mathbb{P}$ in the weak adapted topology. □

Utility Maximization. Recall from Example 4.2 that $U : \mathbb{R} \to \mathbb{R}$ is a strictly increasing and concave, and that we had introduced (the one-dimensional version of)

$$u(\mathbb{P}) := \max\left\{\mathbb{E}_{\mathbb{P}}[U(\pi \cdot (S_2 - S_1))] \,:\, \pi \text{ is } S_1\text{-measurable, } \|\pi\| \leq 1\right\}.$$

We take $\mathbb{X}_t = \mathbb{R}^k$ with D_t the metric obtained from the Euclidean norm, so defining $\mathcal{AW}_1$ per Definition 4.36.

Theorem 4.46. *Suppose U is furthermore L-Lipschitz. Then*

$$|u(\mathbb{P}) - u(\mathbb{Q})| \leq L\,\mathcal{AW}_1(\mathbb{P},\mathbb{Q}).$$

Proof. Without loss of generality let us suppose that $0 \leq u(\mathbb{P}) - u(\mathbb{Q})$. Let π be a $(1/n)$ optimizer for $u(\mathbb{P})$, meaning that $u(\mathbb{P}) \leq \mathbb{E}_{\mathbb{P}}[U(\pi(X_2 - X_1))] + 1/n$. Consider now $q \in \mathsf{Cpl}_{bc}(\mathbb{P},\mathbb{Q})$ which is $(1/n)$-optimal for $\mathcal{AW}_1(\mathbb{P},\mathbb{Q})$, meaning this time that

$$\int (|x_1 - y_1| + |x_2 - y_2|)\, q(dx_1, dx_2, dy_1, dy_2) \leq \mathcal{AW}_1(\mathbb{P},\mathbb{Q}) + 1/n.$$

We define

$$\hat{\pi}(y_1) := \int \pi(x_1) q_{y_1}(dx_1) = \int \pi(x_1) q_{y_1,y_2}(dx_1),$$

where equality occurs by causality. In particular $\hat{\pi}$ is a feasible (usually sub-optimal) element for the problem $u(\mathbb{Q})$. Hence

$$\begin{aligned}
-1/n + u(\mathbb{P}) - u(\mathbb{Q}) &\leq \int \left[U(\pi(x_1)(x_2 - x_1)) - U(\hat{\pi}(y_1)(y_2 - y_1))\right] dq \\
&= \int \left[U(\pi(x_1)(x_2 - x_1)) - U\left(\int \pi(x_1)(y_2 - y_1) q_{y_1,y_2}(dx_1, dx_2)\right)\right] dq \\
&\leq \int \left[U(\pi(x_1)(x_2 - x_1)) - U(\,\pi(x_1)(y_2 - y_1)\,)\right] dq \\
&\leq L \int (|x_1 - y_1| + |x_2 - y_2|)\, dq,
\end{aligned}$$

where the first inequality is by Jensen and the second one by the Lipschitz assumption plus the boundedness of π. Hence

$$u(\mathbb{P}) - u(\mathbb{Q}) \leq 1/n + L/n + L\,\mathcal{AW}_1(\mathbb{P},\mathbb{Q}),$$

and we conclude sending $n \to \infty$. □

Observe that the key in the above argument was that we could *project* the adapted process π into another close-by adapted process $\hat{\pi}$ by means of a causal transport. Adaptedness of $\hat{\pi}$ would not be true, if we had used a non-causal transport.

Super-hedging. The setting is that of Example 4.3, but in the mutidimensional generalization (as for utility maximization).

If $C(S_1, S_2)$ is an option, then "super-replicating it at cost at most m under model $\mathbb{P}$" means finding $\pi = \pi(S_1)$ such that $C \leq m + \pi[S_2 - S_1]$ ($\mathbb{P}$-a.s.). A smoothed-out version of it, under model $\mathbb{Q}$ rather than $\mathbb{P}$, is to find some $\tilde{\pi} = \tilde{\pi}(S_1)$ such that $\mathbb{E}_{\mathbb{Q}}[(C(S_1, S_2) - m - \tilde{\pi}[S_2 - S_1])_+)]$ is as small as possible.

As in the previous part, we take $\mathbb{X}_t = \mathbb{R}^k$ with D_t the metric obtained from the Euclidean norm, so defining $\mathcal{AW}_1$ per Definition 4.36.

Proposition 4.47. *Suppose C is L-Lipschitz, and that C can be super-replicated at cost at most m under model $\mathbb{P}$ by a bounded $\pi = \pi(S_1)$. Then $\tilde{\pi} = \tilde{\pi}(S_1)$ exists such that $\|\tilde{\pi}\|_\infty \leq \|\pi\|_\infty$ and*

$$\mathbb{E}_{\mathbb{Q}}[(C(S_1, S_2) - m - \tilde{\pi}[S_2 - S_1])_+)] \leq (L + \|\pi\|_\infty)\, \mathcal{AW}_1(\mathbb{P}, \mathbb{Q}). \tag{4.27}$$

Proof. Take q_n which is $1/n$-optimal for $\mathcal{AW}_1(\mathbb{P}, \mathbb{Q})$ and define

$$\tilde{\pi}_n(y_1) := \int \pi(x_1)(q_n)_{y_1}(dx_1) = \int \pi(x_1)(q_n)_{y_1, y_2}(dx_1),$$

with equality by causality. Denote $\mu_n = (C(S) - m - \tilde{\pi}_n[S_2 - S_1])(\mathbb{Q})$ and $\nu = (C(S) - m - \pi[S_2 - S_1])(\mathbb{P})$. By definition $0 = \mathbb{E}_{\mathbb{P}}[(C(S_1, S_2) - m - \pi[S_2 - S_1])_+)]$. As the function $\tilde{\varphi} := [\cdot]_+$ is convex and 1-Lipschitz, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[(C(S_1, S_2) - m - \tilde{\pi}_n[S_2 - S_1])_+)] - \mathbb{E}_{\mathbb{P}}[(C(S_1, S_2) - m - \pi[S_2 - S_1])_+)] &= \int \tilde{\varphi} d(\mu_n - \nu) \\ &\leq d^w(\mu_n, \nu),\end{aligned}$$

as follows from the convex Kantorovic-Rubinstein formula of Equation (4.23), and where $d^w(\mu_n, \nu)$ is the value of the weak transport problem with cost $\mathbb{R} \times \mathcal{P}(\mathbb{R}) \ni (x, R) \mapsto \left|x - \int y R(dy)\right|$.

On the other hand, let $(X, Y) \sim q_n$ and denote $a(X) = C(X) - m - \pi(X_1)[X_2 - X_1]$, $b(Y) = C(Y) - m - \tilde{\pi}_n(Y_1)[Y_2 - Y_1]$, so by definition

$$\begin{aligned} d^w(\mu_n, \nu) &\leq \mathbb{E}_{q_n}\left[\left|b(Y) - \mathbb{E}_{q_n}[a(X)|b(Y)]\right|\right] \\ &= \mathbb{E}_{q_n}\left[\left|\mathbb{E}_{q_n}[C(Y) - C(X)|Y] - \mathbb{E}_{q_n}\left[\tilde{\pi}_n(Y_1)[Y_2 - Y_1] - \pi(X_1)[X_2 - X_1]|Y\right]\right|\right] \\ &\leq \mathbb{E}_{q_n}\left[|C(Y) - C(X)|\right] + \mathbb{E}_{q_n}\left[|\pi(X_1)[Y_2 - Y_1] - \pi(X_1)[X_2 - X_1]|\right], \end{aligned}$$

where we used the tower property and Jensen's inequality. Hence

$$d^w(\mu_n, \nu) \leq (L + \|\pi\|_\infty)\mathbb{E}_{q_n}\left[\|X - Y\|\right] \leq (L + \|\pi\|_\infty)(1/n + \mathcal{AW}_1(\mathbb{P}, \mathbb{Q})),$$

and putting all together we find

$$\mathbb{E}_{\mathbb{Q}}[(C(S_1, S_2) - m - \tilde{\pi}_n[S_2 - S_1])_+)] \leq (L + \|\pi\|_\infty)(1/n + \mathcal{AW}_1(\mathbb{P}, \mathbb{Q})). \tag{4.28}$$

Clearly $\|\tilde{\pi}_n\|_\infty \leq \|\pi\|_\infty$, and in particular $\{\tilde{\pi}_n\}_n$ is bounded in $L^2(\mathbb{Q})$. By Mazur's Lemma there is a sequence $\{\hat{\pi}_n\}_n$ and $\tilde{\pi}$ in the closed convex hull of $\{\tilde{\pi}_n\}_n$, such that $\hat{\pi}_n \to \tilde{\pi}$ strongly in $L^2(\mathbb{Q})$. Thus $\tilde{\pi}$ is S_1-measurable and $\|\tilde{\pi}\|_\infty \leq \|\pi\|_\infty$. An application of Fatou's Lemma and the convexity of $[\cdot]_+$, together with (4.28), yield

$$\mathbb{E}_{\mathbb{Q}}[(C(S_1, S_2) - m - \tilde{\pi}[S_2 - S_1])_+)] \leq (L + \|\pi\|_\infty)\mathcal{AW}_1(\mathbb{P}, \mathbb{Q}).$$

□

4.4.3. *Stretched Brownian motion revisited.* Recall the standard stretched Brownian motion from Example 2.50. We considered a Brownian motion B with $B_0 \sim \alpha$, and an increasing map f s.t. $f(\alpha * \gamma) = \nu$. Then, the standard stretched Brownian motion is defined as

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

where $\mathcal{F}$ denotes the natural Brownian filtration. We showed that stretched Brownian motion is a standard stretched Brownian motion on each of its irreducible components.

Lemma 4.48. *Let M be stretched Brownian motion. There holds*

$$Law(B - B_0, M) \in \mathsf{Cpl}_{bc}(\mathbb{W}, Law(M)).$$

Proof. Let us consider standard stretched Brownian first. Since $M_t = f_t(B_t)$ by construction, M is adapted to the Brownian filtration $\mathcal{F}$. Hence, $\mathrm{Law}(B - B_0, M)$ is causal. For the converse causality, observe that $\{B_{t+h} - B_t, h \geq 0\}$ is independent of $\{B_s, s \leq t\}$, so that given $\{M_s, s \leq t\}$ it follows that $\{B_s - B_0, s \leq t\}$ and $\{M_s, s \leq 1\}$ are independent.

The case of stretched Brownian motion follows from first conditioning on the value of M_0 and then following the reasoning above. □

We claim that the stretched Brownian motion can be interpreted as an adapted projection of Brownian motion onto the set of martingale laws with prescribed initial and terminal marginal. Indeed, recall that stretched Brownian motion is the optimizer to

$$\sup\left\{\mathbb{E}\int_0^1 \sigma_t(M_t)dt : M_t = M_0 + \int_0^t \sigma_s dB_s, M_0 \sim \mu, M_1 \sim \nu\right\}. \tag{4.29}$$

Pick any admissable martingale M and observe (as in the proof of Theorem 2.53) that by integration by parts

$$\mathbb{E}\left[\int_0^1 \sigma_t dt\right] = \mathbb{E}[M_1 \cdot (B_1 - B_0)] = -\frac{1}{2}\mathbb{E}[(M_1 - (B_1 - B_0))^2] + \frac{1}{2}\mathbb{E}M_1^2 + \frac{1}{2}\mathbb{E}[(B_1 - B_0)^2],$$

so that the optimization problem (4.29) is equivalent to (writing M_μ^ν for the set of all continuous martingales with initial law μ and terminal law ν)

$$\inf_{q\in \mathsf{Cpl}_{bc}(\mathbb{W},\mathrm{Law}(M)),M\in\mathsf{M}_\mu^\nu} \mathbb{E}_q[(M_1 - (B_1 - B_0))^2]. \tag{4.30}$$

4.4.4. *Existence of dynamic Cournot-Nash equilibria.* We will define a notion of a game, and its equilibrium solution, built on causal optimal transport. For that matter, we introduce the "cost function"

$$F : \mathsf{X}\times\mathsf{Y}\times\mathcal{P}(\mathsf{Y}) \to \mathbb{R}. \tag{4.31}$$

Recall that $\mathsf{X} = \Pi_{t=1}^N \mathbb{X}_t$ and $\mathsf{Y} = \Pi_{t=1}^N \mathbb{Y}_t$. We fix $\eta \in \mathcal{P}(\mathsf{X})$ and $K \subseteq \mathcal{P}(\mathsf{Y})$.

Definition 4.49 (Cournot-Nash)**.** *A causal transport $\hat{q} \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$ with $p_1(q) = \eta$ and $p^2(q) \in K$ is called* dynamic Cournot-Nash equilibrium *for a type-η population and action-ambiguity set K, if*

$$\hat{q} \in \arg\min \int_{\mathsf{X}\times\mathsf{Y}} F(x, y, p^2(\hat{q}))\, q(dx, dy), \tag{4.32}$$

where minimization is done over causal transports $q \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$ with $p_1(q) = \eta$ and $p^2(q) \in K$.

Let us explain the *game interpretation* of Definition 4.49. We think of a continuum of negligible players. Each player has at every time $t = 1, \dots, N$ a *type* $\in \mathbb{X}_t$, so X is the state space of types, and is private knowledge of the player. However, the distribution of types in the population of player, namely η, is known and fixed in advance. A player must choose at each time t an action/strategy in $\mathbb{Y}_t$, and so overall an element of Y must be chosen. The cost to a player of type x of choosing action y depends on the decisions of the whole population of players, and so if the distribution of actions in the population is described by ν, then this cost is $F(x, y, \nu)$. Unique η, the distribution ν is not fixed in advance, but it must belong to K. Finally, if we suppose that the type of a player is only revealed progressively in time, so that the associated action cannot anticipate the future evolution of the type, this forces the coupling of type and action to be causal. Clearly Equation (4.32) is a fixed-point condition. It states: the joint distribution $\hat{q}$ of types and actions should be cost minimizing, given the implied action distribution $\nu := p^2(\hat{q})$.

We fix $\mathcal{W}$ a metric compatible with weak convergence on $\mathcal{P}(\mathsf{Y})$, and denote $\mathsf{Cpl}_c(\eta, K) := \cup_{\nu\in K}\mathsf{Cpl}_c(\eta, \nu)$. We have

Proposition 4.50. *Assume*

- *F is bounded and $\mathcal{F}(\cdot, \cdot, \nu)$ is continuos for each ν;*
- *For each $K_1 \subseteq \mathsf{X}, K_2 \subseteq \mathsf{Y}$ compact, there is a constant ℓ such that*

$$\sup_{x\in K_1, y\in K_2} |F(x, y, \nu) - F(x, y, \bar{\nu})| \le \ell\, \mathcal{W}(\nu, \bar{\nu}).$$

Then a dynamic Cournot-Nash equilibrium for a type-η population and action-ambiguity set K exists.

Proof. By Lemma 4.11, the set $S := \mathsf{Cpl}_c(\eta, K)$ is weakly compact. For $\tilde{q} \in S$ we define

$$\Phi(\tilde{q}) := \arg\min \int_{\mathsf{X}\times\mathsf{Y}} F(x, y, p^2(\tilde{q}))\, q(dx, dy),$$

where minimization is over $q \in S$. Clearly $\Phi(\tilde{q})$ is a convex and compact set, and by Theorem 4.10, it is also non-empty. One says that Φ is a convex-, nonempy-, compact-valued correspondence. If we can prove the existence of $\hat{q} \in S$ such that $\hat{q} \in \Phi(\hat{q})$, we conclude the proof. This is the content of the celebrated Kakutani-Fan-Glicksberg fixed-point Theorem (see [**?**, Corollary 17.55]). However we still need to verify one hypothesis: that the graph of Φ is closed. So let $\tilde{q}_n \to \tilde{q}$, $q_n \to q$, and $q_n \in \Phi(\tilde{q}_n)$. The aim is to prove that $q \in \Phi(\tilde{q})$. Let $m \in S$ arbitrary, so that by definition

$$\int_{\mathsf{X}\times\mathsf{Y}} F(x, y, p^2(\tilde{q}_n))\, q_n(dx, dy) \le \int_{\mathsf{X}\times\mathsf{Y}} F(x, y, p^2(\tilde{q}_n))\, m(dx, dy).$$

We only explain how to pass to the limit for the l.h.s. since convergence of the r.h.s. is more direct. This clearly finishes the proof. For simplicity denote $\tilde{F}_n$ and $\tilde{F}$ as shorthand for $F(x, y, p^2(\tilde{q}_n))$ and $F(x, y, p^2(\tilde{q}))$. Then

$$\int \tilde{F}_n q_n - \int \tilde{F} dq = \int (\tilde{F}_n - \tilde{F}) dq_n + \int \tilde{F} d(q_n - q).$$

The term $\int \tilde{F} d(q_n - q) \to 0$ since $\tilde{F}$ is continuous and bounded. On the other hand, as $(q_n)_n$ is tight, for $\varepsilon > 0$ there are K_1, K_2 compact such that $q_n(K_1 \times K_2) \ge 1 - \varepsilon$ for all n. Thus

$$\left| \int (\tilde{F}_n - \tilde{F}) dq_n \right| \le \int |\tilde{F}_n - \tilde{F}| dq_n \le 2\varepsilon \|F\|_\infty + \int_{K_1\times K_2} |\tilde{F}_n - \tilde{F}| dq_n,$$

and by assumption $\int_{K_1\times K_2} |\tilde{F}_n - \tilde{F}| dq_n \le \ell \mathcal{W}(p^2(\tilde{q}_n), p^2(\tilde{q}))$, which goes to zero. So $\int \tilde{F}_n q_n \to \int \tilde{F} dq$ indeed. □

4.4.5. *Talagrand functional inequality.* If $\mathbb{X}$ is a Polish space, and $\mu, \nu \in \mathcal{P}(\mathbb{X})$, then the relative entropy of μ with respect to ν is defined by

$$H(\mu|\nu) := \int \log\left(\frac{d\mu}{d\nu}\right) d\mu = \int \log\left(\frac{d\mu}{d\nu}\right) \frac{d\mu}{d\nu} d\nu$$

if $\mu \ll \nu$, and otherwise $H(\mu|\nu) := +\infty$. Let us fix a compatible metric $d_{\mathbb{X}}$ on $\mathbb{X}$ and use it to define the p-Wasserstein distance $\mathcal{W}_p^{\mathbb{X}}$.

Definition 4.51. *We say that $\nu \in \mathcal{P}_p(\mathbb{X})$ satisfies (Talagrand's transport-information) $\mathcal{T}_p(c)$ inequality if*

$$\forall \mu \in \mathcal{P}_p(\mathbb{X}) : \ \mathcal{W}_p^{\mathbb{X}}(\mu, \nu) \le \sqrt{2cH(\mu|\nu)}.$$

We now prove that $\mathcal{T}_p(c)$ inequalities *tensorize*, by means of causal optimal transport arguments. On $\mathsf{X} := \mathbb{X}^N$ we consider the ℓ^p metric

$$d_{\mathsf{X}}(x, y) = \sqrt[p]{d_{\mathbb{X}}(x_1, y_1)^p + \cdots + d_{\mathbb{X}}(x_N, y_N)^p},$$

with associated p-Wasserstein distance $\mathcal{W}_p^{\mathsf{X}}$

Proposition 4.52. *Suppose $\nu \in \mathcal{P}(\mathbb{X})$ satisfies the $\mathcal{T}_p(c)$ inequality for some $1 \le p \le 2$. Then $\nu^{\otimes N} \in \mathcal{P}_p(\mathsf{X})$ likewise satisfies the $\mathcal{T}_p(c)$ inequality, namely*

$$\forall \tilde{\mu} \in \mathcal{P}_p(\mathsf{X}) : \ \mathcal{W}_p^{\mathsf{X}}(\tilde{\mu}, \nu^{\otimes N}) \le \sqrt{2cH(\tilde{\mu}, \nu^{\otimes N})}.$$

For the proof we will need the decomposition property of the relative entropy. To this end, note that for two Polish spaces $\mathbb{A}, \mathbb{B}$, a map $T : \mathbb{A} \to \mathbb{B}$ and $\mathbb{P} \in \mathcal{P}(\mathbb{A})$ we can disintegrate

$$\mathbb{P}(dx) = \mathbb{P}_y(dx) T(\mathbb{P})(dy),$$

so that $\mathbb{P}_y = \mathbb{P}(\cdot|T = y)$.

Lemma 4.53. *If $\mathbb{A}, \mathbb{B}$ are Polish spaces, $T : \mathbb{A} \to \mathbb{B}$ Borel measurable, and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{A})$, then using the notation from above*

$$H(\mathbb{P}|\mathbb{Q}) = H(T(\mathbb{P})|T(\mathbb{Q})) + \int H(\mathbb{P}_y|\mathbb{Q}_y)dT(\mathbb{P})(y).$$

Proof. Wlog we can assume that $\mathbb{P} \ll \mathbb{Q}$ (otherwise there is a set of positive $T(\mathbb{P})$ measure such that $\mathbb{P}_y$ is not absolutely continuous wrt $\mathbb{Q}_y$ so that both sides are infinite). Denote $\rho = d\mathbb{P}/d\mathbb{Q}$, $\bar{\rho} = dT(\mathbb{P})/dT(\mathbb{Q})$, and $\rho_y = d\mathbb{P}_y/d\mathbb{Q}_y$. For any measurable f we obtain by disintegration

$$\int f d\mu = \int f\rho(x)\mathbb{P}_y(dx)dT(\mathbb{P})(dy) = \int f(x)\rho_y(x)\bar{\rho}(x)\mathbb{P}_y(dx)dT(\mathbb{P})(dy).$$

so that on the set $\{T = y\}$ on which $\mathbb{P}_y$ is concentrated there holds $\rho(x) = \rho_y(x)\bar{\rho}(y)$. Taking logarithm and integrating wrt $\mathbb{P}(dx) = \mathbb{P}_y(dx)T(\mathbb{P})(dy)$ yields the result. □

Proof of Proposition 4.52. By Lemma 4.53, we find

$$H(\tilde{\mu}|\nu^{\otimes N}) = H(p_1(\tilde{\mu})|\nu) + \int H(\tilde{\mu}^{2:N}_{x_1}|\nu^{\otimes(N-1)})p_1(\tilde{\mu})(dx),$$

where $\tilde{\mu}^{2:N}_{x_1}$ is shorthand for the distribution of $x_2, \dots, x_N$ under $\tilde{\mu}$ given x_1. If we iterate this argument $N-2$ times, we find

$$H(\tilde{\mu}|\nu^{\otimes N}) = H(p_1(\tilde{\mu})|\nu) + \sum_{t=1}^{N-1} \int H(\tilde{\mu}_{x_1,\dots,x_t}|\nu)\tilde{\mu}(dx_1, \dots, x_t).$$

By assumption $H(\tilde{\mu}_{x_1,\dots,x_t}|\nu) \geq \mathcal{W}_p^{\mathbb{X}}(\tilde{\mu}_{x_1,\dots,x_t}, \nu)^2/(2c)$, and so

$$2cH(\tilde{\mu}|\nu^{\otimes N}) \geq \mathcal{W}_p^{\mathbb{X}}(p_1(\tilde{\mu}), \nu)^2 + \sum_{t=1}^{N-1} \int \mathcal{W}_p^{\mathbb{X}}(\tilde{\mu}^{x_1,\dots,x_t}, \nu)^2\tilde{\mu}(dx_1, \dots, x_t).$$

Using that $p/2 \leq 1$ and Jensen's inequality we find

$$\begin{aligned}\sqrt{2cH(\tilde{\mu}|\nu^{\otimes N})} &\geq \left(\int \left[\mathcal{W}_p^{\mathbb{X}}(p_1(\tilde{\mu}), \nu)^2 + \sum_{t=1}^{N-1} \mathcal{W}_p^{\mathbb{X}}(\tilde{\mu}_{x_1,\dots,x_t}, \nu)^2\right]^{p/2} d\tilde{\mu}\right)^{1/p} \\ &\geq \left(\int \mathcal{W}_p^{\mathbb{X}}(p_1(\tilde{\mu}), \nu)^p + \sum_{t=1}^{N-1} \mathcal{W}_p^{\mathbb{X}}(\tilde{\mu}_{x_1,\dots,x_t}, \nu)^p d\tilde{\mu}\right)^{1/p},\end{aligned}$$

where we also used that $\sqrt[2]{a_1^2 + \cdots + a_N^2} \geq \sqrt[p]{a_1^p + \cdots + a_N^p}$ if $p \leq 2$. We now define de bicausal transport problem (between $\tilde{\mu}$ and $\nu^{\otimes N}$) with cost function d_{X}^p:

$$\mathcal{AW}_p(\tilde{\mu}, \nu^{\otimes N})^p := \inf_{q \in \mathsf{Cpl}_{bc}(\tilde{\mu}, \nu^{\otimes N})} \int \sum_{t=1}^{N} d_{\mathbb{X}}(x_t, y_t)^p q(dx, dy),$$

which by the dynamic programming principle of Proposition 4.13 is equal to

$$\mathcal{W}_p^{\mathbb{X}}(p_1(\tilde{\mu}), \nu)^p + \sum_{t=1}^{N-1} \int \mathcal{W}_p^{\mathbb{X}}(\tilde{\mu}_{x_1,\dots,x_t}, \nu)^p \tilde{\mu}(dx_1, \dots, x_t).$$

All in all

$$\sqrt{2cH(\tilde{\mu}|\nu^{\otimes N})} \geq \mathcal{AW}_p(\tilde{\mu}, \nu^{\otimes N}) \geq \mathcal{W}_p(\tilde{\mu}, \nu^{\otimes N}),$$

since the infimum for $\mathcal{W}_p$ is computed over a larger set. □

Remark 4.54. The constant c appearing in Proposition 4.52 is independent of N, which immediately suggests infinite-dimensional counterparts. The most renowned transport-information inequalities are $\mathcal{T}_1(c)$ and $\mathcal{T}_2(c)$, the first of which is equivalent to the concentration of measure phenomenon. Talagrand's original application concerned $\nu = \gamma_1$, the 1-dimensional standard Gaussian distribution, for which the $\mathcal{T}_2(1)$ inequality can be verified by calculus arguments, and so by Proposition 4.52 the same holds for multidimensional Gaussians.

4.5. **Exercises.**

Problem 45. (Bicausal Couplings)

Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^3)$ be defined as

$$\mu := \frac{1}{4}\left(\delta_{(-1,-1,-2)} + \delta_{(-1,-1,0)} + \delta_{(1,1,0)} + \delta_{(1,1,2)}\right) \quad \text{and}$$
$$\nu := \frac{1}{4}\left(\delta_{(0,-1,-2)} + \delta_{(0,-1,0)} + \delta_{(0,1,0)} + \delta_{(0,1,2)}\right).$$

a) Find a coupling of μ and ν that is causal but not bicausal.
b) Find a bicausal coupling of μ and ν that is not the product coupling.

Problem 46. (Causal Map)

Let $N \geq 1$ and $\mu \in \mathcal{P}(\mathsf{X}^N)$. For all $1 \leq i \leq N$, let $T_i : \mathsf{X}^i \to \mathsf{Y}$ be a measureable map. Moreover, we define

$$T : \mathsf{X}^N \to \mathsf{Y}^N$$
$$(x_1, ..., x_N) \mapsto (T_1(x_1), T_2(x_1, x_2), ..., T_n(x_1, ..., x_n)).$$

Show that $(id, T)_{\#}\mu$ is a causal coupling of μ and $T_{\#}\mu$.

Problem 47. (Characterization of bicausal couplings)

Prove Proposition 4.12.

Problem 48. (On Proposition 4.15)

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^2)$ be defined as

$$\mu := \frac{1}{25}\left(9\delta_{(-1,-1)} + 6\delta_{(-1,1)} + 6\delta_{(1,-1)} + 4\delta_{(1,1)}\right) \quad \text{and}$$
$$\nu := \frac{1}{4}\left(\delta_{(-1,-1)} + \delta_{(-1,1)} + \delta_{(1,-1)} + \delta_{(1,1)}\right).$$

Moreover, we consider the cost function $c(x_1, x_2, y_1, y_2) := \mathbb{1}_{(x_1,x_2)\neq(y_1,y_2)}$.

Show that the value of the causal transport problem is strictly smaller than the value of the bicausal transport problem. Why does this not contradict Proposition 4.15?

Problem 49. (Stationary bicausal optimal transport)

Consider $\mathbb{X}_t = \mathbb{R} = \mathbb{Y}_t$ for $t \in \mathbb{N}$ and $\mathsf{X} = \mathbb{R}^{\mathbb{N}} = \mathsf{Y}$. Fix $c : \mathbb{R}\times\mathbb{R} \to [0, \infty)$ and $\beta \in [0, 1]$. Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^{\mathbb{N}})$ we define $\mathsf{Cpl}_{bc}(\mu, \nu)$ in the obvious way and consider

$$BOT_{c,\beta}(\mu, \nu) := \inf_{q\in\mathsf{Cpl}_{bc}(\mu,\nu)} \int \left(\sum_{t\in\mathbb{N}} \beta^t c(x_t, y_t)\right) q(dx, dy).$$

Justify that

$$BOT_{c,\beta}(\mu, \nu) = \inf_{q\in\mathsf{Cpl}(p_1(\mu),p_1(\nu))} \int \{c(x_0, y_0) + \beta BOT_{c,\beta}(\mu_{x_0}, \nu_{x_0})\} q(dx_0, dy_0),$$

where μ_{x_0} denotes the law of $(x_1, x_2, x_3, \dots)$ given x_0 under μ and so for ν.

Problem 50. (Markov Property is not closed)

For $N = 3$ consider $\mu_\varepsilon = 1/2\delta_{(1,\varepsilon,1)} + 1/2\delta_{(-1,-\varepsilon,-1)}$ for $\varepsilon \geq 0$. Show that μ_ε has the Markov property if and only of $\varepsilon > 0$. Show as well that $\mathcal{AW}(\mu_\varepsilon, \mu_0) \to 0$ as $\varepsilon \to 0$, so concluding that the Markov property is not closed under convergence in adapted Wasserstein distance.

Problem 51. (Turning a Borel set into an open set)

Let (X, τ) be a Polish space (i.e. separable and the topology τ is generated by some complete metric) and $A \subseteq X$ be a Borel measurable set. Define a stronger/finer topology $\bar{\tau}$ on X such that

- $(X, \bar{\tau})$ is still a Polish space, and
- $A \in \bar{\tau}$.

Appendix A. Some useful measure theory

Theorem A.1 (Disintegration theorem)**.** *Let* X, Y *be Polish spaces, and let* q *be a finite Borel measure on* $\mathsf{X}\times\mathsf{Y}$*. Denote by* μ *and* ν *the marginals of* q *on the first and second factor respectively. Then, there exist two measurable families of probability measures* $(q_x)_{x\in\mathsf{X}}$ *and* $(q_y)_{y\in\mathsf{Y}}$ *such that*

$$q(dx, dy) = q_x(dy)\mu(dx) = q_y(dx)\nu(dy).$$

For a proof we refer to e.g. Theorem 5.1.3 in [AGS08] or III-70 in [DM78].

Recall Definition 4.27 on the intensity operator I. With slight abuse of notation we still call I the operator from $\mathcal{P}(\mathcal{P}(\mathsf{Y}))$ to $\mathcal{P}(\mathsf{Y})$, obtained when we ignore the first coordinate of $\mathcal{X}\times\mathcal{P}(\mathcal{Y})$ resp. $\mathcal{X}\times\mathcal{Y}$. The following is a classical result which can be found in [Szn91, p. 178, Ch. II].

Lemma A.2. *A set* $\mathcal{A}\subseteq\mathcal{P}(\mathcal{P}(\mathsf{Y}))$ *is tight if and only if the set of its intensities* $I(\mathcal{A})$ *is tight in* $\mathcal{P}(\mathsf{Y})$*.*

For our purposes, we need a refined form of Lemma A.2. Recall the embedding operator J from Definition 4.26.

Lemma A.3. $\Pi\subseteq\mathcal{P}(\mathsf{X}\times\mathsf{Y})$ *is tight if and only if* $J(\Pi)\subseteq\mathcal{P}(\mathsf{X}\times\mathcal{P}(\mathsf{Y}))$ *is tight. Similarly,* $\Lambda\in\mathcal{P}(\mathsf{X}\times\mathcal{P}(\mathsf{Y}))$ *is tight if and only if* $I(\Lambda)\subseteq\mathcal{P}(\mathsf{X}\times\mathsf{Y})$ *is tight.*

Proof. Say Γ is tight. Since continuous maps preserve relative compactness in Hausdorff spaces, we immediately deduce tightness of $I(\Lambda)$.

Say Π is tight. Then so are the sets $\Pi^{\mathsf{X}}\subseteq\mathcal{P}(\mathsf{X})$ and $\Pi^{\mathsf{Y}}\subseteq\mathcal{P}(\mathsf{Y})$ consisting respectively of the X- and Y-marginals of the elements in Π. Denote now respectively by $\Pi^{\mathsf{X}}_J\subseteq\mathcal{P}(\mathsf{X})$ and $\Pi^{\mathsf{Y}}_J\subseteq\mathcal{P}(\mathcal{P}(\mathsf{Y}))$ the set of X- and $\mathcal{P}(Y)$-marginals of the elements in $J(\Pi)$. Clearly $\Pi^{\mathsf{X}}_J=\Pi^{\mathsf{X}}$. By Lemma A.2, the set Π^{Y}_J is tight in $\mathcal{P}(\mathcal{P}(\mathsf{Y}))$ if and only if the set $I(\Pi^{\mathsf{Y}}_J)$ is tight in $\mathcal{P}(\mathsf{Y})$. However, if m is equal to the $\mathcal{P}(Y)$-marginal of $J(\pi)$, then $I(m)$ is equal to the Y-marginal of π. It follows that $I(\Pi^{\mathsf{Y}}_J)\subseteq\Pi^{\mathsf{Y}}$ is tight and hence so is Π^{Y}_J. Since the marginals of $J(\Pi)$ are tight, we conclude that $J(\Pi)$ itself is tight. Conversely, if $J(\Pi)$ is tight then $I(J(\Pi))=\Pi$ is tight, by the first paragraph.

Finally, if $I(\Lambda)$ is tight, then as in the previous paragraph, the set of X- and Y-marginals of $I(\Lambda)$ are tight, and then by Lemma A.2 the set of X- and $\mathcal{P}(\mathsf{Y})$-marginals of Λ are tight. We conclude that Λ is tight. □

Appendix B. Some useful analysis

We recall that a function $f:\mathsf{X}\to(-\infty,\infty]$ defined on a topological space X is called lower semicontinuous (lsc) if for each $c\in\mathbb{R}$ the set $\{f\le c\}$ is closed. If the topology is metrizable, then this notion is equivalent to the condition $\liminf_{x_n\to x} f(x_n)\ge f(x)$.

Lemma B.1. *Let* X *be a Polish space and* $f:\mathsf{X}\to(-\infty,\infty]$ *be lsc and bounded from below. Then*

$$\mathcal{P}(\mathsf{X})\ni\mu\mapsto\int f d\mu$$

is lsc w.r.t. the weak topology (and bounded from below).

Proof. Take d any metric compatible with the Polish topology, and put $f_n(x)=(\inf_y(f(y)+nd(x,y)))\wedge n$. Then, $f_n\in C_b(\mathsf{X})$ (even Lipschitz), $f_n\le f$ and $f\le\liminf f_n$ (here we need the lower bound and lsc of f) so that $f=\sup_n f_n$. Hence, if $\mu_n\to\mu$ weakly,

$$\liminf_n\int f d\mu_n\ge\sup_k\liminf_n\int f_k d\mu_n=\sup_k\int f_k d\mu=\int\sup_k f_k d\mu=\int f d\mu.$$

□

Appendix C. Pending proofs

Towards the proof of Lemma 4.31, we will need:

Lemma C.1. *Let $\mathcal{A}$ be a Polish space. Then the family of functions*

$$\left\{\mathcal{P}(\mathcal{A}) \ni \mu \mapsto G\left(\int_{\mathcal{A}} h_1 d\mu, \dots, \int_{\mathcal{A}} h_L d\mu\right) : \begin{array}{l} L \in \mathbb{N}, G \in C_b(\mathbb{R}^L) \\ (h_i)_{i\leq L} \subseteq C_b(\mathcal{A}) \end{array} \right\} \tag{C.1}$$

is convergence determining for the weak topology on $\mathcal{P}(\mathcal{P}(\mathcal{A}))$, that is, a sequence of probability measures $(\mu_n)_n$ in $\mathcal{P}(\mathcal{P}(\mathcal{A}))$ converges weakly to a probability measure $\mu \in \mathcal{P}(\mathcal{P}(\mathcal{A}))$ if and only if $\int F d\mu_n \to \int F d\mu$ for all F in (C.1).

This follows from the Stone-Weierstrass theorem in case of compact $\mathcal{A}$ and extends to general Polish spaces e.g. via compactification.

Lemma C.2. *Let $\mathcal{A}$ be a Polish space and $\sigma : \mathbb{R} \to [0,1]$ any continuous and increasing function such that $\lim_{z\to-\infty} \sigma(z) = 0$ and $\lim_{z\to\infty} \sigma(z) = 1$. Then the family of functions*

$$\left\{\mathcal{P}(\mathcal{A}) \ni \mu \mapsto \sigma\left(\int_{\mathcal{A}} h d\mu\right) : \ h \in C_b(\mathcal{A})\right\} \tag{C.2}$$

is convergence determining for the weak topology on $\mathcal{P}(\mathcal{P}(\mathcal{A}))$.

Proof. Let L, G, and $(h_i)_{i\leq L}$ as in (C.1). Moreover, let $m \in \mathbb{R}$ such that $|h_i| \leq m$ for all $1 \leq i \leq L$ and define $A := [-m, m]^L$. Then $A \subseteq \mathbb{R}^L$ is compact and satisfies

$$\left(\int h_1 d\mu, \dots, \int h_L d\mu\right) \in A \quad \text{for all } \mu \in \mathcal{P}(\mathcal{A}).$$

By the universal approximation result of Cybenko [Cyb89, Theorem 2], the set

$$\left\{x \mapsto \sum_{i=1}^m u_i \sigma(v_i \cdot x + w_i) : \begin{array}{l} m \in \mathbb{N}, (u_i)_{i\leq m} \subseteq \mathbb{R}, \\ (v_i)_{i\leq m} \subseteq \mathbb{R}^L, (w_i)_{i\leq m} \subseteq \mathbb{R} \end{array} \right\}$$

is dense in $C(A)$ w.r.t. the supremum norm. As a result, it is enough to replace G in (C.1) by functions of the form $x \mapsto \sum_{i=1}^m u_i \sigma(v_i \cdot x + w_i)$. Evaluating the latter function on the vector $x = (\int h_1 d\mu, \dots, \int h_L d\mu)$ yields

$$\begin{aligned} \sum_{i=1}^m u_i \sigma\left(\sum_{k=1}^L v_i^k \int h_k d\mu + w_i\right) &= \sum_{i=1}^m u_i \sigma\left(\int \left(\sum_{k=1}^{L+1} v_i^k h_k\right) d\mu\right) \\ &= \sum_{i=1}^m u_i \sigma\left(\int \bar{h}_i d\mu\right), \end{aligned}$$

upon defining $v_i^{L+1} := w_i$, $h_{L+1} := 1$, and finally $\bar{h}_i := \sum_{k=1}^{L+1} v_i^k h_k$ for every i. The result follows from Lemma C.1. □

Proof of Lemma 4.31. As $C_b(\mathbb{X}_1)$ is convergence determining for $\mathcal{P}(\mathbb{X}_1)$, and $\{\nu \mapsto \sigma(\int_{\mathbb{X}_2} h d\nu) : h \in C_b(\mathcal{X}_2)\}$ is, by Lemma C.2, convergence determining for $\mathcal{P}(\mathcal{P}(\mathbb{X}_2))$, it follows e.g. from [EK09, Proposition 4.6 (p.115)] that

$$\left\{(x_1, \nu) \mapsto h(x_1)\sigma\left(\int g(x_2)\nu(dx_2)\right) : h \in C_b(\mathbb{X}_1),\ g \in C_b(\mathbb{X}_2)\right\}, \tag{C.3}$$

is convergence determining for the weak topology on $\mathcal{P}(\mathbb{X}_1 \times \mathcal{P}(\mathbb{X}_2))$. By definition of J and the weak adapted topology, we conclude. □

References

[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.

[Bal00] Erik J. Balder. Lectures on Young measure theory and its applications in economics. *Rend. Istit. Mat. Univ. Trieste*, 31(suppl. 1):1–69, 2000. Workshop on Measure Theory and Real Analysis (Italian) (Grado, 1997).

[Bei12] M. Beiglböck. Cyclical monotonicity and the ergodic theorem. *submitted*, 2012.

[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.

[BNS19] Mathias Beiglböck, Marcel Nutz, and Florian Stebegg. Fine properties of the optimal skorokhod embedding problem. *arXiv preprint arXiv:1903.03887*, 2019.

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

[BVBHK17] Julio Backhoff-Veraguas, Mathias Beiglböck, Martin Huesmann, and Sigrid Källblad. Martingale benamou–brenier: a probabilistic perspective. *arXiv:1708.04869*, 2017.

[CGS10] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From knothe's transport to brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[CSS76] Stamatis Cambanis, Gordon Simons, and William Stout. Inequalities for ek (x, y) when the marginals are fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 36(4):285–294, 1976.

[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.

[Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[DM78] C. Dellacherie and P.-A. Meyer. *Probabilities and potential*, volume 29 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam, 1978.

[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.

[EK09] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[Gal18] Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

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

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

[HS18] Martin Huesmann and Florian Stebegg. Monotonicity preserving transformations of mot and sep. *Stochastic Processes and their Applications*, 128(4):1114–1134, 2018.

[HT19] Martin Huesmann and Dario Trevisan. A benamou–brenier formulation of martingale optimal transport. *Bernoulli*, 25(4A):2729–2757, 2019.

[Kal02] O. Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

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

[McC94] R. McCann. A convexity theory for interacting gases and equilibrum crystals. *PhD thesis, Princeton University*, 1994.

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

[PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[Roc97] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[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.

[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.

[Str11] Daniel W. Stroock. *Probability theory*. Cambridge University Press, Cambridge, second edition, 2011. An analytic view.

[Szn91] Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

[Tre16] Dario Trevisan. Well-posedness of multidimensional diffusion processes with weakly differentiable coefficients. *Electronic Journal of Probability*, 21, 2016.

[Váz07] Juan Luis Vázquez. *The porous medium equation: mathematical theory*. Oxford University Press, 2007.

[Vil03] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[Vil09] C. Villani. *Optimal Transport. Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.