Calculus of Variations

Existence, duality, and cyclical monotonicity for weak transport costs

J. Backhoff-Veraguas[1] · **M. Beiglböck**[1] · **G. Pammer**[1]

Abstract

The optimal weak transport problem has recently been introduced by Gozlan et al. (J Funct Anal 273(11):3327–3405, 2017). 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 and Juillet (On a mixture of brenier and strassen theorems. arXiv:1808.02681, 2018) to general probability measures on $\mathbb{R}^d$ under minimal assumptions. A driving idea behind our proofs is to consider the set of transport plans with a new ('adapted') topology which seems better suited for the weak transport problem and allows to carry out arguments which are close to the proofs in the classical setup.

1 Introduction

1.1 Notation

This article is concerned with the optimal transport problem for weak costs, as initiated by Gozlan et al. [25]. To state it [see (1.1) below] we introduce some basic notation. We write $\mathcal{P}(Z)$ for the set of probability measures on a Polish space Z is and equip $\mathcal{P}(Z)$ with the usual weak topology. Throughout X and Y are Polish spaces, $\mu \in \mathcal{P}(X)$, and $\nu \in \mathcal{P}(Y)$. 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 $(\pi_x)_{x \in X}$.

✉ M. Beiglböck
  mathias.beiglboeck@gmail.com

1   University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

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 and lower semicontinuous in an appropriate sense, and that $C(x, \cdot)$ is convex. The weak transport problem is then 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. [24,25] are mainly motivated by applications to geometric inequalities. Indeed, particular costs of the form (1.1) were already considered by Marton [29,30] and Talagrand [40,41]. Further papers directly related to [25] include [21,23,36–38]. 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 [7], in [1,2,6,15] it is used to investigate martingale optimal transport problems, in [3] it is applied to prove stability of pricing and hedging in mathematical finance, it appears in the characterization of optimal mechanism for the multiple good monopolist [19] and motivates the investigation of linear transfers in [17]. A more classical example is given by entropy-regularized optimal transport (i.e. the Schrödinger problem); see [28] and the references therein.

1.3 Main results

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. Sect. 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 $\mathcal{P}(X \times Y)$. In the present paper we circumvent this difficulty by embedding $\mathcal{P}(X \times Y)$ into the bigger space $\mathcal{P}(X \times \mathcal{P}(Y))$. This idea is borrowed from the investigation of process distances (cf. [4,5,33]) 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 Sect. 2 the following basic existence results.

Theorem 1.1 (Existence I) *Assume that $C : X \times \mathcal{P}(Y) \to \mathbb{R} \cup \{+\infty\}$ 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. prove existence of minimizers 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 Alibert et al. [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 and $t \in [1, \infty)$, 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. [42, Chapter 7]). In the sequel we make the 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 $C : X \times \mathcal{P}^t_{d_Y}(Y) \to \mathbb{R} \cup \{+\infty\}$ 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 Sect. 2.

1.3.2 Duality

We fix a compatible metric d_Y on Y and introduce the space

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

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 ([16, 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.

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 for all* $\mu \in \mathcal{P}(X)$ *and* $\nu \in \mathcal{P}^t_{d_Y}(Y)$

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

The proof of Theorem 1.3 is provided in Sect. 3. We also refer to this section for a comparison of earlier duality results of Gozlan et al. [25, Theorem 9.6] and Alibert et al. [2, Theorem 4.2].

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 monotonicity*; see [22,35] as well as the monographs [34,42,43]. More recently, variants of this 'monotonicity priniciple' have been applied in transport problems for finitely or infinitely many marginals [11,18,26, 32,44], the martingale version of the optimal transport problem [12,13,31], the Skorokhod embedding problem [9] and the distribution constrained optimal stopping problem [14].

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 have that every *C*-monotone transport plan is optimal under certain regularity assumptions. See Theorems 5.3 and 5.6 respectively.

We note that related concepts already appeared in [6, Proposition 4.1] (where necessity of a 2-step optimality condition is established) and in [23] (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.

We remark that the 2-step monotonicity principle for weak transport costs has already proved vital in [6] for the construction of a martingale counterpart to the Brenier theorem and the Benamou–Brenier formula. On the other hand, we conjecture that this monotonicity principle could be used in order to generalize [23] to non-quadratic costs.

1.3.4 A general Brenier–Strassen theorem

As an application of our abstract results we extend the Brenier–Strassen theorem [23, 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 [23] 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 $\leq_c$ 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* $\mu^* \leq_c \nu$ *such that*

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

Moreover, there exists a convex function $\varphi \colon \mathbb{R}^d \to \mathbb{R}$ *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. [24] for $d = 1$ and by Alfonsi et al. [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 [23] under the assumption of compactly supported measures μ, ν. The proof of Theorem 1.4 is given in Sect. 6.

Note added in revision In an updated version of [23], Gozlan and Juillet have also removed the compactness assumption in Theorem 1.4. Their proof is based on duality arguments and in particular differs from the one given here.

2 Existence of minimizers

A principal idea behind the proofs of this paper is to endow the set of transport plans $\mathcal{P}(X \times Y)$ with a topology that is finer than the usual weak topology and which appropriately accounts for the asymmetric role of X and Y in the context of weak transport. This can be formalized by embedding $\mathcal{P}(X \times Y)$ into the bigger space $\mathcal{P}(X \times \mathcal{P}(Y))$. I.e., given a transport plan π, we will consider its disintegration $(\pi_x)_{x \in X}$ (w.r.t. its first marginal) and view it as a *Monge-type coupling* in the larger space $\mathcal{P}(X \times \mathcal{P}(Y))$. It turns out that on this 'extended' space the minimization problems Theorems 1.1 and 1.2 can be handled more efficiently.

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 $\pi(dx \times Y)$, we consider the measurable map

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

We define the embedding $J \colon \mathcal{P}(X \times Y) \to \mathcal{P}(X \times \mathcal{P}(Y))$ by setting for $\pi \in \mathcal{P}(X \times Y)$ with X-marginal $\mu(dx) = \pi(dx \times Y)$

$$J(\pi) := (\kappa_\pi)_\#(\mu). \tag{2.1}$$

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 $\mathcal{P}(Y)$.

The intensity $I(P) \in \mathcal{P}(Y)$ of $P \in \mathcal{P}(\mathcal{P}(Y))$ is uniquely determined by

$$I(P)(f) = \int_{\mathcal{P}(Y)} p(f) P(dp) \quad \forall f \in C_b(Y). \tag{2.2}$$

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

$$\Lambda(\mu, \nu) := \left\{ P \in \mathcal{P}(X \times \mathcal{P}(Y)) \mid \mathrm{proj}_X P = \mu, \ I(\mathrm{proj}_{\mathcal{P}(Y)}(P)) = \nu \right\}. \tag{2.3}$$

Similar to (2.2), 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

$$\int_{X \times Y} f(x, y) \hat{I}(P)(dx, dy) = \int_{X \times \mathcal{P}(Y)} \int_Y f(x, y) p(dy) P(dx, dp) \quad \forall f \in C_b(X \times Y).$$
(2.4)

Note that while J is in general not continuous (cf. Example 2.2), the mappings I and $\hat{I}$ are continuous.

Using (2.1) and (2.4) we find that

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

Also note that $\hat{I}$ is the left-inverse of J, i.e., $\hat{I} \circ J(\pi) = \pi$ for $\pi \in \mathcal{P}(X \times Y)$. 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-bounded, and convex in the second argument. Then*

$$V(\mu, \nu) = \hat{V}(\mu, \nu),$$
(2.5)

where V was defined in (1.1) and

$$\hat{V}(\mu, \nu) := \inf_{P \in \Lambda(\mu, \nu)} \int_{X \times \mathcal{P}(Y)} C(x, p) P(dx, dp).$$
(2.6)

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 (2.4) 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

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

$$\geq \int_X C(x, \hat{I}(P)_x) \mu(dx)$$

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

$\square$

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 $\pi_x^k(dy) = \frac{\pi^k(x,dy)}{\pi^k(x \times Y)}$ 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

$$\lim_k |J(\pi^k)(f) - J(\pi)(f)|$$

$$\leq \limsup_k \sum_x |f(x, \pi_x^k)(\pi^k(x \times Y) - \pi(x \times Y))| + \sum_x |f(x, \pi_x^k) - f(x, \pi_x)|\pi(x \times Y)$$

$$= 0.$$

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

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

$$J(\lim_k \pi^k) = J\left(\frac{1}{2}\delta_x(p+q)\right) = \delta_{\left(x, \frac{1}{2}(p+q)\right)},$$

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 [39, 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 $\mathcal{P}_{d_Y}^t(Y)$ with the $\mathcal{W}_t$-topology instead of the weak topology.

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

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 $\mathcal{A} \subseteq \mathcal{P}_{d_Y}^t(Y)$ is relatively compact if and only if it is tight and*

$$\exists y' \in Y \, \forall \varepsilon > 0 \, \exists R > 0: \sup_{\mu \in \mathcal{A}} \int_{B_R(y')^c} d_Y(y, y')^t \mu(dy) < \varepsilon. \qquad (2.7)$$

Note that if (2.7) holds for some $y' \in Y$ it automatically holds for any $y' \in Y$.

Proof of Lemma 2.4 Since continuous maps preserve relative compactness in Hausdorff spaces, the first implication follows by continuity of I. To show the reverse implication, let $I(\mathcal{A})$ be relatively compact in $\mathcal{P}^t_{d_Y}(Y)$. First we show for fixed $y' \in Y$ that

$$\forall \varepsilon > 0 \, \exists R_\varepsilon > 0 \colon \; \sup_{P \in \mathcal{A}} \int_{\{p \colon \mathcal{W}_t(p, \delta_{y'})^t \geq R_\varepsilon\}} \mathcal{W}_t(p, \delta_{y'})^t \, P(dp) < \varepsilon. \qquad (2.8)$$

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

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

$$\int_{\mathcal{P}^t_{d_Y}(Y)} \int_{B_r(y')^c} d_Y(y, y')^t \, p(dy) P(dp) = \int_{B_r(y')^c} d_Y(y, y')^t I(P)(dy) < \frac{\varepsilon}{2}, \qquad (2.9)$$

where $B_r(y') = \{y \in Y \colon d_Y(y, y') < r\}$. Set $R_\varepsilon = \frac{2r^t K}{\varepsilon}$ and $A_{R_\varepsilon} = \left\{ p \in \mathcal{P}^t_{d_Y}(Y) \colon \mathcal{W}_t(p, \delta_{y'})^t \geq R_\varepsilon \right\}$, then

$$\sup_{P \in \mathcal{A}} P(A_{R_\varepsilon}) \leq \sup_{P \in \mathcal{A}} \frac{1}{R_\varepsilon} \int_{A_{R_\varepsilon}} \mathcal{W}_t(p, \delta_{y'})^t \, P(dp) \leq \frac{K}{R_\varepsilon}$$

and

$$\sup_{P \in \mathcal{A}} \int_{A_{R_\varepsilon}} \int_{B_r(y')} d_Y(y, y')^t \, p(dy) P(dp) \leq \sup_{P \in \mathcal{A}} P(A_{R_\varepsilon}) r^t \leq \frac{\varepsilon}{2}. \qquad (2.10)$$

Putting (2.9) and (2.10) together shows (2.8).

It remains to show that $\mathcal{A}$ is tight in $\mathcal{P}(\mathcal{P}^t_{d_Y}(Y))$. By Lemma 2.3 we have that $\mathcal{A}$ is tight in $\mathcal{P}(\mathcal{P}(Y))$, i.e., given $\varepsilon > 0$ there is a compact set $K_\varepsilon \subseteq \mathcal{P}(Y)$ such that for all $P \in \mathcal{A}$ we have $P(K_\varepsilon) \geq 1 - \varepsilon$. We will construct a set $\tilde{K}_\varepsilon \subseteq K_\varepsilon$ which is compact in $\mathcal{P}^t_{d_Y}(Y)$ and satisfies $P(\tilde{K}_\varepsilon) \geq 1 - 2\varepsilon$ in $P \in \mathcal{A}$. To this end, take a sequence of radii $(R_n)_{n \in \mathbb{N}}$ such that

$$\sup_{P \in \mathcal{A}} P\left(\left\{ p \colon \int_{\{y \colon d_Y(y, y')^t > R_n\}} d_Y(y, y')^t \, p(dy) \geq \frac{1}{n} \right\} \right) < \frac{\varepsilon}{2^n},$$

which is possible since

$$P\left(\left\{ p \colon \int_{\{y \colon d_Y(y, y')^t > R_n\}} d_Y(y, y')^t \, p(dy) \geq \frac{1}{n} \right\} \right) \leq n \int_{\{y \colon d_Y(y, y')^t > R_n\}} d_Y(y, y')^t I(P)(dy),$$

can be chosen sufficiently small, uniformly for $P \in \mathcal{A}$. The set

$$\tilde{K}_\varepsilon := \left\{ p \in K_\varepsilon \colon \int_{\{y \colon d_Y(y, y')^t > R_n\}} d_Y(y, y')^t \, p(dy) \leq \frac{1}{n}, \quad n \in \mathbb{N} \right\}$$

is compact in $\mathcal{P}^t_{d_Y}(Y)$ (cf. Lemma 2.5). Finally, given $P \in \mathcal{A}$ we obtain

$$P(\tilde{K}_\varepsilon) \geq P(K_\varepsilon) - \sum_n P\left(\left\{ p \colon \int_{\{y \colon d_Y(y, y')^t > R_n\}} d_Y(y, y') p(dy) \geq \frac{1}{n} \right\} \right) \geq 1 - 2\varepsilon$$

as desired $\qquad \square$

Proof of Lemma 2.5 '$\Rightarrow$': 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. Suppose for contradiction that (2.7) fails, i.e. there exist $y' \in Y$ and $\varepsilon > 0$ such that for all $N \in \mathbb{N}$ there is $\mu_N \in \mathcal{A}$ s.t.

$$\int_{B_N(y')^c} d_Y(y, y')^t \mu_N(dy) \geq \varepsilon.$$

In particular,

$$\lim_{R \to \infty} \liminf_N \int_{B_R(y')^c} d_Y(y', y)^t \mu_N(dy) \geq \varepsilon. \tag{2.11}$$

Due to relative compactness we find for any sequence in $\mathcal{A}$ an accumulation point. Then, from the definition of $\mathcal{W}_t$-convergence, see [43, Definition 6.8 (iii)], we deduce

$$\lim_{R \to \infty} \liminf_N \int_{B_R(y')^c} d_Y(y', y)^t \mu_N(dy) = 0,$$

which contradicts (2.11). Hence, (2.7) is satisfied.

'$\Leftarrow$': Let $\mathcal{A}$ be tight such that (2.7) 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

$$\int d_Y(y, y')^t \mu(dy) = \lim_{R \to \infty} \int R \wedge d_Y(y, y')^t \mu(dy)$$

$$\leq \lim_{R \to \infty} \liminf_{n \to \infty} \int R \wedge d_Y(y, y')^t \mu_n(dy) \leq \sup_n \int d_Y(y, y')^t \mu_n(dy).$$

Hence, by (2.7) we can choose (for $\varepsilon = 1$, say) $R > 0$ such that

$$\int_Y d_Y(y, y')^t \mu(dy) \leq \sup_n \int_{B_R(y')} d_Y(y, y')^t \mu_n(dy) + 1 < \infty,$$

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 $k \geq k_0$

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

Thus we have for $k \geq k_0$

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

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

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

$$d((x, y), (x_0, y_0)) = d_X(x, x_0) + d_Y(y, y_0), \tag{2.12}$$

$$\hat{d}((x, p), (x_0, p_0)) = d_X(x, x_0) + \mathcal{W}_t(p, p_0). \tag{2.13}$$

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. Conversely, if $\Lambda \in \mathcal{P}^t_{\hat{d}}(X \times \mathcal{P}^t_{d_Y}(Y))$ is relatively compact then $\hat{I}(\Lambda) \subseteq \mathcal{P}^t_d(X \times Y)$ is relatively compact.*

Proof Since continuous maps preserve relative compactness in Hausdorff spaces, we immediately deduce relative compactness of $\hat{I}(\Lambda)$, and 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.

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 $I(\Pi^Y_J)$ 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. $\qquad\square$

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.

$$p, q \in \mathcal{P}^t_{d_Y}(Y), \alpha \in [0, 1] \Rightarrow C(x, \alpha p + (1 - \alpha)q) \le \alpha C(x, p) + (1 - \alpha)C(x, q), \tag{2.14}$$

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*

$$\mathcal{P}^t_{\hat{d}}(X \times \mathcal{P}^t_{d_Y}(Y)) \ni P \mapsto \int_{X \times \mathcal{P}^t_{d_Y}(Y)} C(x, p) P(dx, dp) \tag{2.15}$$

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

$$\mathcal{P}_d^t(X \times Y) \ni \pi \mapsto \int_X C(x, \pi_x)\pi(dx \times Y) \tag{2.16}$$

is lower semicontinuous.

Proof Let $P^k \to P$ in $\mathcal{P}_{\hat{d}}^t(X \times \mathcal{P}_{d_Y}^t(Y))$. Similar to [20, Theorem A.3.12], we can approximate C from below by d-Lipschitz functions and obtain lower semicontinuity of (2.15), i.e.,

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

To show lower semicontinuity of (2.16), let $\pi^k \to \pi$ in $\mathcal{P}_d^t(X \times Y)$ and denote $P^k = J(\pi^k)$. We may assume that $\liminf_k \int_X C(x, \pi_x^k)\pi^k(dx \times Y) = \lim_k \int_X C(x, \pi_x^k)\pi^k(dx \times Y)$ by selecting a subsequence. By Lemma 2.6 we know that $\{P^k\}_k$ is relatively compact in $\mathcal{P}_{\hat{d}}^t(X \times \mathcal{P}_{d_Y}^t(Y))$. Denote by P an accumulation point of $\{P^k\}_k$. From now on we work along a subsequence converging to P. Observe that

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

Hence, we find by the first part that

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

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

$$\liminf_k \int_X C(x, \pi_x^k)\pi^k(dx \times Y) \geq \int_{X \times \mathcal{P}(Y)} C(x, p) \, P_x(dp)\pi(dx \times Y)$$

$$\geq \int_X C\left(x, \int_{\mathcal{P}(Y)} p(dy) P_x(dp)\right)\pi(dx \times Y).$$

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

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

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

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

$\square$

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* (2.6)]. *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 $\{P_k^* : k \in \mathbb{N}\}$ 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*

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

Remark 3.2 A proof of Theorem 1.3 can be obtained by means of [25, 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

$$F(m) := \inf_{P \in \Lambda(\mu,m)} \int_{X \times \mathcal{P}(Y)} \widetilde{C}(x, p)P(dx, dp)$$

$$= \inf_{P \in \Lambda(\mu,m)} \int_{X \times \mathcal{P}(Y)} C(x, p)P(dx, dp) + \int_Y d_Y(y, y_0)^t m(dy). \quad (3.4)$$

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 induced 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

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

Observe that $F^*(\psi) = \lim_{k \to +\infty} F^*(\psi \wedge k)$, by monotone convergence. We may apply the Fenchel duality theorem [45, Theorem 2.3.3], and then replace Φ_t by $\Phi_{b,t}$, obtaining:

$$F(m) = \sup_{\psi \in \Phi_t} m(\psi) - F^*(\psi)$$

$$= \sup_{-\psi \in \Phi_{b,t}} m(\psi) - F^*(\psi)$$

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

Now we show that

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

Rewriting (3.5) yields

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

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 [16, Proposition 7.47] and by [16, 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 $x \in X$

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

Then, we immediately obtain

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

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

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

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

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

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

$$
\inf_{P \in \Lambda(\mu,m)} P(C) = F(m) - \mathcal{W}_t(m, \delta_{y_0})^t
$$

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

$$
= \sup_{\psi \in \Phi_{b,t}} -m(\psi) + \int_X R_C\psi(x)\mu(dx),
$$

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 a simple 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 the following positive result.[1]

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.*

---

[1] In a preliminary version of this article the restriction property Proposition 4.2 was used to derive Theorem 1.4 from the compact version given by Gozlan and Juillet [23]. Following the insightful suggestion of the anonymous referee, we now give a more self contained argument that does not require Proposition 4.2 / [23]. We have decided to keep Proposition 4.2 since it might be of some independent interest.

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

$$\int C(x, \pi_x^*)\mu(dx) = \int C\left(x, \left(1 - \frac{d\tilde{\mu}}{d\mu}(x)\right)\pi_x + \frac{d\tilde{\mu}}{d\mu}(x)\chi_x\right)\mu(dx)$$

$$\leq \int C(x, \pi_x)\mu(dx) + \int [C(x, \chi_x) - C(x, \pi_x)]\tilde{\mu}(dx)$$

$$< \int C(x, \pi_x)\mu(dx),$$

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

5 C-Monotonicity for weak transport costs

Cyclical monotonicity plays a crucial role in classical optimal transport [22,35]. This has inspired similar development for weak transport costs in [6,23]:

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 $x_1, \ldots, x_N$ in Γ and measures $m_1, \ldots, m_N$ in $\mathcal{P}(Y)$ with $\sum_{i=1}^{N} m_i = \sum_{i=1}^{N} \pi_{x_i}$, the following inequality holds:

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

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 $\{1, \ldots, N\}$. If $\vec{z} := (z_1 \ldots, z_n)$ is any N-vector, and $\sigma \in S_N$, we naturally overload the notation by defining

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

Recall the notation (1.1) for the weak transport problem, and the following lemma, which is employed prominently in the proof of Theorem 5.3.

Lemma 5.2 ([10], Proposition 2.1]) *Let* $X_1, \ldots, X_n$, $n \geq 2$, *be Polish spaces equipped with probability measures* $\mu_i \in \mathcal{P}(X_i)$, $i = 1, \ldots, n$. *Then for any analytic set* $B \subseteq X_1 \times \cdots \times X_n$ *one of the following holds:*

(a) *For every $i = 1, \ldots, n$ there is a μ_i-null set $A_i \subseteq X_i$ s.t.*

$$B \subseteq \bigcup_{i=1}^{n} \mathrm{proj}_{X_i}^{-1}(A_i).$$

(b) *There exists a coupling $\pi \in \Pi(\mu_1, \ldots, \mu_n)$ with $\pi(B) > 0$.*

The previous lemma is originally stated only for Borel sets, but the same proof technique also works for analytic sets.

Our main result, concerning the necessity of C-monotonicity is the following:

Theorem 5.3 *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

$$\mathcal{D}_N := \left\{ ((x_1, \ldots, x_N), (m_1, \ldots, m_N)) \in X^N \times \mathcal{P}(Y)^N : \right.$$

$$\left. \sum_{i=1}^{N} \pi^*_{x_i} = \sum_{i=1}^{N} m_i \text{ and } \sum_{i=1}^{N} C(x_i, \pi^*_{x_i}) > \sum_{i=1}^{N} C(x_i, m_i) \right\},$$

is an analytic set. Write

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

By Jankov-von Neumann uniformization [27, Theorem 18.1] there is an analytically measurable function $f_N \colon D_N \to \mathcal{P}(Y)^N$ such that $\mathrm{graph}(f_N) \subseteq \mathcal{D}_N$. We can extend f_N to X^N by defining it on $X^N \setminus D_N$ as the Borel-measurable map $\vec{x} \mapsto (\pi^*_{x_1}, \ldots, \pi^*_{x_N})$. Observe that for all $\sigma \in S_N$, we have $(\sigma, \sigma)(\mathcal{D}_N) = \mathcal{D}_N$. Thanks to this, and Lemma 5.4 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 ith element of the vector $f_N(\vec{x}) \in \mathcal{P}(Y)^N$.

Assume that there exists a coupling $Q \in \Pi(\mu^N) = \Pi(\mu, \ldots, \mu)$ 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

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

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

$$\int_X \mu(dx_1)\tilde{\pi}_{x_1}(dy) = \int_{X^N} Q(dx_1, \ldots, dx_N) f_N^1(\vec{x})(dy)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int_{X^N} \sigma_i(Q)(dx_1, \ldots, dx_N) f_N^i(\vec{x})(dy)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int_{X^N} Q(dx_1, \ldots, dx_N) \pi_{x_i}^*(dy)$$

$$= \nu(dy).$$

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

$$\int_X C(x, \tilde{\pi}_x)\mu(dx) = \int_X \mu(dx_1) C\left(x_1, \int_{X^{N-1}} Q_{x_1}(dx_2, \ldots, dx_N) f_N^1(\vec{x})\right)$$

$$\le \int_{X^N} Q(d\vec{x}) C\left(x_1, f_N^1(\vec{x})\right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int_{X^N} Q(d\vec{x}) C\left(x_i, f_N^i(\vec{x})\right)$$

$$< \frac{1}{N} \sum_{i=1}^{N} \int_{X^N} Q(d\vec{x}) C\left(x_i, \pi_{x_i}\right) = \int_X C(x, \pi_x)\mu(dx),$$

yielding a contradiction to the optimality of π^*.

We conclude that no measure Q with the stated properties exists. By Lemma 5.2, we obtain that D_N is contained in a set of the form $\bigcup_{k=1}^{N} \text{proj}_k^{-1}(M_N)$ where $\mu(M_N) = 0$ and proj_k denotes the projection from X^N to its kth 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. $\qquad\square$

The missing bit in the above proof is Lemma 5.4. By [27, 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 $h_X \colon F \to X$. 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:

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

Lemma 5.4 *The set*

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

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

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

Proof of Lemma 5.4 Let $\hat{A} = \{\vec{a} \in \mathcal{N}^N : a_1 \leq a_2 \leq \cdots \leq a_N\}$, and define $g \colon \mathcal{N}^N \to S_N$ by $g(\vec{a}) = \sigma$ where $\sigma \in S_N$ satisfies

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

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

where the order $\leq_\sigma^i$ is defined depending on σ by

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

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

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

Then the set

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

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 5.5, proving the continuity[2] 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

$$\hat{f} \colon X^N \to Y^N,$$

$$\vec{x} \mapsto (g \circ g_X^N(\vec{x}))^{-1} \left(f \circ (g_X^N)^{-1} \circ G \circ g_X^N(\vec{x}) \right),$$

which is well defined since $G \circ g_X^N(\vec{x}) \in \hat{A}$, so that $(g_X^N)^{-1} \circ G \circ g_X^N(\vec{x}) \in A$. 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

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

$\square$

Lemma 5.5 *Let each of $a, b \in \mathcal{N}^N$ be increasing vectors.[3] Then for any permutation $\sigma \in S_N$ we have*

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

---

[2] 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 5.5.

[3] 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 $\mathcal{N}$.

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

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

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

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

and particularly

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

On the other hand, clearly

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

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

5.2 C-monotonicity: sufficiency

The conditions under which Theorem 5.3 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 [42, Chapter 7].

Theorem 5.6 *Let $\nu \in \mathcal{P}^1_{d_Y}(Y)$. Assume that $C: X \times \mathcal{P}^1_{d_Y}(Y) \to \mathbb{R}$ satisfies Condition* (A+) *and is $\mathcal{W}_1$-Lipschitz in the second argument is the sense that for some $L \geq 0$:*

$$|C(x, p) - C(x, q)| \leq L\mathcal{W}_1(p, q), \quad \forall x \in X, \forall p, q \in \mathcal{P}^1_{d_Y}(Y). \tag{5.4}$$

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 5.7 *Let $v \in \mathcal{P}^1_{d_Y}(Y)$. 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*

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

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

Proof of Theorem 5.6 Let π be C-monotone. There is an increasing sequence $(K_n)_{n \in \mathbb{N}}$ of compact sets on Y such that $v(K_n) \nearrow 1$. 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:

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

where we used that

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

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 5.7 in this context and obtain

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

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 $\pi_x(\varphi_k) + C(x, \pi_x) - R_C \varphi_k(x) \geq 0$. 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 $R_C\psi(x) = \inf_{p \in \mathcal{P}^1_{d_Y}(A)} p(\psi) + C(x, p)$. Indeed, by [16, Proposition 7.50] there is for any $\varepsilon > 0$ an analytically measurable function $T_\varepsilon \colon Y \to A$ with

$$\tilde{\psi}(T_\varepsilon(y)) + Ld_Y(T_\varepsilon(y), y) \le \psi(y) + \varepsilon,$$

from which after integrating with respect to p and using the definition of the Wasserstein distance we deduce

$$p(\psi) - T_\varepsilon(p)(\psi) + C(x, p) - C(x, T_\varepsilon(p))$$
$$\ge -\varepsilon + L\mathcal{W}_1(p, T_\varepsilon(p)) + C(x, p) - C(x, T_\varepsilon(p)) \ge -\varepsilon,$$

where we used (5.4) in the last inequality. Therefore, it is actually possible to restrict infimum in $R_C\psi(x)$ to $\mathcal{P}^1_{d_Y}(A)$, and we 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_0, 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 $(\psi_\alpha)_{\alpha \in I}$ 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)–(5.10)] we obtain that $\varphi \in \Psi_G$.

Note that Ψ_G is a closed subset of $\prod_{y \in A}[-Ld(y, y_0), Ld(y, y_0)]$ 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 $G_1, \ldots, G_n$ are finite then

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

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. $\qquad\square$

Proof of Lemma 5.7 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. $\varphi_k \vee \left(b_k + \frac{1}{k}\right)$ instead. Then

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

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

$$0 = \lim_k \int_X [\pi_x^*(\varphi_k) + C(x, \pi_x^*) - R_C\varphi_k(x)]\mu(dx)$$

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

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

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

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

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

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

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

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

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

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

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 consequently $Z(\bar{z}) \subseteq Z(z_i)$. Since then $\inf\{\varphi_k(z) : z \in Z(z_i)\} \leq \inf\{\varphi_k(z) : z \in Z(\bar{z})\}$ and from (5.13), we deduce $\inf\{\varphi_k(z) : z \in Z(\bar{z})\} \geq \varphi_k(\bar{z})$. Thus $Z(\bar{z}) = \emptyset$, implying $\bar{z} \in A_k$ and yielding a contradiction to $y \in \text{proj}_1(\mathcal{Y}_k)^c$. All in all, we have proven that $A_k^c = \text{proj}_1(\mathcal{Y}_k)$.

By Jankov-von Neumann uniformization [27, Theorem 18.1] there is an analytically measurable selection $T_k\colon \text{proj}_1(\mathcal{Y}_k) \to A_k$. We set T_k on $A_k = \text{proj}_1(\mathcal{Y}_k)^c$ 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:

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

We introduce a family of L-Lipschitz continuous functions by

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

where equality holds thanks to $\text{proj}_1(\mathcal{Y}_k) = A_k^c$, since for $z \in A_k^c$ we find $(z, \hat{z}) \in \mathcal{Y}_k$, and so

$$
\varphi_k(z) + Ld_Y(y, z) > \varphi_k(\hat{z}) + L(d_Y(y, z) + d_Y(z, \hat{z})) \geq \varphi_k(\hat{z}) + Ld_Y(y, \hat{z}).
$$

Then $\varphi_k \geq \psi_k$ where equality holds precisely on A_k. Similarly to before, we find a measurable selection $\hat{T}_k\colon Y \to A_k$ 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

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

$\square$

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 5.3) there exists a measurable set $\Gamma \subseteq X$ with $\mu(\Gamma) = 1$ such that for any finite number of points $x_1, \ldots, x_N$ in Γ and measures $m^1, \ldots, m^N$ 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

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

Let us introduce $p(dx, dz) := \mu(dx)\delta_{T(x)}(dz)$ and observe that its z-marginal is $T(\mu)$. By Rockafellar's theorem ([42, 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 ([42, Theorem 2.12]) the coupling p attains $\mathcal{W}_2(\mu, T(\mu))$. Since by Jensen clearly $T(\mu) \le_c \nu$, 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 [23, Proposition 1.1].

We can use (6.2) and argue verbatim as in [23, 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

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

Take any $\mathcal{W}_2$-approximative sequence $(\mu^k)_{k \in \mathbb{N}}$ of μ such that for all $k \in \mathbb{N}$

$$\mu^k \ll \lambda \ll \mu^k,$$

where λ denotes the d-dimensional Lebesgue measure. This can be easily achieved by scaled convolution with a non-degenerate Gaussian kernel. By stability of the considered weak transport problem [8, Theorem 1.5], and using the previously shown, we obtain for each μ^k a 1-Lipschitz map T^k defined this time everywhere in $\mathbb{R}^d$ with

$$\mathcal{W}_2(\mu^k, T^k(\mu^k))^2 = V(\mu^k, \nu),$$

and $T^k(\mu^k) \to T(\mu)$ in $\mathcal{W}_1$. By Brenier's theorem [42, Theorem 2.12 (ii)] we find for each $k \in \mathbb{N}$ some convex function $\varphi^k : \mathbb{R}^d \to \mathbb{R}$, $\varphi(0) = 0$, and $\nabla \varphi^k(x) = T^k(x)$ λ-a.e. x. By continuity of T^k we have $\nabla \varphi^k(x) = T^k(x)$ for all $x \in \mathbb{R}^d$.

We want to show that $(\varphi^k)_{k \in \mathbb{N}}$ is suitably relatively compact. By tightness of μ^k and $T^k(\mu^k)$ we find compact sets $K_1, K_2 \subseteq \mathbb{R}^d$ with

$$\inf_k \mu^k(K_1) > \frac{1}{2}, \quad \inf_k T^k(\mu^k)(K_2) > \frac{1}{2}.$$

In particular, the sets $(T^k(K_1) \cap K_2)_{k \in \mathbb{N}}$ are all non-empty. The compactness of K_1 and K_2, and the 1-Lipschitz property of each T^k, imply then the existence of $x \in K_1$ such that $\sup_k |T^k(x)| < \infty$. Hence, $(T^k)_{k \in \mathbb{N}}$ is pointwise bounded and uniformly 1-Lipschitz. Thanks to Arzelà-Ascoli's theorem and a diagonalization argument, we can select a subsequence $(T^{k_j})_{j \in \mathbb{N}}$ of $(T^k)_{k \in \mathbb{N}}$ which converges locally uniformly to some 1-Lipschitz function $\tilde{T} : \mathbb{R}^d \to \mathbb{R}^d$. Since being a gradient field is preserved under locally uniform limits, we have that $\tilde{T}$ is a gradient field, and φ^{k_j} converges pointwise to some φ with $\varphi(0) = 0$ and $\nabla \varphi = \tilde{T}$. In particular φ is convex and of class $C^1(\mathbb{R}^d)$.

Finally, for any $f \in C_b(\mathbb{R}^d)$ and $\varepsilon > 0$, we find an index $j_0 \in \mathbb{N}$ such that for all $j \ge j_0$:

$$|T^{k_j}(\mu^{k_j})(f) - \tilde{T}(\mu)(f)| \le |T^{k_j}(\mu^{k_j})(f) - \tilde{T}(\mu^{k_j})(f)| + |\tilde{T}(\mu^{k_j})(f) - \tilde{T}(\mu)(f)| < \varepsilon,$$

where the first summand can be chosen sufficiently small for large j by locally uniform convergence of T^{k_k} to $\tilde{T}$ and the second one by weak convergence of μ^{k_j} to μ. All in all,

we deduce that $T^{k_j}(\mu^{k_j})$ converges weakly to $\tilde{T}(\mu)$, which must therefore match $T(\mu)$. Furthermore, $\mu(dx)\delta_{\tilde{T}(x)}(dy)$ defines an optimizer for the weak transport problem (1.1) between μ and ν with cost (1.5). By uniqueness of the optimizers we conclude $T = \tilde{T}$ μ-almost surely. In particular, T is μ-almost everywhere the gradient of the convex function $\varphi \in C^1(\mathbb{R}^d)$. $\qquad\square$

References

1. Alfonsi, A., Corbetta, J., Jourdain, B.: Sampling of probability measures in the convex order and approximation of Martingale Optimal Transport problems. arXiv e-prints, Sept (2017)
2. Alibert, J.-J., Bouchitte, G., Champion, T.: A new class of cost for optimal transport planning. hal-preprint, (2018)
3. Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., Eder, M.: Adapted wasserstein distances and stability in mathematical finance. arXiv e-prints, page arXiv:1901.07450, Jan (2019)
4. Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., Eder, M.: All adapted topologies are equal. ArXiv e-prints, page arXiv:1905.00368, May (2019)
5. Backhoff-Veraguas, J., Beiglböck, M., Eder, M., Pichler,A.: Fundamental properties of process distances. ArXiv e-prints, (2017)
6. Backhoff-Veraguas, J., Beiglböck, M., Huesmann, M., Källblad,S.: Martingale Benamou–Brenier: a probabilistic perspective. ArXiv e-prints, Aug (2018)
7. Backhoff-Veraguas, J., Beiglböck, M., Lin, Y., Zalashko, A.: Causal transport in discrete time and applications. SIAM J. Optim. **27**(4), 2528–2562 (2017)
8. Backhoff-Veraguas, J., Beiglböck, M., Pammer,G.: Weak monotone rearrangement on the line. ArXiv e-prints, page arXiv:1902.05763, Feb (2019)
9. Beiglböck, M., Cox, A., Huesmann, M.: Optimal transport and Skorokhod embedding. Invent. Math. **208**(2), 327–400 (2017)
10. Beiglböck, M., Goldstern, M., Maresch, G., Schachermayer, W.: Optimal and better transport plans. J. Funct. Anal. **256**(6), 1907–1927 (2009)
11. Beiglböck, M., Griessler,C.: A land of monotone plenty. Annali della SNS, Apr (2016) **(to appear)**
12. Beiglböck, M., Juillet, N.: On a problem of optimal transport under marginal martingale constraints. Ann. Probab. **44**(1), 42–106 (2016)
13. Beiglböck, M., Nutz, M., Touzi, N.: Complete duality for martingale optimal transport on the Line. Ann. Probab. (2016) **(to appear)**
14. Beiglboeck, M., Eder, M., Elgert, C., Schmock, U.: Geometry of distribution-constrained optimal stopping problems. Probab. Theory Relat. Fields (2018) **(to appear)**
15. Beiglboeck, M., Juillet, N.: Shadow couplings. ArXiv e-prints, Sept (2016)
16. Bertsekas, D.P., Shreve, S.E.: Stochastic Optimal Control, the Discrete Time Volume 139 Case of Mathematics in Science and Engineering. Academic Press, Inc., New York (1978)
17. Bowles,M., Ghoussoub, N.: A Theory of transfers: duality and convolution. ArXiv e-prints, page arXiv:1804.08563, Apr (2018)
18. Colombo, M., De Pascale, L., Di Marino, S.: Multimarginal optimal transport maps for one-dimensional repulsive costs. Can. J. Math. **67**(2), 350–368 (2015)
19. Daskalakis, C., Deckelbaum, A., Tzamos, C.: Strong duality for a multiple-good monopolist. Econometrica **85**(3), 735–767 (2017)
20. Dupuis, P., Ellis, R.S.: A Weak Convergence Approach to the Theory of Large Deviations, vol. 902. Wiley, Hoboken (2011)
21. Fathi, M., Shu, Y.: Curvature and transport inequalities for Markov chains in discrete spaces. Bernoulli **24**(1), 672–698 (2018)
22. Gangbo, W., McCann, R.: The geometry of optimal transportation. Acta Math. **177**(2), 113–161 (1996)

23. Gozlan, N., Juillet,N.: On a mixture of brenier and strassen theorems. ArXiv preprint arXiv:1808.02681, (2018)
24. Gozlan, N., Roberto, C., Samson, P.-M., Shu, Y., Tetali, P.: Characterization of a class of weak transport-entropy inequalities on the line. Ann. Inst. Henri Poincaré Probab. Stat. **54**(3), 1667–1693 (2018)
25. Gozlan, N., Roberto, C., Samson, P.-M., Tetali, P.: Kantorovich duality for general transport costs and applications. J. Funct. Anal. **273**(11), 3327–3405 (2017)
26. Griessler, C.: c-cyclical monotonicity as a sufficient criterion for optimality in the multi-marginal Monge–Kantorovich problem. ArXiv e-prints, Jan (2016)
27. Kechris, A.S.: Classical Descriptive Set Theory, Volume 156 of Graduate Texts in Mathematics. Springer, New York (1995)
28. Léonard, C.: A survey of the Schrödinger problem and some of its connections with optimal transport. Discret. Contin. Dyn. Syst. **34**(4), 1533–1574 (2014)
29. Marton, K.: A measure concentration inequality for contracting Markov chains. Geom. Funct. Anal. GAFA **6**(3), 556–571 (1996)
30. Marton, K., et al.: Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. Ann. Probab. **24**(2), 857–866 (1996)
31. Nutz, M., Stebegg, F.: Canonical Supermartingale Couplings. Ann. Probab. Sept (2018) **(to appear)**
32. Pass, B.: On the local structure of optimal measures in the multi-marginal optimal transportation problem. Calc. Var. Partial Differ. Equ. **43**(3–4), 529–536 (2012)
33. Pflug, G.C., Pichler, A.: A distance for multistage stochastic optimization models. SIAM J. Optim. **22**(1), 1–23 (2012)
34. Rachev, S., Rüschendorf, L.: Mass Transportation Problems. Vol. I. Probability and Its Applications. Springer, New York (1998)
35. Rachev, S.T., Rüschendorf, L.: A characterization of random variables with minimum L^2-distance. J. Multivar. Anal. **32**(1), 48–54 (1990)
36. Samson, P.-M.: Transport-entropy inequalities on locally acting groups of permutations. Electron. J. Probab. **22**(62), 33 (2017)
37. Shu,Y.: From Hopf–Lax formula to optimal weak transfer plan. ArXiv preprint arXiv:1609.03405, (2016)
38. Shu, Y.: Hamilton–Jacobi equations on graph and applications. Potential Anal. **48**(2), 125–157 (2018)
39. Sznitman, A.-S.: Topics in propagation of chaos. In: Ecole d'été de probabilités de Saint-Flour XIX—1989, pp. 165–251. Springer, (1991)
40. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces. Publ. Math. de l'Institut des Hautes Etudes Sci. **81**(1), 73–205 (1995)
41. Talagrand, M.: New concentration inequalities in product spaces. Invent. Math. **126**(3), 505–563 (1996)
42. Villani, C.: Topics in optimal transportation, Volume 58 of Graduate Studies in Mathematics. American Mathematical Society, Providence (2003)
43. Villani, C.: Optimal Transport Old and New, Volume 338 of Grundlehren der Mathematischen Wissenschaften. Springer, Berlin (2009)
44. Zaev,D.: On the Monge–Kantorovich problem with additional linear constraints. ArXiv e-prints, (2014)
45. Zalinescu, C.: Convex Analysis in General Vector Spaces. World scientific, Singapore (2002)