EXISTENCE, DUALITY, AND CYCLICAL MONOTONICITY FOR WEAK TRANSPORT COSTS

J. BACKHOFF-VERAGUAS, M. BEIGLBÖCK, AND G. PAMMER

Abstract. The optimal weak transport problem has recently been introduced by Gozlan et. al. [19]. We provide general existence and duality results for these problems on arbitrary Polish spaces, as well as a necessary and sufficient optimality criterion in the spirit of cyclical monotonicity. As an application we extend the Brenier-Strassen Theorem of Gozlan-Juillet [17] to general probability measures on $\mathbb{R}^d$ under minimal assumptions.

Keywords: Optimal Transport, cyclical monotonicity, Brenier's Theorem, Strassen's Theorem, Weak transport costs, Duality.

1. Introduction

1.1. **Notation.** This article is concerned with the optimal transport problem for weak costs, as initiated by Gozlan et.al. [19]. To state it (see (1.1) below) we introduce some basic notation. On a Polish space Z the set of probability measures is denoted by $\mathcal{P}(Z)$. Denoting by $C_b(Z)$ the space of real-valued continuous bounded functions on Z, we use the probabilists terminology of 'weak convergence' for the weak topology that $C_b(Z)$ induces on $\mathcal{P}(Z)$. For Polish spaces X, Y and probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ we write $\Pi(\mu,\nu)$ for the set of all couplings on $X \times Y$ with marginals μ and ν. Given a coupling π on $X \times Y$ we denote a regular disintegration with respect to the first marginal by $(\pi_x)_{x\in X}$. We consider cost functionals of the form

$$C : X \times \mathcal{P}(Y) \to \mathbb{R} \cup \{+\infty\};$$

usually it is assumed that C is lower bounded, lower semicontinuity in an appropriate sense, and that $C(x,\cdot)$ is convex. With these ingredients, the weak transport problem is defined as

$$V(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_X C(x,\pi_x)\mu(dx). \tag{1.1}$$

1.2. **Literature.** The initial works of Gozlan et al. [19, 18] are mainly motivated by applications to geometric inequalities. Indeed, particular costs of the form (1.1) were already considered by Marton [23, 22] and Talagrand [33, 34]. Further papers directly related to [19] include [30, 29, 31, 15, 17]. Notably the weak transport problem (1.1) also yields a natural framework to investigate a number of related problems: it appears in the recursive formulation of the causal transport problem [3], in [1, 2, 11] it is used to provide a new perspective on (discrete time) martingale optimal transport, in [5] it is employed as a tool to study a martingale transport problem in continuous time.

1.3. **Main results.** Throughout X and Y are Polish spaces, $\mu \in \mathcal{P}(X)$, and $\nu \in \mathcal{P}(Y)$.

We will establish analogues of three fundamental facts in optimal transport theory: existence of optimizers, duality, and characterization of optimizers through c-cyclical monotonicity. We make the important comment, that these concepts (in particular existence and duality) have been previously studied for the weak transport problem. However, the results available so far may be too restrictive for certain applications.

Our goal is to establish these results at a level of generality that mimics the framework usually considered in the optimal transport literature (i.e. lower bounded, lower semicontinuous cost function). We emphasize that this extension is in fact required to treat specific examples of interest, cf. Section 1.3.4 below.

We briefly hint at the novel viewpoint which makes this extension possible: In a nutshell, the technicalities of the weak transport problem appear intricate and tedious since kernels $(\pi_x)_x$ are notoriously ill behaved with respect to weak convergence of measures on $\mathbb{P}(X \times Y)$. In the present paper we circumvent this difficulty by embedding $\mathbb{P}(X \times Y)$ into the bigger space $\mathbb{P}(X \times \mathbb{P}(Y))$. This idea is borrowed from the investigation of process distances (cf. [26, 4]) and will allow us to carry out proofs that closely resemble familiar arguments from classical optimal transport.

1.3.1. *Primal Existence.* As a first contribution we will establish in Section 2 the following basic existence results.

Theorem 1.1 (Existence I)**.** *Assume that $C\colon X \times \mathcal{P}(Y) \to \mathbb{R} \cup \{+\infty\}$ is jointly lower semicontinuous, bounded from below, and convex in the second argument. Then, the problem*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_X C(x, \pi_x) \mu(dx),$$

admits a minimizer.

Notably, Gozlan et.al. provide existence of minimizer under the assumption that $\pi \mapsto \int C(x, \pi_x)\, d\mu(x)$ is continuous on the set of all transport plans with first marginal μ, whereas our aim is to establish existence based on properties of the function C. We also note that Theorem 1.1 was first established by [2] in the case where X, Y are compact spaces.

In fact the assumptions of Theorem 1.1 may be more restrictive than they initially appear. Indeed, as the cost function defined in (1.5) below is not lower semicontinuous with respect to weak convergence, we will need to employ a refined version of Theorem 1.1 to carry out our application in Theorem 1.4 below.

Given a compatible metric d_Y on the Polish space Y, we write $\mathcal{P}^t_{d_Y}(Y)$ for the set of probability measures $\nu \in \mathcal{P}(Y)$ such that $\int d_Y(y, y_0)^t\, \nu(dy) < \infty$ for some (and then any) $y_0 \in Y$ and denote the t-Wasserstein metric on $\mathcal{P}^t_{d_Y}(Y)$ by $\mathcal{W}_t$ (see e.g. [35, Chapter 7]). In the sequel we make the important convention that, whenever we refer to $\mathcal{P}^t_{d_Y}(Y)$, it is assumed that this set is equipped with the topology generated by $\mathcal{W}_t$. On the other hand, regarding the Polish space X, we fix from now on a compatible bounded metric d_X.

Theorem 1.2 (Existence II)**.** *Assume that $\nu \in \mathcal{P}^t_{d_Y}(Y)$. Let $C\colon X \times \mathcal{P}^t_{d_Y}(Y) \to \mathbb{R} \cup \{+\infty\}$ be jointly lower semicontinuous with respect to the product topology on $X \times \mathcal{P}^t_{d_Y}(Y)$, bounded from below, and convex in the second argument. Then, the problem*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_X C(x, \pi_x) \mu(dx),$$

admits a minimizer.

We emphasize that Theorem 1.1 is a special case of Theorem 1.2. To see this, just take d_Y to be a compatible bounded metric. We also note that if C is strictly convex in the second argument and $V(\mu, \nu) < \infty$, then the minimizer $\pi^* \in \Pi(\mu, \nu)$ is unique. We report our proofs in Section 2.

1.3.2. *Duality.* We fix a compatible metric d_Y on Y and introduce the space

$$\Phi_{b,t} := \{\psi : Y \to \mathbb{R} \text{ cont. s.t. } \exists a, b, \ell \in \mathbb{R}, y_0 \in Y, \ell \le \psi(\cdot) \le a + b d_Y(y_0, \cdot)^t\},. \tag{1.2}$$

To each $\psi \in \Phi_{b,t}$ we associate the function

$$R_C\psi(x) := \inf_{p \in \mathcal{P}^t_{d_Y}(Y)} p(\psi) + C(x, p). \tag{1.3}$$

We remark that $R_C\psi(\cdot)$ is universally measurable if C is measurable ([12, Proposition 7.47]) and so the integral $\mu(R_C\psi)$ is well defined for all $\mu \in \mathcal{P}(Y)$ if C is lower-bounded. We will compare the following duality result with those in [19, Theorem 9.6] and [2, Theorem 4.2] in Section 3, where the proof is provided.

Theorem 1.3. *Let $C\colon X\times\mathcal{P}^t_{d_Y}(Y)\to\mathbb{R}\cup\{\infty\}$ be jointly lower semicontinuous with respect to the product topology on $X\times\mathcal{P}^t_{d_Y}(Y)$, bounded from below, and convex in the second argument. Then we have $\mu\in\mathcal{P}(X)$ and $\nu\in\mathcal{P}^t_{d_Y}(Y)$*

$$V(\mu,\nu)=\sup_{\psi\in\Phi_{b,t}}\mu(R_C\psi)-\nu(\psi). \tag{1.4}$$

1.3.3. *C-monotonicity.* Besides primal existence and duality, another fundamental result in classical optimal transport is the characterization of optimality through the notion of *cyclical montonicity*; see [27, 16] as well as the monographs [28, 35, 36]. More recently, variants of this 'monotonicity priniciple' have been applied in transport problems for finitely or infinitely many marginals [25, 13, 20, 7, 37], the martingale version of the optimal transport problem [8, 24, 9], the Skorokhod embedding problem [6] and the distribution constrained optimal stopping problem [10].

We provide in Definition 5.1 below, a concept analogous to cyclical monotonicity (which we call C-monotonicity) for weak transport costs C . We show that every optimal transport plan is C-monotone in a very general setup. Conversely, we show that every C-monotone transport plan is optimal under certain regularity assumptions. See Theorems 5.2 and 5.5 respectively.

We note that related concepts already appeared in [5, Proposition 4.1] (where necessity of a 2-step optimality condition is established) and in [17] (necessity in the case of compactly supported measures and a quadratic cost criterion). To the best of our knowledge, our sufficient criterion is the first of its kind for weak transport costs.

1.3.4. *A general Brenier-Strassen theorem.* As an application of our abstract results we extend the Brenier-Strassen theorem [17, Theorem 1.2] of Gozlan and Juillet to the case of general probabilities on $X=Y=\mathbb{R}^d$ under the assumption that μ has finite second moment and ν has finite first moment. We thus drop the condition in [17] that the marginals have compact support. For this part we set

$$C(x,\rho):=\left|x-\int y\rho(dy)\right|^2, \tag{1.5}$$

and write $\le_c$ for the convex order of probability measures.

Theorem 1.4. *Let $\mu\in\mathcal{P}^2(\mathbb{R}^d)$ and $\nu\in\mathcal{P}^1(\mathbb{R}^d)$. There exists a unique $\mu^*\le_c\nu$ such that*

$$\mathcal{W}_2(\mu^*,\mu)^2=\inf_{\eta\le_c\nu}\mathcal{W}_2(\eta,\mu)^2=V(\mu,\nu). \tag{1.6}$$

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

Existence of μ^* and the expression (1.6) were first proved by Gozlan et al [18] for $d=1$ and by Alfonsi, Corbetta, Jourdain [1] for arbitrary $d\in\mathbb{N}$. Indeed a general version of (1.6), appealing to $\mathcal{W}_p$ and probabilities $\mu,\nu\in\mathcal{P}^p(\mathbb{R}^d)$ is provided in [1]. All other statements in the above theorem were originally established by Gozlan and Juillet [17] under the assumption of compactly supported measures μ,ν. The proof of Theorem 1.4 is given in Section 6.

2. Existence of minimizers

The principal idea in this section is to make use of the natural embedding of $\mathcal{P}(X\times Y)$ into $\mathcal{P}(X\times\mathcal{P}(Y))$, which we explain in (2.1) below. It turns out that on this 'extended' space the minimization problems Theorem 1.2 and Theorem 1.1 can be handled more efficienlty.

We need to introduce additional notation: for a probability measure $\pi\in\mathcal{P}(X\times Y)$ with not further specified marginals, we write $\pi(dx\times Y)$ and $\pi(X\times dy)$ for its X-marginal and Y-marginal respectively. At several instances we use the projection from a product space onto one of its components. This map is usually denoted by $\mathrm{proj}_\bullet$ where the subscript describes the component, e.g. $\mathrm{proj}_X\colon X\times Y\to X$ stands for the projection onto the X-component. Denoting by $(\pi_x)_{x\in X}$ a regular disintegration of π with respect to $\pi(dx\times Y)$, then we can consider the measurable map

$$\begin{aligned}\kappa_\pi\colon X&\to X\times\mathcal{P}(Y)\\ x&\mapsto(x,\pi_x).\end{aligned}$$

We define the embedding

$$J\colon\mathcal{P}(X{\times}Y)\to\mathcal{P}(X\times\mathcal{P}(Y)), \tag{2.1}$$

$$\pi\mapsto\pi(dx\times Y)\circ\kappa_\pi^{-1}. \tag{2.2}$$

The map J is well-defined since κ_π is $\pi(dx\times Y)$-almost surely unique. Note that elements in $\mathcal{P}(X\times Y)$ precisely correspond to those elements of $\mathcal{P}(X\times\mathcal{P}(Y))$ which are concentrated on a graph of a measurable function from X to $\mathcal{P}(Y)$.

We introduce the intensity $I(P)\in\mathcal{P}(Y)$ of $P\in\mathcal{P}(\mathcal{P}(Y))$, uniquely determined by

$$I(P)(f)=\int_{\mathcal{P}(Y)}p(f)P(dp)\quad\forall f\in C_b(Y). \tag{2.3}$$

Observe that I is continuous. The set of all probability measures $P\in\mathcal{P}(X\times\mathcal{P}(Y))$ with X-marginal μ and 'Y-marginal intensity' ν is denoted by

$$\Lambda(\mu,\nu):=\left\{P\in\mathcal{P}(X\times\mathcal{P}(Y))\mid\mathrm{proj}_X P=\mu,\ I(\mathrm{proj}_{\mathcal{P}(Y)}(P))=\nu\right\}. \tag{2.4}$$

Similar to (2.3), we define the intensity of $P\in\mathcal{P}(X\times\mathcal{P}(Y))$ as the unique measure $\hat I(P)\in\mathcal{P}(X\times Y)$ such that

$$\int_{X\times Y}f(x,y)\hat I(P)(dx,dy)=\int_{X\times\mathcal{P}(Y)}\int_Y f(x,y)p(dy)P(dx,dp)\quad\forall f\in C_b(X\times Y). \tag{2.5}$$

Using (2.1) and (2.5) we find that

$$\Lambda(\mu,\nu)=\hat I^{-1}(\Pi(\mu,\nu))\text{ and }J(\Pi(\mu,\nu))\subseteq\Lambda(\mu,\nu).$$

We now describe the relation between minimization problems on $\Pi(\mu,\nu)$ and $\Lambda(\mu,\nu)$:

Lemma 2.1. *Let $C\colon X\times\mathcal{P}(Y)\to\mathbb{R}\cup\{-\infty,+\infty\}$ be measurable, lower-bouunded, and convex in the second argument. Then*

$$V(\mu,\nu)=\hat V(\mu,\nu), \tag{2.6}$$

where V was defined in (1.1) *and*

$$\hat V(\mu,\nu):=\inf_{P\in\Lambda(\mu,\nu)}\int_{X\times\mathcal{P}(Y)}C(x,p)P(dx,dp). \tag{2.7}$$

Proof. For any $\pi\in\Pi(\mu,\nu)$ we have $J(\pi)\in\Lambda(\mu,\nu)$ and

$$\int_X C(x,\pi_x)\mu(dx)=\int_{X\times\mathcal{P}(Y)}C(x,p)J(\pi)(dx,dp).$$

Thus,

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_X C(x,\pi_x)\mu(dx)\geq\inf_{P\in\Lambda(\mu,\nu)}\int_{X\times\mathcal{P}(Y)}C(x,p)P(dx,dp).$$

Now, letting $P\in\Lambda(\mu,\nu)$, we easily derive from (2.5) that $\hat I(P)\in\Pi(\mu,\nu)$ and $\hat I(P)_x=\int_{\mathcal{P}(Y)}p\,P_x(dp)$ for μ-a.e x. Using convexity we conclude

$$\begin{aligned}\int_{X\times\mathcal{P}(Y)}C(x,p)P(dx,dp)&=\int_X\int_{\mathcal{P}(Y)}C(x,p)P_x(dp)\mu(dx)\\ &\geq\int_X C(x,\hat I(P)_x)\mu(dx)\\ &\geq\inf_{\pi\in\Pi(\mu,\nu)}\int_X C(x,\pi_x)\mu(dx).\end{aligned}$$

□

2.1. **Existence of minimizers.** The purpose of this subsection is to establish Theorem 1.2, or more precisely, a strengthened version of it; see Theorem 2.9 below. To this end we need a number of auxiliary results.

We start by stressing that, in general, the embedding J is not continuous. In fact:

Example 2.2. The map J is continuous if and only if X is discrete or $|Y| = 1$. Indeed, given X discrete and a sequence $(\pi^k)_{k\in\mathbb{N}} \in \mathcal{P}(X\times Y)^{\mathbb{N}}$ which weakly converges to π, we have that $\pi^k(x\times Y) \to \pi(x\times Y)$ from which $\pi^k_x(dy) = \frac{\pi^k(x,dy)}{\pi^k(x\times Y)}$ converges weakly to $\pi_x(dy) = \frac{\pi(x,dy)}{\pi(x\times Y)}$ if $\pi(x\times Y) > 0$. Consequently if $f \in C_b(X\times\mathcal{P}(Y))$, then

$$\begin{aligned}
&\lim_k |J(\pi^k)(f) - J(\pi)(f)| \\
&\quad\le \limsup_k \sum_x |f(x,\pi^k_x)(\pi^k(x\times Y) - \pi(x\times Y))| + \sum_x |f(x,\pi^k_x) - f(x,\pi_x)|\pi(x\times Y) \\
&\quad= 0.
\end{aligned}$$

Therefore $(J(\pi^k))_{k\in\mathbb{N}}$ converges weakly to $J(\pi)$. On the other hand, suppose there is a sequence $(x_k)_{k\in\mathbb{N}} \in X^{\mathbb{N}}$ of distinct points converging to some $x \in X$, as well as $p, q \in \mathcal{P}(Y)$ with $p \neq q$. For $k \in \mathbb{N}$ define a probability measure on $\mathcal{P}(X\times Y)$ by

$$\pi^k(dx,dy) := \frac{1}{2}(\delta_{x_{k+1}}(dx)p(dy) + \delta_{x_k}(dx)q(dy)).$$

A short computation yields

$$\begin{aligned}
\lim_k J(\pi^k) &= \lim_k \frac{1}{2}\left(\delta_{(x_{k+1},p)} + \delta_{(x_k,q)}\right) = \frac{1}{2}\left(\delta_{(x,p)} + \delta_{(x,q)}\right), \\
J(\lim_k \pi^k) &= J\Big(\frac{1}{2}\delta_x(p+q)\Big) = \delta_{(x,\frac{1}{2}(p+q))},
\end{aligned}$$

which shows that J is discontinuous.

On the bright side, J possesses a crucial feature: it maps relatively compact sets to relatively compact sets. We prove this in Lemma 2.6 below. But first we need to digress into the characterization of tightness on $\mathcal{P}(\mathcal{P}(Y))$ and subspaces thereof. The following can be found in [32, p. 178, Ch. II].

Lemma 2.3. *A set $\mathcal{A} \subseteq \mathcal{P}(\mathcal{P}(Y))$ is tight if and only if the set of its intensities $I(\mathcal{A})$ is tight in $\mathcal{P}(Y)$.*

We need to refine Lemma 2.3 for our purposes, since we equip $\mathcal{P}^t_{d_Y}(Y)$ with the $\mathcal{W}_t$-topology instead of the weak topology.

Lemma 2.4. *A set $\mathcal{A} \subseteq \mathcal{P}^t_{\mathcal{W}_t}(\mathcal{P}^t_{d_Y}(Y))$ is relatively compact if and only if the set of its intensities $I(\mathcal{A})$ is relatively compact in $\mathcal{P}^t_{d_Y}(Y)$.*

The proof of Lemma 2.4 heavily relies on the following lemma, for which we include a proof for sake of completeness.

Lemma 2.5. *A set $\mathcal{A} \subseteq \mathcal{P}^t_{d_Y}(Y)$ is relatively compact if and only if it is tight and*

$$\exists y' \in Y\ \forall\varepsilon > 0\ \exists R > 0\colon \sup_{\mu\in\mathcal{A}} \int_{B_R(y')^c} d_Y(y,y')^t \mu(dy) < \varepsilon \tag{2.8}$$

Proof of Lemma 2.4. The first implication follows by continuity of I and Lemma 2.3 provides tightness. Given $I(\mathcal{A})$ is relatively compact in $\mathcal{P}^t_{d_Y}(Y)$, it remains to show for fixed $y' \in Y$ that

$$\forall\varepsilon > 0\ \exists R > 0\colon \sup_{P\in\mathcal{A}} \int_{\{p\colon \mathcal{W}_t(p,\delta_{y'})^t \ge R\}} \mathcal{W}_t(p,\delta_{y'})^t P(dp) < \varepsilon.$$

Fix $\varepsilon > 0$. There exist $K > 0$ and $r > 0$ such that for all $P \in \mathcal{A}$

$$\int_{\mathcal{P}^t_{d_Y}(Y)} \mathcal{W}_t(p, \delta_{y'})^t P(dp) = \int_Y d_Y(y, y')^t I(P)(dy) \le K$$

$$\int_{\mathcal{P}^t_{d_Y}(Y)} \int_{B^c_r(y')} d_Y(y, y')^t p(dy) P(dp) = \int_{B^c_r(y')} d_Y(y, y')^t I(P)(dy) < \frac{\varepsilon}{2}. \tag{2.9}$$

Set $R = \frac{2r^t K}{\varepsilon}$ and $A_R := \left\{p \in \mathcal{P}^t_{d_Y}(Y) \colon \mathcal{W}_t(p, \delta_{y'})^t \ge R\right\}$, then

$$\sup_{P \in \mathcal{A}} P(A_R) \le \sup_{P \in \mathcal{A}} \frac{1}{R} \int_{A_R} \mathcal{W}_t(p, \delta_{y'})^t P(dp) \le \frac{K}{R}$$

and

$$\sup_{P \in \mathcal{A}} \int_{A_R} \int_{B_r(y')} d_Y(y, y')^t p(dy) P(dp) \le \sup_{P \in \mathcal{A}} P(A_R) r^t \le \frac{\varepsilon}{2} \tag{2.10}$$

Putting (2.9) and (2.10) together completes the proof. □

Proof of Lemma 2.5.

'⇒': Since the topology induced by $\mathcal{W}_t$ on $\mathcal{P}^t_{d_Y}(Y)$ is finer than the weak topology on $\mathcal{P}^t_{d_Y}(Y)$, relative compactness in $\mathcal{W}_t$ implies relative compactness with respect to the weak topology. Therefore, Prokhorov's theorem yields tightness. Note that (2.8) follows immediately from the definition of convergence in $\mathcal{W}_t$.

'⇐': Let $\mathcal{A}$ be tight such that (2.8) holds. Then, any sequence $(\mu_k)_{k \in \mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$ has an accumulation point $\mu \in \mathcal{P}(Y)$ with respect to the weak topology. Without loss of generality assume that $\mu_k \to \mu$ for $k \to \infty$. By monotone convergence and the Portmanteau theorem, we have for $O \subseteq Y$ open

$$\begin{aligned}
\int_O d_Y(y, y')^t \mu(dy) &= \lim_{R \to \infty} \int_O R \wedge d_Y(y, y')^t \mu(dy) \\
&\le \lim_{R \to \infty} \liminf_{n \to \infty} \int_O R \wedge d_Y(y, y')^t \mu_n(dy) \\
&\le \sup_n \int_O d_Y(y, y')^t \mu_n(dy).
\end{aligned}$$

Hence, by (2.8) we can choose (for $\varepsilon = 1$, say) $R > 0$ such that

$$\int_Y d_Y(y, y')^t \mu(dy) \le \sup_n \int_{B_R(y')} d_Y(y, y')^t \mu_n(dy) + \varepsilon < \infty,$$

which shows that $\mu \in \mathcal{P}^t_{d_Y}(Y)$.

Next, fix $\varepsilon > 0$. Pick $R > 0$ such that

$$\int_Y d_Y(y, y')^t - R^t \wedge d_Y(y, y')^t \mu(dy) < \varepsilon,$$

$$\sup_n \int_{B_R(y')^C} d_Y(y, y')^t \mu_n(dy) < \varepsilon.$$

By weak convergence we know that

$$\lim_k \int_Y R^t \wedge d_Y(y, y')^t \mu_k(dy) \to \int_Y R^t \wedge d_Y(y, y')^t \mu(dy).$$

Hence we may pick k_0 such that for all $k \ge k_0$

$$\left| \int_Y R^t \wedge d_Y(y, y')^t (\mu_k - \mu)(dy) \right| < \varepsilon.$$

Thus we have for $k \ge k_0$

$$\left| \int_Y d_Y(y, y')^t (\mu_k - \mu)(dy) \right| < 3\varepsilon.$$

Since ε was arbitrary, we obtain that the t-moments are converging, which implies convergence in $\mathcal{W}_t$. □

We recall that on Y we are usually given a compatible complete metric d_Y, whereas on X we fix a compatible bounded metric d_X. We thus endow the product spaces $X \times Y$ and $X \times \mathcal{P}^t_{d_Y}(Y)$ with natural (product) metrices d and $\hat{d}$ defined respectively by

$$d((x,y),(x_0,y_0)) = d_X(x,x_0) + d_Y(y,y_0), \tag{2.11}$$

$$\hat{d}((x,p),(x_0,p_0)) = d_X(x,x_0) + \mathcal{W}_t(p,p_0). \tag{2.12}$$

We can now state and prove the crucial property of J:

Lemma 2.6. *If $\Pi \subseteq \mathcal{P}^t_d(X \times Y)$ is relatively compact then $J(\Pi) \subseteq \mathcal{P}^t_{\hat{d}}(X \times \mathcal{P}^t_{d_Y}(Y))$ is relatively compact.*

Proof. By continuous mapping (see [14, Theorem A.3.10]) the sets $\Pi^X \subseteq \mathcal{P}(X)$ and $\Pi^Y \subseteq \mathcal{P}^t_{d_Y}(Y)$, consisting respectively of the X- and Y-marginals of the elements in Π, are tight. Then, relative compactness of Π^X and Π^Y can be readily derived by courtesy of Lemma 2.5 and the structure of the product metric d, cf. (2.11).

Denote now respectively by $\Pi^X_J \subseteq \mathcal{P}(X)$ and $\Pi^Y_J \subseteq \mathcal{P}^t_{\mathcal{W}_t}(\mathcal{P}^t_{d_Y}(Y))$ the set of X- and $\mathcal{P}(Y)$-marginals of the elements in $J(\Pi)$. Clearly $\Pi^X_J = \Pi^X$. By Lemma 2.4, the set Π^Y_J is relatively compact in $\mathcal{P}^t_{\mathcal{W}_t}(\mathcal{P}^t_{d_Y}(Y))$ if and only if the set $J(\Pi^Y_J)$ is relatively compact in $\mathcal{P}^t_{d_Y}(Y)$. 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 $I(\Pi^Y_J) \subseteq \Pi^Y$ is relatively compact and so is Π^Y_J. Since the marginals of $J(\Pi)$ are relatively compact, we conclude that $J(\Pi)$ itself is relatively compact. □

It is convenient to introduce the following assumptions, which we will often require:

Definition 2.7 (A). *Given Polish spaces X, Y, we say that a function*

$$C \colon X \times \mathcal{P}^t_{d_Y}(Y) \to \mathbb{R} \cup \{+\infty\}$$

satisfies Condition (A) *if the following hold:*

- *C is lower semicontinuous with respect to the product topology of*

$$(X, d_X) \times (\mathcal{P}^t_{d_Y}(Y), \mathcal{W}_t),$$

- *C is bounded from below.*

If in addition for all $x \in X$ the map $p \mapsto C(x,p)$ is convex, i.e.

$$p, q \in \mathcal{P}^t_{d_Y}(Y), \alpha \in [0,1] \Rightarrow \ C(x, \alpha p + (1-\alpha)q) \le \alpha C(x,p) + (1-\alpha)C(x,q), \tag{2.13}$$

then we say that C satisfies Condition (A+).

We now show that under Condition (A+) the cost functional defining the weak transport problem is lower semicontinuous:

Proposition 2.8. *Let $C \colon X \times \mathcal{P}^t_{d_Y}(Y) \to \mathbb{R} \cup \{+\infty\}$ satisfy condition* (A). *Then the map*

$$\mathcal{P}^t_{\hat{d}}(X \times \mathcal{P}^t_{d_Y}(Y)) \ni P \mapsto \int_{X \times \mathcal{P}^t_{d_Y}(Y)} C(x,p) P(dx, dp) \tag{2.14}$$

is lower semicontinuous. If C satisfies condition (A+) *then the map*

$$\mathcal{P}^t_d(X \times Y) \ni \pi \mapsto \int_X C(x, \pi_x) \pi(dx \times Y) \tag{2.15}$$

is lower semicontinuous.

Proof. Let $P^k \to P$ in $\mathcal{P}^t_{\hat{d}}(X \times \mathcal{P}^t_{d_Y}(Y))$. Similar to [14, Theorem A.3.12], we can approximate C from below by $\hat{d}$-Lipschitz functions and obtain lower semicontinuity of (2.14), i.e.,

$$\liminf_k \int_{X \times \mathcal{P}(Y)} C(x,p)\, P^k(dx, dp) \ge \int_{X \times \mathcal{P}(Y)} C(x,p)\, P(dx, dp).$$

To show lower semicontinuity of (2.15), let $\pi^k \to \pi$ in $\mathcal{P}^t_d(X\times Y)$ and denote $P^k = J(\pi^k)$. We may assume that $\liminf_k \int_X C(x,\pi^k_x)\pi^k(dx\times Y) = \lim_k \int_X C(x,\pi^k_x)\pi^k(dx\times Y)$ by selecting a subsequence. By Lemma 2.6 we know that $\{P^k\}_k$ is relatively compact in $\mathcal{P}^t_{\hat d}(X\times\mathcal{P}^t_{d_Y}(Y))$. Denote by P an accumulation point of $\{P^k\}_k$. From now on we work along a subsequence converging to P. Observe that

$$\int_X C(x,\pi^k_x)\pi^k(dx\times Y) = \int_{X\times\mathcal{P}(Y)} C(x,p)\,P^k(dx,dp).$$

Hence, we find by the first part that

$$\liminf_k \int_{X\times\mathcal{P}(Y)} C(x,p)\,P^k(dx,dp) \ge \int_{X\times\mathcal{P}(Y)} C(x,p)\,P(dx,dp).$$

Observe that the X-marginal of P equals the X-marginal of π, so by convexity of $C(x,\cdot)$ we then have

$$\begin{aligned}\liminf_k \int_X C(x,\pi^k_x)\pi^k(dx\times Y) &\ge \int_{X\times\mathcal{P}(Y)} C(x,p)\,P_x(dp)\pi(dx\times Y)\\ &\ge \int_X C\Big(x,\int_{\mathcal{P}(Y)} p(dy)P_x(dp)\Big)\pi(dx\times Y).\end{aligned}$$

Now, if f is continuous bounded on $X\times Y$, we have

$$\int_{X\times Y} f(x,y)\pi^k(dx,dy) \to \int_{X\times Y} f(x,y)\pi^k(dx,dy).$$

But the function $F(x,p) := \int_Y f(x,y)p(dy)$ is easily seen to be continuous and bounded in $X\times\mathcal{P}(Y)$. Hence $\int F dP^k \to \int F dP$ and by the structure of F we deduce

$$\int_{X\times Y} f(x,y)\pi(dx,dy) = \int F dP = \int_{X\times\mathcal{P}(Y)}\int_Y f(x,y)p(dy)P(dx,dp).$$

This shows for the disintegration $(\pi_x)_{x\in X}$ of π that $\pi_x(dy) = \int_{\mathcal{P}(Y)} p(dy)\,P_x(dp)$ for $\pi(dx\times Y)$-almost every x. So we conclude

$$\liminf_k \int_X C(x,\pi^k_x)\pi^k(dx\times Y) \ge \int_X C(x,\pi_x)\pi(dx\times Y).$$

□

We are finally ready to provide our main existence result:

Theorem 2.9. *Let $C\colon X\times\mathcal{P}^t_{d_Y}(Y)\to\mathbb{R}\cup\{+\infty\}$ satisfy Condition* (A)*. If $\Lambda\subseteq\mathcal{P}^t_{\hat d}(X\times\mathcal{P}^t_{d_Y}(Y))$ is compact, then there exists a minimizer $P^*\in\Lambda$ of*

$$\inf_{P\in\Lambda}\int_{X\times\mathcal{P}(Y)} C(x,p)P(dx,dp).$$

In particular $\mathcal{P}(X)\times\mathcal{P}^t_{d_Y}(Y)\ni(\mu,\nu)\mapsto\hat V(\mu,\nu)$ is lower semicontinuous and $\hat V(\mu,\nu)$ is attained (recall (2.7)*). Assume now that C fulfils Condition* (A+) *and $\Pi\subseteq\mathcal{P}^t_d(X\times Y)$ is compact. Then there exists a minimizer $\pi^*\in\Pi$ of*

$$\inf_{\pi\in\Pi}\int_X C(x,\pi_x)\pi(dx\times Y).$$

In particular $\mathcal{P}(X)\times\mathcal{P}^t_{d_Y}(Y)\ni(\mu,\nu)\mapsto V(\mu,\nu)$ is lower semicontinuous and $V(\mu,\nu)$ is attained (recall (1.1)*).*

Proof. The existence of minimizers in Λ and Π are direct consequences of their compactness and the lower semicontinuity of the objective functionals (Proposition 2.8).

We move to the study of $\hat V$. Let $(\mu_k,\nu_k)\to(\mu,\nu)$ in $\mathcal{P}(X)\times(\mathcal{P}^t_{d_Y},\mathcal{W}_t)$. For any $k\in\mathbb{N}$ we find an optimizer P^*_k of $\hat V(\mu_k,\nu_k)$. Note that the set $\{P^*_k\colon k\in\mathbb{N}\}$ is relatively compact in

$\mathcal{P}^t_{\hat{d}}(X \times \mathcal{P}^t_{d_Y}(Y))$. Therefore, we can find again a converging subsequence with limit point in $\Pi(\mu,\nu)$. Without loss of generality we assume

$$\liminf_k \hat{V}(\mu_k,\nu_k) = \lim_k \hat{V}(\mu_k,\nu_k).$$

Using lower semicontinuity of the objective functional shows the assertion for $\hat{V}$. By Lemma 2.1 the lower semicontinuity of V is immediate. □

Of course Theorems 1.1 and 1.2 are particular cases of the second half of Theorem 2.9. More generally: if A is compact in $\mathcal{P}(X)$ and B is compact in $(\mathcal{P}^t_{d_Y}(Y), \mathcal{W}_t)$, then $\Pi := \bigcup_{\mu\in A,\nu\in B} \Pi(\mu,\nu)$ is compact in $\mathcal{P}^t_d(X\times Y)$ and Theorem 2.9 applies.

3. Duality

We denote by Φ_t the set of continuous functions on Y which satisfy the growth constraint

$$\exists y_0 \in Y,\ \exists a,b \in \mathbb{R}_+,\ \forall y \in Y:\ |\psi(y)| \le a + b d_Y(y,y_0)^t,$$

and by $\Phi_{b,t}$ the subset of functions in Φ_t which are bounded from below. Further, we recall the notion of C-conjugate : The C-conjugate of a measurable function $\psi\colon Y \to \mathbb{R}$, denoted $R_C\psi$, is given by

$$R_C\psi(x) := \inf_{p\in\mathcal{P}^t_{d_Y}(Y)} p(\psi) + C(x,p). \tag{3.1}$$

We obtain Theorem 1.3 as a particular case of the following:

Theorem 3.1. *Let $C\colon X\times\mathcal{P}^t_{d_Y}(Y) \to \mathbb{R}\cup\{+\infty\}$ satisfy Condition* (A)*, then*

$$\inf_{P\in\Lambda(\mu,\nu)} \int_{X\times\mathcal{P}(Y)} C(x,p)P(dx,dp) = \sup_{\psi\in\Phi_{b,t}} -\nu(\psi) + \int_X R_C\psi(x)\mu(dx). \tag{3.2}$$

If moreover C satisfies Condition (A+)*, then*

$$V(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_X C(x,\pi_x)\mu(dx) = \sup_{\psi\in\Phi_{b,t}} -\nu(\psi) + \int_X R_C\psi(x)\mu(dx). \tag{3.3}$$

Remark 3.2. A proof of Theorem 1.3 can be obtained by means of [19, Theorem 9.6], since we may verify the hypotheses therein thanks to our Proposition 2.8. We prefer to obtain the slightly stronger Theorem 3.1 via self-contained arguments. The primal-dual equality (3.3) was obtained in [2, Theorem 4.2] in the case when X,Y are compact spaces.

Proof of Theorem 3.1. Fix $y_0 \in Y$. Define the auxiliary cost function $\widetilde{C}\colon X\times\mathcal{P}^t_{d_Y}(Y)$ by

$$\widetilde{C}(x,p) := C(x,p) + \mathcal{W}_t(p,\delta_{y_0})^t$$

and $F\colon \mathcal{P}^t_{d_Y}(Y) \to \mathbb{R}\cup\{+\infty\}$ by

$$\begin{aligned} F(m) &:= \inf_{P\in\Lambda(\mu,m)} \int_{X\times\mathcal{P}(Y)} \widetilde{C}(x,p)P(dx,dp) \\ &= \inf_{P\in\Lambda(\mu,m)} \int_{X\times\mathcal{P}(Y)} C(x,p)P(dx,dp) + \int_Y d_Y(y,y_0)^t m(dy). \end{aligned} \tag{3.4}$$

Since the integrand $\widetilde{C}$ is bounded from below and lower semicontinuous we can apply Proposition 2.8 and find that F is lower semicontinuous on $\mathcal{P}^t_{d_Y}(Y)$. Note that for any $\alpha\in[0,1]$ and $m_1,m_2\in\mathcal{P}^t_{d_Y}(Y)$ we have

$$P_i \in \Lambda(\mu,m_i),\ i=1,2 \implies \alpha P_1 + (1-\alpha)P_2 \in \Lambda(\mu, \alpha m_1 + (1-\alpha)m_2),$$

and, particularly, it follows that F is convex. We can extend F to the set $\mathcal{M}^t_{d_Y}(Y)$ of bounded signed measures with finited t-moment (i.e. $m\in\mathcal{M}^t_{d_Y}(Y)$ implies $\int_Y d_Y(y,y_0)^t|m|(dy) < \infty$ for some y_0) by setting $F(m) = +\infty$ if $m \notin \mathcal{P}^t_{d_Y}(Y)$. We equip the space $\mathcal{M}^t_{d_Y}(Y)$ with the topology induce by Φ_t. It follows that the extension of F is still convex and lower semicontinuous. Now, the spaces Φ_t and $\mathcal{M}^t_{d_Y}(Y)$ are in separating duality. Define the convex conjugate $F^*\colon \Phi_t \to \mathbb{R} \cup \{+\infty\}$ of F by

$$F^*(\psi) = \sup_{m \in \mathcal{P}^t_{d_Y}(Y)} m(\psi) - F(m). \tag{3.5}$$

Observe that $F^*(\psi) = \lim_{k\to+\infty} F^*(\psi \wedge k)$, by monotone convergence. We may apply the Fenchel duality theorem [38, Theorem 2.3.3], and then replace Φ_t by $\Phi_{b,t}$, obtaining:

$$\begin{aligned} F(m) &= \sup_{\psi\in\Phi_t} m(\psi) - F^*(\psi) \\ &= \sup_{-\psi\in\Phi_{b,t}} m(\psi) - F^*(\psi) \\ &= \sup_{\psi\in\Phi_{b,t}} m(-\psi) - F^*(-\psi). \end{aligned}$$

Now we show that

$$F^*(-\psi) = -\int_X R_{\widetilde{C}}\psi(x)\mu(dx). \tag{3.6}$$

Rewriting (3.5) yields

$$\begin{aligned} F^*(-\psi) &= \sup_{m\in\mathcal{P}^t_{d_Y}(Y)} m(-\psi) - \inf_{P\in\Lambda(\mu,m)} \int_{X\times\mathcal{P}(Y)} \widetilde{C}(x,p)P(dx,dp) \\ &= \sup_{\substack{m\in\mathcal{P}^t_{d_Y}(Y)\\ P\in\Lambda(\mu,m)}} -\int_X \left(\int_{\mathcal{P}(Y)} p(\psi) + \widetilde{C}(x,p)P_x(dp) \right) \mu(dx) \\ &= -\inf_{\substack{m\in\mathcal{P}^t_{d_Y}(Y)\\ P\in\Lambda(\mu,m)}} \int_X \left(\int_{\mathcal{P}(Y)} p(\psi) + \widetilde{C}(x,p)P_x(dp) \right) \mu(dx) \\ &\le -\int_X R_{\widetilde{C}}\psi(x)\mu(dx). \end{aligned}$$

To show the converse inequality, we assume without loss of generality that $\int_X \widetilde{R_C}\psi(x)\mu(dx) < +\infty$. For all $x \in X$ the value of $R_{\widetilde{C}}\psi(x)$ is finite , because ψ is bounded from below. Fix $\varepsilon > 0$. The map $R_{\widetilde{C}}\psi(\cdot)$ is lower semianalytic by [12, Proposition 7.47] and by [12, Proposition 7.50] there exists an analytically measurable probability kernel $(\tilde{p}_x)_{x\in X} \in (\mathcal{P}^t_{d_Y}(Y))^X$ such that for all $x \in X$

$$p_x(\psi) + \widetilde{C}(x, p_x) \le R_{\widetilde{C}}\psi(x) + \varepsilon.$$

Then, we immediately obtain

$$\int_X p_x(\psi) + \widetilde{C}(x,p_x)\mu(dx) \le \int_X R_{\widetilde{C}}\psi(x)\mu(dx) + \varepsilon.$$

The term $\delta_{p_x}(dp)\mu(dx)$ uniquely defines a probability measures $\tilde{P}$ on $X \times \mathcal{P}(Y)$. Since $\widetilde{C}$ and ψ are bounded from below, we infer that

$$\mathcal{W}_t(\mathrm{proj}_Y \hat{I}(\tilde{P}), \delta_{y_0})^t = \int_{X\times\mathcal{P}(Y)} \mathcal{W}_t(p, \delta_{y_0})^t \tilde{P}(dx,dp) < +\infty,$$

and in particular $\mathrm{proj}_Y \hat{I}(\tilde{P}) \in \mathcal{P}^t_{d_Y}(Y)$. Clearly $\tilde{P} \in \Lambda(\mu, \mathrm{proj}_Y \hat{I}(\tilde{P}))$, so

$$\begin{aligned} -\int_X \widetilde{R_C}\psi(x)\mu(dx) &\le \mathrm{proj}_Y(\hat{I}(\tilde{P}))(-\psi) - \int_{X\times\mathcal{P}(Y)} C(x,p) + \mathcal{W}_t(p,\delta_{y_0})^t \tilde{P}(dx,dp) + \varepsilon \\ &\le \mathrm{proj}_Y(\hat{I}(\tilde{P}))(-\psi) - F\left(\mathrm{proj}_Y \hat{I}(\tilde{P})\right) + \varepsilon \\ &\le F^*(-\psi) + \varepsilon, \end{aligned}$$

and since ε was arbitrary, we have shown (3.6).

So far, we know that

$$F(m) = \sup_{\psi\in\Phi_{b,t}} -m(\psi) + \int_X R_{\widetilde{C}}\psi(x)\mu(dx).$$

Define $f(y) := d_Y(y, y_0)^t$ and note that $R_C(\psi + f)(x) = R_{\widetilde{C}}\psi(x)$ for all $x \in X$, as well as $\psi + f \in \Phi_{b,t}$ for $\psi \in \Phi_{b,t}$. From (3.4) we get

$$\begin{aligned}\inf_{P\in\Lambda(\mu,m)} P(C) &= F(m) - \mathcal{W}_t(m, \delta_{y_0})^t \\ &= \sup_{\psi\in\Phi_{b,t}} -m(\psi + f) + \int_X R_{\widetilde{C}}\psi(x)\mu(dx) \\ &= \sup_{\psi\in\Phi_{b,t}} -m(\psi) + \int_X R_C\psi(x)\mu(dx),\end{aligned}$$

which shows (3.2).

If for all $x \in X$ the map $C(x, \cdot)$ is convex, then (3.3) follows by Lemma 2.1 and (3.2). □

4. On the restriction property

The restriction property of optimal transport roughly states that if a coupling is optimal, then the conditioning of the coupling to a subset is also optimal given its marginals. This property fails for weak optimal transport, as we illustrate with an example:

Example 4.1. Let $X = Y = \mathbb{R}$, $\mu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, $\nu = \frac{1}{4}\delta_{-2} + \frac{1}{2}\delta_0 + \frac{1}{4}\delta_2$ and $C(x,\rho) = \left(x - \int y\rho(dy)\right)^2$. We consider the weak transport problem with these ingredients, and observe that an optimal coupling is given by

$$\pi = \frac{1}{4}[\delta_{(1,2)} + \delta_{(1,0)} + \delta_{(-1,0)} + \delta_{(-1,-2)}],$$

since it produces a cost equal to zero. Consider the set $K = \{(x, y) : y \neq 0\}$ and $\tilde{\pi}(dx, dy) = \pi(dx, dy|K)$ the conditioning of π to the set K, i.e. $\tilde{\pi}(S) := \frac{\pi(S\cap K)}{\pi(K)}$. It follows that

$$\tilde{\pi} = \frac{1}{2}[\delta_{(1,2)} + \delta_{(1,-2)}],$$

and denoting by $\tilde{\mu}$ and $\tilde{\nu}$ the first and second marginals of $\tilde{\pi}$, we have $\tilde{\mu} = \mu$ and $\tilde{\nu} = \frac{1}{2}\delta_2 + \frac{1}{2}\delta_{-2}$. With $\tilde{\mu}$ and $\tilde{\nu}$ and again the cost C as ingredients, an optimizer for the weak transport problem is given by

$$\hat{\pi} = \frac{3}{8}\delta_{(1,2)} + \frac{1}{8}\delta_{(1,-2)} + \frac{1}{8}\delta_{(-1,2)} + \frac{3}{8}\delta_{(-1,-2)},$$

since this time this coupling produces a cost equal to zero. On the other hand the cost of $\tilde{\pi}$ is equal to 1, and so $\tilde{\pi}$ is not optimal between is marginals.

However, we can state a positive result, used in Section 6 for the proof of Theorem 1.4:

Proposition 4.2. *Suppose that π is optimal between the marginals μ and ν, $V(\mu, \nu) < \infty$, and that $C(x, \cdot)$ is convex. Let $0 \leq \tilde{\mu} \leq \mu$ be a non-negative measure such that $0 \not\equiv \tilde{\mu}$ and define $\hat{\mu} = \tilde{\mu}/\tilde{\mu}(X)$. Then $\hat{\pi}(dx, dy) := \hat{\mu}(dx)\pi_x(dy)$ is optimal between its marginals.*

Proof. By contradiction, suppose there exists a coupling χ with the same marginals as $\hat{\pi}$ such that

$$\int C(x, \chi_x)\hat{\mu}(dx) < \int C(x, \hat{\pi}_x)\hat{\mu}(dx).$$

Now define $\pi^* := \pi + \tilde{\mu}(X)[\chi - \hat{\pi}] = \pi - \tilde{\mu}.\pi_x + \tilde{\mu}(X)\chi$. Observe that π^* has marginals μ, ν, and $\pi^*(X \times Y) = 1$. We also have $\pi^* \geq 0$ since $\tilde{\mu} \leq \mu$, so π^* is a probability measure. Of course $0 \leq \frac{d\tilde{\mu}}{d\mu} \leq 1$ and clearly $\pi^*_x = \left(1 - \frac{d\tilde{\mu}}{d\mu}(x)\right)\pi_x + \frac{d\tilde{\mu}}{d\mu}(x)\chi_x$. Therefore

$$\begin{aligned}\int C(x, \pi^*_x)\mu(dx) &= \int C\left(x, \left(1 - \frac{d\tilde{\mu}}{d\mu}(x)\right)\pi_x + \frac{d\tilde{\mu}}{d\mu}(x)\chi_x\right)\mu(dx) \\ &\leq \int C(x, \pi_x)\mu(dx) + \int [C(x, \chi_x) - C(x, \pi_x)]\tilde{\mu}(dx) \\ &< \int C(x, \pi_x)\mu(dx),\end{aligned}$$

where we used convexity in the first inequality and that $V(\mu,\nu) < \infty$ in the second one. □

5. C-Monotonicity for weak transport costs

Cyclical monotonicity plays a crucial role in classical optimal transport [27, 16]. This has inspired similar development for weak transport costs in [5, 17]:

Definition 5.1 (C-monotonicity)**.** *We say that a coupling $\pi \in \Pi(\mu,\nu)$ is C-monotone if there exists a measurable set $\Gamma \subseteq X$ with $\mu(\Gamma) = 1$, such that for any finite number of points $x_1,\dots,x_N$ in Γ and measures $m_1,\dots,m_N$ in $\mathcal{P}(Y)$ with $\sum_{i=1}^N m_i = \sum_{i=1}^N \pi_{x_i}$, the following inequality holds:*

$$\sum_{i=1}^N C(x_i,\pi_{x_i}) \le \sum_{i=1}^N C(x_i,m_i).$$

We first show that C-monotonicity is necessary for optimality under minimal assumptions. We then provide strengthened assumptions under which C-monotonicity is sufficient.

5.1. **C-monotonicity: necessity.** We denote by S_N the set of permutations of the set $\{1,\dots,N\}$. If $\vec{z} := (z_1\dots,z_n)$ is any N-vector, and $\sigma \in S_N$, we naturally overload the notation by defining

$$\sigma(\vec{z}) := (z_{\sigma(1)},\dots,z_{\sigma(N)}).$$

Recall the notation (1.1) for the weak transport problem. Our main result, concerning the necessity of C-monotonicity is the following:

Theorem 5.2. *Let C be jointly measurable and $C(x,\cdot)$ be convex and lower semicontinuous for all x. Assume that π^* is optimal for $V(\mu,\nu)$ and $|V(\mu,\nu)| < \infty$. Then π^* is C-monotone.*

Proof. Let $N \in \mathbb{N}$. Then

$$\mathcal{D}_N := \Big\{((x_1,\dots,x_N),(m_1,\dots,m_N)) \in X^N \times \mathcal{P}(Y)^N :$$
$$\sum_{i=1}^N \pi^*_{x_i} = \sum_{i=1}^N m_i \text{ and } \sum_{i=1}^N C(x_i,\pi^*_{x_i}) > \sum_{i=1}^N C(x_i,m_i)\Big\},$$

is an analytic set. Write

$$D_N := \mathrm{proj}_{X^N}(\mathcal{D}_N).$$

By Jankov-von Neumann uniformization [21, Theorem 18.1] there is an analytically measurable function $f_N\colon D_N \to \mathcal{P}(Y)^N$ such that graph$(f_N) \subseteq \mathcal{D}_N$. We can extend f_N to X^N by defining it on $X^N \setminus D_N$ as the Borel-measurable map $\vec{x} \mapsto (\pi^*_{x_1},\dots,\pi^*_{x_N})$. Observe that for all $\sigma \in S_N$, we have $(\sigma,\sigma)(\mathcal{D}_N) = \mathcal{D}_N$. Thanks to this, and Lemma 5.3 below, we can assume without loss of generality that f_N satisfies

$$f_N \circ \sigma = \sigma \circ f_N \quad \forall \sigma \in S_N.$$

We write $f_N^i(\vec{x})$ for the i-th element of the vector $f_N(\vec{x}) \in \mathcal{P}(Y)^N$.

Assume that there exists a coupling $Q \in \Pi(\mu^N) = \Pi(\mu,\dots,\mu)$ such that $Q(D_N) > 0$. We now show that this is in conflict with optimality of π^*. We clearly may assume that Q is symmetric, i.e. such that for all $\sigma \in S_N$ we have $Q(B) = Q(\sigma(B))$ for all $B \in \mathcal{B}(X^N)$ (in other words $\sigma(Q) = Q$). We define the possible contender $\tilde{\pi}$ of π^* by

$$\tilde{\pi}(dx_1,dy) := \mu(dx_1)\int_{X^{N-1}} Q_{x_1}(dx_2,\dots,dx_n) f_N^1(x_1,\dots,x_N)(dy), \tag{5.1}$$

which is legitimate owing to all measurability precautions we have taken. We will prove

(1) $\tilde{\pi} \in \Pi(\mu,\nu)$,
(2) $\int \mu(dx) C(x,\pi^*_x) > \int \mu(dx) C(x,\tilde{\pi}_x)$.

Ad (1): Evidently the first marginal of $\tilde{\pi}$ is μ. Write $\sigma_i \in S_N$ for the permutation that merely interchanges the first and i-th component of a vector. By the symmetric properties of Q and f_N we find

$$\begin{aligned}\int_X \mu(dx_1)\tilde{\pi}_{x_1}(dy) &= \int_{X^N} Q(dx_1,\dots,dx_N) f_N^1(\vec{x})(dy)\\ &= \frac{1}{N}\sum_{i=1}^N \int_{X^N} \sigma_i(Q)(dx_1,\dots,dx_N) f_N^i(\vec{x})(dy)\\ &= \frac{1}{N}\sum_{i=1}^N \int_{X^N} Q(dx_1,\dots,dx_N)\pi^*_{x_i}(dy)\\ &= \nu(dy).\end{aligned}$$

Ad (2): On D_N holds by construction the strict inequality

$$\sum_{i=1}^N C(x_i, f_N^i(\vec{x})) < \sum_{i=1}^N C(x_i, \pi_{x_i}).$$

Using convexity of $C(x,\cdot)$ and the symmetry properties of Q and f_N, we find

$$\begin{aligned}\int_X C(x,\tilde{\pi}_x)\mu(dx) &= \int_X \mu(dx_1) C\left(x_1, \int_{X^{N-1}} Q_{x_1}(dx_2,\dots,dx_N) f_N^1(\vec{x})\right)\\ &\le \int_{X^N} Q(d\vec{x}) C(x_1, f_N^1(\vec{x}))\\ &= \frac{1}{N}\sum_{i=1}^N \int_{X^N} Q(d\vec{x}) C(x_i, f_N^i(\vec{x}))\\ &< \frac{1}{N}\sum_{i=1}^N \int_{X^N} Q(d\vec{x}) C(x_i, \pi_{x_i}) = \int_X C(x,\pi_x)\mu(dx),\end{aligned}$$

yielding a contradiction to the optimality of π^*.

We conclude that no measure Q with the stated properties exists. By "Kellerer's lemma" [8, Proposition 2.1], which is also true for analytic sets, we obtain that D_N is contained in a set of the form $\bigcup_{k=1}^N \mathrm{proj}_k^{-1}(M_N)$ where $\mu(M_N) = 0$ and proj_k denotes the projection from X^N to its k-th component. Since $N \in \mathbb{N}$ was arbitrary, we can define the set $\Gamma := (\bigcup_{N\in\mathbb{N}} M_N)^C$ with $\mu(\Gamma) = 1$, which has the desired property. □

The missing bit in the above proof is Lemma 5.3. By [21, Theorem 7.9] there exists for every Polish space X a closed subset F of the Baire space $\mathcal{N} := \mathbb{N}^{\mathbb{N}}$ and a continuous bijection $h_X : F \to X$. On the Baire space the lexicographic order naturally provides a total order. Hence, X inherits the total order of $F \subseteq \mathcal{N}$ by virtue of h_X and its Borel-measurable inverse $h_X^{-1} := g_X$, namely:

$$x, y \in X : x \le y \Leftrightarrow h_X^{-1}(x) = g_X(x) \le h_X^{-1}(y) = g_X(y).$$

Lemma 5.3. *The set*

$$A = \left\{\vec{x} \in X^N : x_1 \le x_2 \le \ldots \le x_N\right\},$$

is Borel-measurable. Given $f : A \subseteq X^N \to Y^N$ an analytically measurable function, there exists an analytically measurable extension $\hat{f} : X^N \to Y^N$ such that for any $\sigma \in S_N$

$$\hat{f} \circ \sigma = \sigma \circ \hat{f}.$$

Proof of Lemma 5.3. Let $\hat{A} = \left\{\vec{a} \in \mathcal{N}^N : a_1 \le a_2 \le \ldots \le a_N\right\}$, and define $g : \mathcal{N}^N \to S_N$ by $g(\vec{a}) = \sigma$ where $\sigma \in S_N$ satisfies

- $\sigma(\vec{a}) \in \hat{A}$
- for each i, j such that $0 \le i < j \le N$ it holds

$$a_i = a_j \implies \sigma(i) < \sigma(j).$$

With these precautions $g(\vec{a}) = \sigma$ is indeed well defined. For each $\sigma \in S_N$ we define also $B_\sigma \subseteq \mathcal{N}^N$ by

$$B_\sigma := \left\{\vec{a} \in \mathcal{N}^N : g(\vec{a}) = \sigma\right\} = \left\{\vec{a} \in \mathcal{N}^N : a_{\sigma(1)} \leq^1_\sigma a_{\sigma(2)} \leq^2_\sigma \dots \leq^{N-1}_\sigma a_{\sigma(N)}\right\},$$

where the order $\leq^i_\sigma$ is defined depending on σ by

$$\leq^i_\sigma := \begin{cases} \leq & \sigma(i) \leq \sigma(i+1), \\ < & \text{else.} \end{cases}$$

It follows from this representation that B_σ is Borel-measurable. We introduce

$$X^N \ni \vec{x} \mapsto g^N_X(\vec{x}) := (g_X(x_1), g_X(x_2), \dots, g_X(x_N)) \in F^N \subseteq \mathcal{N}^N.$$

Then the set

$$A_\sigma := \{\vec{x} \in X^N : g \circ g^N_X(\vec{x}) = \sigma\} = (g^N_X)^{-1}(B_\sigma),$$

is Borel-measurable. In particular, $A_{id} = A$ is Borel-measurable. Note that $\cup_\sigma A_\sigma = X^N$ and $A_{\sigma_1} \cap A_{\sigma_2} = \emptyset$ if $\sigma_1 \not\equiv \sigma_2$. We can apply Lemma 5.4, proving the continuity[1] of

$$\mathcal{N}^N \ni \vec{a} \mapsto G(a) := g(\vec{a})(\vec{a}) \in \mathcal{N}^N.$$

We define the candidate for the desired extension of f by

$$\begin{aligned} \hat{f} \colon X^N &\to Y^N, \\ \vec{x} &\mapsto (g \circ g^N_X(\vec{x}))^{-1}\left(f \circ (g^N_X)^{-1} \circ G \circ g^N_X(\vec{x})\right), \end{aligned}$$

which is well defined since $G \circ g^N_X(\vec{x}) \in \hat{A}$, so that $(g^N_X)^{-1} \circ G \circ g^N_X(\vec{x}) \in A$. As a composition of analytically measurable function, $\hat{f}$ inherits this property. It is also clear that $\hat{f}(\vec{x}) = f(\vec{x})$ if $\vec{x} \in A$. Finally, for any $\sigma \in S_N$ and $\vec{x} \in X^N$, we easily find

$$\sigma^{-1}(\hat{f} \circ \sigma(\vec{x})) = \hat{f}(\vec{x}).$$

□

Lemma 5.4. *Let each of $a, b \in \mathcal{N}^N$ be increasing vectors.*[2] *Then for any permutation $\sigma \in S_N$ we have*

$$\max_{i \in \{1,\dots,N\}} d_\mathcal{N}(a_i, b_i) \leq \max_{i \in \{1,\dots,N\}} d_\mathcal{N}(a_i, b_{\sigma(i)}), \tag{5.2}$$

where the metric $d_\mathcal{N}$ on $\mathcal{N}$ is given by

$$d_\mathcal{N}(a,b) = \begin{cases} 0 & a = b \\ \frac{1}{\min\{n \in \mathbb{N} \colon a(n) \neq b(n)\}} & \textit{else.} \end{cases}$$

Proof. We show the assertion by induction. For $N = 1$ (5.2) holds trivially. Now assume that (5.2) holds for $N = k$. Given $\sigma \in S_{k+1}$ and $a, b \in \mathcal{N}^{k+1}$ increasing, we know that any $\tilde{\sigma} \in S_k$ yields

$$\max_{i \in \{1,\dots,k\}} d_\mathcal{N}(a_i, b_i) \leq \max_{i \in \{1,\dots,k\}} d_\mathcal{N}(a_i, b_{\tilde{\sigma}(i)}).$$

If $\sigma(k+1) = k+1$ the assertion follows by the inductive hypothesis. So let $\sigma(k+1) \neq k+1$ and write $k_1 = \sigma(k+1)$ and $k_2 = \sigma^{-1}(k+1)$. Define a permutation $\hat{\sigma} \in S_k$ by

$$\hat{\sigma}(i) = \begin{cases} \sigma(i) & i \neq k_1 \\ k_2 & i = k_1 \end{cases}$$

Since that $a_{k_2} \leq a_{k+1}$ and $b_{k_1} \leq b_{k+1}$, then

$$\begin{aligned} a_{k_2} \leq b_{k_1} &\implies a_{k_2} \leq b_{k_1} \leq b_{k+1} \implies d_\mathcal{N}(a_{k_2}, b_{k_1}) \leq d_\mathcal{N}(a_{k_2}, b_{k+1}), \\ a_{k_2} \geq b_{k_1} &\implies a_{k+1} \geq a_{k_2} \geq b_{k_1} \implies d_\mathcal{N}(a_{k_2}, b_{k_1}) \leq d_\mathcal{N}(a_{k+1}, b_{k_1}), \end{aligned}$$

---

[1] In fact one obtains $\max_{i \in \{1,\dots,N\}} d_\mathcal{N}(g(\vec{a})(\vec{a})_i, g(\vec{b})(\vec{b})_i) \leq \max_{i \in \{1,\dots,N\}} d_\mathcal{N}(a_i, b_i)$, for $d_\mathcal{N}$ the metric on $\mathcal{N}$ that we recall in Lemma 5.4.

[2] A vector $v = (v_i)_{i=1}^N \in \mathcal{N}^N$ is increasing if for any $1 \leq i < j \leq N$ we have $v_i \leq v_j$, where inequality here is meant in the lexicographic order on $\mathcal{N}$.

and particularly

$$\max_{i\in\{1,\dots,k\}} d_{\mathcal{N}}(a_i, b_{\hat\sigma(i)}) \le \max_{i\in\{1,\dots,k+1\}} d_{\mathcal{N}}(a_i, b_{\sigma(i)}). \tag{5.3}$$

On the other hand, clearly

$$\begin{aligned} a_{k+1} \ge b_{k+1} &\implies d_{\mathcal{N}}(a_{k+1}, b_{k+1}) \le d_{\mathcal{N}}(a_{k+1}, b_{k_1}),\\ a_{k+1} \le b_{k+1} &\implies d_{\mathcal{N}}(a_{k+1}, b_{k+1}) \le d_{\mathcal{N}}(a_{k_2}, b_{k+1}). \end{aligned}$$

This and (5.3) yield $\max_{i\in\{1,\dots,k+1\}} d_{\mathcal{N}}(a_i, b_i) \le \max_{i\in\{1,\dots,k+1\}} d_{\mathcal{N}}(a_i, b_{\sigma(i)})$, so concluding the inductive step. □

5.2. **C-monotonicity: sufficiency.** The conditions under which Theorem 5.2 holds are rather mild. If we assume further continuity properties of C, the next theorem establishes that C-monotonicity is also a sufficient criterion for optimality, resembling the classical case. For weak transport costs, we don't know of any comparable result in the literature.

We recall that, for the given compatible complete metric d_Y on Y, we denote by $\mathcal{W}_1$ the 1-Wasserstein distance [35, Chapter 7].

Theorem 5.5. *Let $\nu \in \mathcal{P}^1_{d_Y}(Y)$. Assume that $C\colon X \times \mathcal{P}^1_{d_Y}(Y) \to \mathbb{R}$ satisfies condition* (A+) *and is $\mathcal{W}_1$-Lipschitz in the second argument is the sense that*

$$\text{for some } L \ge 0\colon\ |C(x,p) - C(x,q)| \le L\mathcal{W}_1(p,q),\ \forall x \in X, \forall p,q \in \mathcal{P}^1_{d_Y}(Y). \tag{5.4}$$

If π is C-monotone then π is an optimizer of $V(\mu,\nu)$.

In the proof we will use the following auxiliary result, which we will establish subsequently:

Lemma 5.6. *Let $\nu \in \mathcal{P}^1_{d_Y}(Y)$. Assume that $C\colon X \times \mathcal{P}^1_{d_Y}(Y) \to \mathbb{R}$ satisfies condition* (A+) *and is $\mathcal{W}_1$-Lipschitz in the sense of* (5.4)*. Then*

$$\inf_{\pi\in\Pi(\mu,\nu)} \int C(x,\pi_x)\,\mu(dx) = \sup_{\substack{\varphi\in\Phi_{b,1}\\ \|\varphi\|_{Lip}\le L}} \mu(R_C\varphi) - \nu(\varphi), \tag{5.5}$$

where $R_C\varphi$ is defined as in (3.1)*.*

Proof of Theorem 5.5. Let π be C-monotone. There is an increasing sequence $(K_n)_{n\in\mathbb{N}}$ of compact sets on Y such that $\nu(K_n) \nearrow 1$. From this we can refine the μ-full measurable set Γ in the definition of C-monotonicity, see Definition 5.1, so that for each $x \in \Gamma$ we have $\lim_n \pi_x(K_n) = 1$ and $\pi_x \in \mathcal{P}^1_{d_Y}(Y)$. Our goal is to construct a dual optimizer $\varphi \in \Phi_1$ to π such that

$$\pi_x(\varphi) + C(x,\pi_x) - R_C\varphi(x) = 0 \quad \forall x \in \Gamma.$$

When this is achieved, Theorem 1.3 and the following arguments show that π is optimal as desired:

$$\begin{aligned} \int_X C(x,\pi_x)\mu(dx) &= \int_\Gamma C(x,\pi_x)\mu(dx) = \int_\Gamma [R_C(\varphi)(x) - \pi_x(\varphi)]\mu(dx)\\ &\le \liminf_{k\to-\infty} \int_X [R_C(\varphi\vee k)(x) - \pi_x(\varphi\vee k)]\mu(dx)\\ &\le \sup_{\varphi\in\Phi_{b,1}} \mu(R_C\varphi) - \nu(\varphi)\\ &\le \inf_{\tilde\pi\in\Pi(\mu,\nu)} \int_X C(x,\tilde\pi_x)\mu(dx), \end{aligned}$$

where we used that

$$\liminf_{k\to-\infty} R_C(\varphi\vee k)(x) = \inf_{k\le 0} R_C(\varphi\vee k)(x) = R_C\varphi(x) \quad \forall x \in X.$$

Let us prove the existence of a dual optimizer in Φ_1. Let $G \subseteq \Gamma$ be a finite subset. By definition of C-monotonicity, we conclude that the coupling $\frac{1}{|G|}\sum_{x_i\in G} \delta_{x_i}(dx)\pi_{x_i}(dy)$ is optimal for the weak transport problem determined by the cost C and its first and second marginals. We can apply Lemma 5.6 in this context and obtain

$$\inf_{\|\varphi\|_{Lip}\le L}\sum_{x\in G}\pi_x(\varphi)+C(x,\pi_x)-R_C\varphi(x)=0. \tag{5.6}$$

We fix $y_0\in K_1$ and, without loss of generality, find a maximizing sequence $(\varphi_k)_{k\in\mathbb{N}}$ of (5.6) such that for all $k\in\mathbb{N}$ the function φ_k is L-Lipschitz and $\varphi_k(y_0)=0$. Note that for all $x\in G$

$$\pi_x(\varphi_k)+C(x,\pi_x)-R_C\varphi_k(x)\to 0,$$

since by definition $\pi_x(\varphi_k)+C(x,\pi_x)-R_C\varphi_k(x)\ge 0$. By the Arzelà-Ascoli theorem we find for any $n\in\mathbb{N}$ a subsequence of $(\varphi_k)_{k\in\mathbb{N}}$ and a L-Lipschitz continuous function ψ_n on K_n such that

$$\lim_j \varphi_{k_j}(y)=\psi_n(y)\quad \forall y\in K_n.$$

Thus by a diagonalization argument we can assume without loss of generality that the maximizing sequence converges uniformly for every K_n to a given L-Lipschitz function $\tilde\psi$ defined on

$$A:=\bigcup_n K_n.$$

We can extend $\tilde\psi$ from A to all of Y, obtaining an everywhere L-Lipschitz function, via

$$\psi(y)=\inf_{z\in A}\tilde\psi(z)+Ld_Y(z,y). \tag{5.7}$$

From (5.7) we find $R_C\psi(x)=\inf_{p\in\mathcal{P}^1_{d_Y}(A)}p(\psi)+C(x,p)$ and conclude

$$\limsup_k R_C\varphi_k(x)\le \inf_{p\in\mathcal{P}^1_{d_Y}(A)}p(\psi)+C(x,p)=R_C\psi(x). \tag{5.8}$$

By dominated convergence, and the fact that $\pi_x(A)=1$, we have

$$\lim_k \pi_x(\varphi_k)=\pi_x(\psi), \tag{5.9}$$

which yields

$$0=\liminf_k \pi_x(\varphi_k)+C(x,\pi_x)-R_C\varphi_k(x)\ge \pi_x(\psi)+C(x,\pi_x)-R_C\psi(x)\ge 0, \tag{5.10}$$

by definition of $R_C\psi(x)$.

For $G\subseteq Y$ define Ψ_G as the set of all L-Lipschitz continuous functions on A, vanishing at the point y', and satisfying

$$\pi_x(\psi)+C(x,\pi_x)-R_C\psi(x)=0\quad \forall x\in G.$$

The previous arguments show that, for each finite $G\subseteq\Gamma$, the set Ψ_G is nonempty. We now check that Ψ_G is closed in the topology of pointwise convergence: Let $(\psi_\alpha)_{\alpha\in\mathcal{I}}$ be a net in Ψ_G which converges pointwise to a function φ on A. Since A is the countable union of compact sets, it is possible to extract a sequence $(\psi_{\alpha_k})_{k\in\mathbb{N}}$ of the net such that

$$\psi_{\alpha_k}\to\varphi\quad \text{pointwise on } A \text{ and uniformly on each } K_n,$$

from which φ is L-Lipschitz on A and can be extended to an L-Lipschitz continuous function ψ on Y, see (5.7). By repeating previous arguments (see (5.8), (5.9) and (5.10)) we obtain that $\varphi\in\Psi_G$.

Note that Ψ_G is a closed subset of $\prod_{y\in A}[-Ld(y,y'),Ld(y,y')]$ which is compact in the topology of pointwise convergence by Tychonoff's theorem. Further, the collection $\{\Psi_G : G\subseteq\Gamma, |G|<\infty\}$ satisfies the finite intersection property, since if $G_1,\dots,G_n$ are finite then

$$\bigcap_{i\le n}\Psi_{G_i}\supseteq \Psi_{\cup_{i\le n}G_i}\neq\emptyset.$$

Therefore it is possible to find $\varphi\in\bigcap_{G\subseteq\Gamma,\ |G|<\infty}\Psi_G$. Again extend φ, from A to Y, by a L-Lipschitz function as usual. Thus, we have found the desired dual optimizer. □

Proof of Lemma 5.6. By Theorem 1.3 we have

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_X C(x,\pi_x)\mu(dx) = \sup_{\varphi\in\Phi_{b,1}} \mu(R_C\varphi) - \nu(\varphi). \tag{5.11}$$

By Theorem 1.2 we find a minimizer $\pi^* \in \Pi(\mu,\nu)$ of $V(\mu,\nu)$. Now we proceed by taking a maximizing sequence $(\varphi_k)_{k\in\mathbb{N}}$ for the right-hand side of (5.11). Note that we can choose each φ_k, in addition to being below-bounded and continuous, in a way such that it attains its infimum, i.e., there exists $y_k \in Y$ such that

$$-\infty < b_k := \inf_{y\in Y}\varphi_k(y) = \varphi_k(y_k). \tag{5.12}$$

Indeed, this can be done by using e.g. $\varphi_k \vee (b_k + \frac{1}{k})$ instead. Then

$$\lim_k \nu\left(\varphi_k - \varphi_k \vee \left(b_k + \frac{1}{k}\right)\right) = 0, \quad R_C\varphi_k \le R_C\left(\varphi_k \vee \left(b_k + \frac{1}{k}\right)\right),$$

and the following computation shows that $(\varphi_k \vee (b_k + \frac{1}{k}))_{k\in\mathbb{N}}$ is another maximizing sequence:

$$\begin{aligned} 0 &= \lim_k \int_X [\pi_x^*(\varphi_k) + C(x,\pi_x^*) - R_C\varphi_k(x)]\mu(dx) \\ &\ge \lim_k \int_X \left[\pi_x^*\left(\varphi_k \vee \left(b_k + \frac{1}{k}\right)\right) + C(x,\pi_x^*) - R_C\left(\varphi_k \vee \left(b_k + \frac{1}{k}\right)\right)(x)\right]\mu(dx) \ge 0. \end{aligned}$$

So let φ_k attain its infimum as in (5.12). We want to show that we can choose the sequence to be Lipschitz with constant L. For this purpose we infer additional properties of potential minimizers of $R_C\varphi_k$. Define for each function φ_k the Borel-measurable sets

$$A_k := \left\{ y \in Y \colon \sup_{y\neq z\in Y} \frac{\varphi_k(y) - \varphi_k(z)}{d_Y(y,z)} \le L \right\} \neq \emptyset,$$

$$\mathcal{Y}_k := \{(y,z) \in Y \times A_k \colon \varphi_k(y) - \varphi_k(z) > L d_Y(y,z)\}.$$

That $A_k \neq \emptyset$ follows since the minimizers of φ_k form a subset. We also stress that

$$\mathrm{proj}_1(\mathcal{Y}_k) = A_k^c.$$

Indeed, it is apparent that $\mathrm{proj}_1(\mathcal{Y}_k) \subseteq A_k^c$. To see the converse, assume $y \in A_k^c \cap \mathrm{proj}_1(\mathcal{Y}_k)^c$. Define $Z(z') := \{z \in Y \colon \varphi_k(z') - \varphi_k(z) > L d_Y(z,z')\}$. If there exists $\tilde{z} \in Z(y) \cap A_k$, we obtain a contradiction to $y \in \mathrm{proj}_1(\mathcal{Y}_k)^c$. Let $z_0 := y$ and inductively set $z_l \in Z(z_{l-1})$ such that

$$\inf_{z\in Z(z_{l-1})} \varphi_k(z) + \frac{1}{2^l} \ge \varphi_k(z_l). \tag{5.13}$$

We have for any natural numbers $0 \le i < n$

$$\varphi_k(z_i) - \varphi_k(z_n) = \sum_{l=i}^n \varphi_k(z_{l-1}) - \varphi_k(z_l) > L \sum_{l=i}^n d_Y(z_{l-1}, z_l). \tag{5.14}$$

The r.h.s. is bounded from below by $Ld_Y(z_i,z_n)$ and so as before we see that $z_n \in A_k$ provides a contradiction. We therefore assume for all l that $z_l \notin A_k$. The above inequality yields by lower-boundedness of φ_k that $(z_l)_{l\in\mathbb{N}}$ is a Cauchy sequence in Y. Writing $\bar{z}$ for its limit point, we conclude from (5.14) that $\varphi_k(z_i) - \varphi_k(\bar{z}) > Ld_Y(z_i,\bar{z})$ and consequentely $Z(\bar{z}) \subseteq Z(z_i)$. Since then $\inf\{\varphi_k(z) : z \in Z(z_i)\} \le \inf\{\varphi_k(z) : z \in Z(\bar{z})\}$ and from (5.13), we deduce $\inf\{\varphi_k(z) : z \in Z(\bar{z})\} \ge \varphi_k(\bar{z})$. Thus $Z(\bar{z}) = \emptyset$, implying $\bar{z} \in A_k$ and yielding a contradiction to $y \in \mathrm{proj}_1(\mathcal{Y}_k)^c$. All in all, we have proven that $A_k^c = \mathrm{proj}_1(\mathcal{Y}_k)$.

By Jankov-von Neumann uniformization [21, Theorem 18.1] there is an analytically measurable selection $T_k \colon \mathrm{proj}_1(\mathcal{Y}_k) \to A_k$. We set T_k on $A_k = \mathrm{proj}_1(\mathcal{Y}_k)^c$ as the identity.

Then T_k maps from Y to A_k and for any $p \in \mathcal{P}^t_{d_Y}(Y)$ we have

$$\begin{aligned}
C(x, T_k(p)) &\leq C(x, p) + L\mathcal{W}_1(p, T_k(p)) \\
&\leq C(x, p) + L \int_Y d_Y(y, T_k(y))p(dy) \\
&\leq C(x, p) + \int_Y [\varphi_k(y) - \varphi_k(T_k(y))]p(dy) \\
&= C(x, p) + p(\varphi_k) - T_k(p)(\varphi_k).
\end{aligned}$$

Therefore, we can assume that potential minimizers of $R_C\varphi_k$ are concentrated on A_k:

$$R_C\varphi_k(x) = \inf_{p\in\mathcal{P}^1_{d_Y}(Y)} p(\varphi_k) + C(x, p) = \inf_{p\in\mathcal{P}^1_{d_Y}(A_k)} p(\varphi_k) + C(x, p). \tag{5.15}$$

Thanks to $\mathrm{proj}_1(\mathcal{Y}_k) = A_k^c$, we introduce a family of L-Lipschitz continuous functions by

$$\psi_k(y) := \inf_{z\in A_k} \varphi_k(z) + Ld_Y(y, z) = \inf_{z\in Y} \varphi_k(z) + Ld_Y(y, z) \quad \forall y \in Y.$$

Then $\varphi_k \geq \psi_k$ where equality holds precisely on A_k. Similarly to before, we find a measurable selection $\hat{T}_k : Y \to A_k$ such that $\psi_k(\hat{T}_k(y)) + Ld_Y(y, \hat{T}_k(y)) \leq \psi_k(y) + \varepsilon$. For any $p \in \mathcal{P}^t_{d_Y}(Y)$ we have

$$C(x, \hat{T}_k(p)) \leq C(x, p) + L \int_Y d_Y(y, \hat{T}_k(y))p(dy) \leq C(x, p) + p(\psi_k) - \hat{T}_k(p)(\psi_k) + \varepsilon.$$

Since ε is arbitrary, by the same argument as in (5.15), we can restrict $\mathcal{P}^1_{d_Y}(Y)$ to $\mathcal{P}^1_{d_Y}(A_k)$ in the definition of $R_C\psi_k$. Hence, $R_C\varphi_k(x) = R_C\psi_k(x)$ and

$$\begin{aligned}
\int_X C(x, \pi^*_x)\mu(dx) &= \lim_k \int_X \left[-\pi^*_x(\varphi_k) + R_C\varphi_k(x)\right]\mu(dx) \\
&\leq \lim_k \int_X \left[-\pi^*_x(\psi_k) + R_C\psi_k(x)\right]\mu(dx) \\
&\leq \lim_k \int_X \left[-\pi^*_x(\psi_k) + \pi^*_x(\psi_k) + C(x, \pi^*_x)\right]\mu(dx) \\
&= \int_X C(x, \pi^*_x)\mu(dx).
\end{aligned}$$

□

6. On the Brenier-Strassen Theorem of Gozlan and Juillet

In this part we take $X = Y = \mathbb{R}^d$, equipped with the Euclidean metric, and

$$C_\theta(x, \rho) := \theta\left(x - \int y\rho(dy)\right),$$

where $\theta : \mathbb{R}^d \to \mathbb{R}_+$ is convex. As usual we denote by $V(\cdot, \cdot)$ the value of the weak transport problem with this cost functional (see (1.1)). We have

Lemma 6.1. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}^1(\mathbb{R}^d)$. Then*

$$\inf_{\eta\leq_c\nu} \inf_{\pi\in\Pi(\mu,\eta)} \int \theta(x - z)\pi(dx, dz) = V(\mu, \nu). \tag{6.1}$$

Proof. Given π feasible for $V(\mu, \nu)$, we define $T(x) := \int y\pi^x(dy)$ and notice that $T(\mu) \leq_c \nu$ by Jensen's inequality. From this we deduce that the l.h.s. of (6.1) is smaller than the r.h.s. For the reverse inequality, let $\varepsilon > 0$ and say $\bar{\eta} \leq_c \nu$ is such that

$$\inf_{\eta\leq_c\nu} \inf_{\pi\in\Pi(\mu,\eta)} \int \theta(x-z)\pi(dx, dz) + \varepsilon \geq \inf_{\pi\in\Pi(\mu,\bar{\eta})} \int \theta(x-z)\pi(dx, dz) \geq \int \theta(x-z)\bar{\pi}(dx, dz) - \varepsilon,$$

for some $\bar{\pi} \in \Pi(\mu, \bar{\eta})$. By Strassen theorem there is a martingale measure $m(dz, dy)$ with first marginal $\bar{\eta}$ and second marginal ν. Define $\pi(dx, dy) := \int_z \bar{\pi}^z(dx)m^z(dy)\bar{\eta}(dz)$, so then π

has x-marginal μ and y-marginal ν, and furthermore $\int y\pi^x(dx) = \int z\bar{\pi}^x(dx)$ (μ-a.s.), by the martingale property of m. Thus, by Jensen's inequality:

$$\int \theta(x-z)\bar{\pi}_x(dz)\mu(dx) \geq \int \theta\left(x - \int z\bar{\pi}_x(dz)\right)\mu(dx) = \int \theta\left(x - \int y\pi_x(dy)\right)\mu(dx) \geq V(\mu,\nu).$$

Taking $\varepsilon \to 0$ we conclude. □

We now provide the proof of Theorem 1.4, in which case $\theta(\cdot) = |\cdot|^2$:

Proof of Theorem 1.4. We have $V(\mu,\nu) < \infty$, since the product coupling yields a finite cost. Lemma 6.1 established the rightmost equality in (1.6). The existence of an optimizer π to $V(\mu,\nu)$ follows from Theorem 1.2. By the necessary monotonicity principle (Theorem 5.2) there exists a measurable set $\Gamma \subseteq X$ with $\mu(\Gamma) = 1$ such that for any finite number of points $x_1,\dots,x_N$ in Γ and measures $m^1,\dots,m^N$ in $\mathcal{P}(\mathbb{R}^d)$ with $\sum_{i=1}^N m^i = \sum_{i=1}^N \pi^{x_i}$ the following inequality holds:

$$\sum_{i=1}^N \left|x^i - \int y\pi^{x^i}(dy)\right|^2 \leq \sum_{i=1}^N \left|x^i - \int ym^i(dy)\right|^2. \tag{6.2}$$

In particular, if we let

$$T(x) := \int y\pi_x(dy),$$

and σ is any permutation, then

$$\sum_i \left|x^i - T(x^i)\right|^2 \leq \sum_{i=1}^N \left|x^i - T(x^{\sigma(i)})\right|^2. \tag{6.3}$$

Let us intruduce $p(dx,dz) := \mu(dx)\delta_{T(x)}(dz)$ and observe that its z-marginal is $T(\mu)$. By Rockafellar's theorem ([35, Theorem 2.27]) the support of p is contained in the graph of the subdifferential of a closed convex function. Then by the Knott-Smith optimality criterion ([35, Theorem 2.12]) the coupling p attains $\mathcal{W}_2(\mu,T(\mu))$. Since by Jensen clearly $T(\mu) \leq_c \nu$, this establishes the remaining equality in (1.6) and shows further that $V(\mu,\nu) = \mathcal{W}_2(\mu,T(\mu))^2$ and $\mu^* := T(\mu)$. The uniqueness of μ^* follows the same argument as in the proof of [17, Proposition 1.1].

We can use (6.2) and argue verbatim as in [17, Remark 3.1] showing that T is actually 1-Lipschitz on Γ. We will now prove that T is (μ-a.s. equal to) the gradient of a continuously differentiable convex function. The key remark is that the coupling p is also optimal for $V(\mu,T(\mu))$. Indeed, we have

$$V(\mu,\nu) \leq \inf_{\eta \leq_c T(\mu)} \mathcal{W}_2(\mu,\nu)^2 = V(\mu,T(\mu)) \leq \int |x - T(x)|^2\mu(dx) = V(\mu,\nu).$$

Let now K_n be an increasing sequene of compact sets such that $\mu(K_n) \nearrow 1$. Denote by μ_n the conditioning of μ to K_n, and let $\nu_n := T(\mu_n)$. Then both μ_n,ν_n have compact support, since T is Lipschitz. By the restiction result Proposition 4.2 we deduce that the coupling $p_n(dx,dy) := \mu_n(dx)\delta_{T(x)}(dy)$ is optimal for $V(\mu_n,\nu_n)$. We may apply [17, Theorem 1.2(b)] and obtain the existence of a convex continuously differentiable function φ_n whose gradient is 1-Lipschitz, and by optimality of p_n and [17, Theorem 1.2(c)], we have

$$T(x) = \nabla\varphi_n(x),\ \mu - a.e.\, x \in K_n. \tag{6.4}$$

This implies in particular $\nabla\varphi_n(x) = T(x)$ for μ-a.e. $x \in K_1$, so there exists $\bar{x} \in K_1 \subseteq K_n$ such that $\nabla\varphi_n(\bar{x}) = T(\bar{x})$ for all n. Thanks to this, and equicontinuity of $(\nabla\varphi_n)_n$, we may apply Arzelà-Ascoli locally, proving via a diagonalization argument that (modulo selection of a subsequence) there exists a 1-Lipschitz function $\bar{T}$ such that $\nabla\varphi_n \to \bar{T}$ locally uniformly. Without loss of generality we may assume $\varphi_n(0) = 0$. Some elementary calculus then shows that φ_n converges pointwise to a function φ. We deduce that φ is convex and differentiable, with $\nabla\varphi = \bar{T}$. It is therefore continuously differentiable. From (6.4) we derive that $T = \nabla\varphi$ μ-a.s. in K_n and so $T = \nabla\varphi$ μ-a.s. In particular $\mu^* = \nabla\varphi(\mu)$.

The arguments for the final sentence of Theorem1.4 are the same as in the proof of [17, Theorem 1.2(c)]. □

References

[1] A. Alfonsi, J. Corbetta, and B. Jourdain. Sampling of probability measures in the convex order and approximation of Martingale Optimal Transport problems. *ArXiv e-prints*, September 2017.

[2] Jean-Jacques Alibert, Guy Bouchitte, and Thierry Champion. A new class of cost for optimal transport planning. *hal-preprint*, 2018.

[3] Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

[4] J. Backhoff Veraguas, M. Beiglböck, M. Eder, and A. Pichler. Fine properties of process distances. *ArXiv e-prints*, 2017.

[5] J. Backhoff Veraguas, M. Beiglböck, M. Huesmann, and S. Källblad. Martingale Benamou–Brenier: a probabilistic perspective. *ArXiv e-prints*, August 2017.

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

[7] M. Beiglböck and C. Griessler. A land of monotone plenty. *Annali della SNS, to appear*, April 2016.

[8] M. Beiglböck and N. Juillet. On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.*, 44(1):42–106, 2016.

[9] M. Beiglböck, M. Nutz, and N. Touzi. Complete Duality for Martingale Optimal Transport on the Line. *Ann. Probab., to appear*, 2016.

[10] M. Beiglboeck, M. Eder, C. Elgert, and U. Schmock. Geometry of Distribution-Constrained Optimal Stopping Problems. *Probab. Theory Relat. Fields, to appear*, 2018.

[11] M. Beiglboeck and N. Juillet. Shadow couplings. *ArXiv e-prints*, September 2016.

[12] D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978. The discrete time case.

[13] Maria Colombo, Luigi De Pascale, and Simone Di Marino. Multimarginal optimal transport maps for one-dimensional repulsive costs. *Canad. J. Math.*, 67(2):350–368, 2015.

[14] Paul Dupuis and Richard S Ellis. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

[15] Max Fathi and Yan Shu. Curvature and transport inequalities for Markov chains in discrete spaces. *Bernoulli*, 24(1):672–698, 2018.

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

[17] Nathael Gozlan and Nicolas Juillet. On a mixture of brenier and strassen theorems. *arXiv preprint arXiv:1808.02681*, 2018.

[18] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, Yan Shu, and Prasad Tetali. Characterization of a class of weak transport-entropy inequalities on the line. *Ann. Inst. Henri Poincaré Probab. Stat.*, 54(3):1667–1693, 2018.

[19] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *J. Funct. Anal.*, 273(11):3327–3405, 2017.

[20] C. Griessler. c-cyclical monotonicity as a sufficient criterion for optimality in the multi-marginal Monge-Kantorovich problem. *ArXiv e-prints*, January 2016.

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

[22] Katalin Marton. A measure concentration inequality for contracting markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, 1996.

[23] Katalin Marton et al. Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.

[24] M. Nutz and F. Stebegg. Canonical Supermartingale Couplings. *Ann. Probab., to appear*, September 2018.

[25] B. Pass. On the local structure of optimal measures in the multi-marginal optimal transportation problem. *Calc. Var. Partial Differential Equations*, 43(3-4):529–536, 2012.

[26] G. Ch. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM J. Optim.*, 22(1):1–23, 2012.

[27] S. T. Rachev and L. Rüschendorf. A characterization of random variables with minimum L^2-distance. *J. Multivariate Anal.*, 32(1):48–54, 1990.

[28] S.T. Rachev and L. Rüschendorf. *Mass transportation problems. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, 1998.

[29] Paul-Marie Samson. Transport-entropy inequalities on locally acting groups of permutations. *Electron. J. Probab.*, 22:Paper No. 62, 33, 2017.

[30] Yan Shu. From hopf-lax formula to optimal weak transfer plan. *arXiv preprint arXiv:1609.03405*, 2016.

[31] Yan Shu. Hamilton-Jacobi equations on graph and applications. *Potential Anal.*, 48(2):125–157, 2018.

[32] Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

[33] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.
[34] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.
[35] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
[36] C. Villani. *Optimal Transport. Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.
[37] D. Zaev. On the Monge-Kantorovich problem with additional linear constraints. *ArXiv e-prints*, 2014.
[38] Constantin Zalinescu. *Convex analysis in general vector spaces*. World scientific, 2002.