Optimal Transport Theory: The Essentials

YoonHaeng Hur

July 21, 2025

Preface Due to its wide range of applications in various fields, optimal transport has attracted much attention from multiple disciplines. Thanks to this, the boundary of optimal transport theory is continuously expanding, leading to a variety of perspectives and approaches. Those who wish to learn the theory of optimal transport for the first time are recommended to consult the following excellent textbooks.

- [Vil03] *Topics in Optimal Transportation* (2003)
- [AGS05] *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (2005)
- [Vil09] *Optimal Transport: Old and New* (2009)
- [San15] *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling* (2015)
- [ABS21] *Lectures on Optimal Transport* (2021)
- [FG21] *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows* (2021)

Depending on the reader's background and interest, preferences may vary. In any case, however, optimal transport theory is a vast and deep subject that requires a solid foundation in mathematical analysis. The goal of this note is to provide a concise and rigorous summary of the essential results in optimal transport theory which I believe are foundational for further study. The target audience is expected to be familiar with measure theory, topology, and convex analysis. Although applied researchers may be interested in more practical aspects of optimal transport, it would be beneficial to try—at least once—to understand the core mathematical foundations.

Notation and Preliminaries

General notation For $n \in \mathbb{N}$, define $[n] = \{1, \dots, n\}$ and let $\mathrm{Perm}(n)$ denote the collection of all permutations of $[n]$. For $a, b \in [-\infty, \infty]$, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For a map $T \colon \mathcal{X} \to \mathcal{Y}$ between any sets $\mathcal{X}$ and $\mathcal{Y}$, let $\mathrm{graph}(T)$ denote its graph, i.e.,

$$\mathrm{graph}(T) = \{(x, T(x)) : x \in \mathcal{X}\} \subset \mathcal{X} \times \mathcal{Y}.$$

Any function taking values in $[-\infty, \infty]$ is said to be proper if it is not identically ∞ or $-\infty$.

Measure and integration Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space. We say μ is concentrated on $A \in \mathcal{A}$ if $\mu(\mathcal{X} \backslash A) = 0$. For any measurable function $f \colon \mathcal{X} \to [-\infty, \infty]$,

$$\int_{\mathcal{X}} f \, \mathrm{d}\mu = \int_{\mathcal{X}} f^+ \, \mathrm{d}\mu - \int_{\mathcal{X}} f^- \, \mathrm{d}\mu \quad \text{if } \int_{\mathcal{X}} f^+ \, \mathrm{d}\mu < \infty \text{ or } \int_{\mathcal{X}} f^- \, \mathrm{d}\mu < \infty,$$

where $f^+ = f \vee 0$ and $f^- = (-f) \vee 0$. We write $f \in L^1(\mu)$ if $\int_{\mathcal{X}} |f| \, \mathrm{d}\mu < \infty$. Let δ_x denote the Dirac measure concentrated at $x \in \mathcal{X}$, i.e., $\delta_x(A) = 1$ if $x \in A$, and $\delta_x(A) = 0$ if $x \notin A$. Lastly, we call μ a probability measure if $\mu(\mathcal{X}) = 1$; in this case, we call $(\mathcal{X}, \mathcal{A}, \mu)$ a probability space.

Pushforward measure Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and let $(\mathcal{Y}, \mathcal{B})$ be a measurable space. Given a measurable map $T \colon \mathcal{X} \to \mathcal{Y}$, let $T_\# \mu$ denote the pushforward measure of μ by T, namely, $T_\# \mu$ is a measure on $(\mathcal{Y}, \mathcal{B})$ such that

$$T_\# \mu(B) = \mu\{x \in \mathcal{X} : T(x) \in B\} \quad \forall B \in \mathcal{B}.$$

In this case, for any measurable $f \colon \mathcal{Y} \to [0, \infty]$,

$$\int_{\mathcal{X}} f \circ T \, \mathrm{d}\mu = \int_{\mathcal{Y}} f \, \mathrm{d}\nu.$$

Topology A topological space is said to be separable if it has a countable dense subset. The product of two separable spaces is separable. A topological space is said to be metrizable if there is a metric that generates the topology; we call such a metric a compatible metric. We call a metrizable topological space simply a metrizable space. A topological space is called a Polish space if it is separable and metrizable by a complete metric. The product of two Polish spaces is a Polish space. A subset of a Polish space is a Polish space if and only if it is a G_δ set; see [Coh13]. For a topological space $\mathcal{X}$, let $C(\mathcal{X})$ denote the collection of all real-valued continuous functions on $\mathcal{X}$, and let $C_b(\mathcal{X})$ denote the collection of all bounded real-valued continuous functions on $\mathcal{X}$. For a metrizable space $\mathcal{X}$ and its compatible metric ρ, let $BL(\mathcal{X}, \rho)$ denote the collection of all bounded Lipschitz functions on $\mathcal{X}$.

Topology and measures For a topological space $\mathcal{X}$, let $\mathscr{B}(\mathcal{X})$ denote the Borel σ-algebra on $\mathcal{X}$, and let $\mathscr{P}(\mathcal{X})$ denote the collection of all Borel probability measures on $\mathcal{X}$, i.e., probability measures defined on $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$.

Support of a measure Let $\mathcal{X}$ be a topological space. The support of $\mu \in \mathscr{P}(\mathcal{X})$ is defined by

$$\operatorname{supp}(\mu) = \{x \in \mathcal{X} : \mu(U) > 0 \quad \forall \text{open neighborhood } U \text{ of } x\}.$$

By definition, the support of $\mu \in \mathscr{P}(\mathcal{X})$ satisfies

$$\mathcal{X} \backslash \operatorname{supp}(\mu) = \bigcup_{\substack{G \text{ is open} \\ \mu(G)=0}} G \quad \Leftrightarrow \quad \operatorname{supp}(\mu) = \bigcap_{\substack{F \text{ is closed} \\ \mu(F)=1}} F.$$

Hence, $\operatorname{supp}(\mu)$ is a closed set. If $\mathcal{X}$ is second-countable, e.g., $\mathcal{X}$ is separable and metrizable, continuity of measures ensures $\mu(\operatorname{supp}(\mu)) = 1$, which implies that $\operatorname{supp}(\mu)$ is the smallest closed set having the total mass 1; equivalently, $\mathcal{X} \backslash \operatorname{supp}(\mu)$ is the largest open set having zero mass.

Euclidean spaces A Euclidean space is always equipped with the standard topology and its subset is equipped with the relative topology inherits from that. For $d \in \mathbb{N}$, let m_d denote the Lebesgue measure on the Borel σ-algebra on $\mathbb{R}^d$; the symbol $\mathrm{d}m_d(x)$ of integration is simply denoted as $\mathrm{d}x$. For $d = 1$, let λ denote the restriction of m_1 to the unit interval $[0, 1]$ so that $\lambda \in \mathscr{P}([0, 1])$.

Contents

1 Theoretical Foundations of Optimal Transport

Optimal transport theory concerns how to transport one probability measure to another with minimal cost. We introduce two notions of transport—transport maps and plans—and the corresponding notions of cost. These notions lead to two optimal transport problems, the Monge problem and the Kantorovich problem, which consist in finding transport maps and plans incurring the smallest cost, respectively. We show that the latter is a relaxed version of the former and highlight important aspects of their connections. Moreover, we provide a probabilistic interpretation of transport, which is useful for understanding intricate measure-theoretic formulations. Lastly, we introduce two essential tools in optimal transport theory called disintegration and gluing techniques, which will be used in the subsequent sections.

Settings Throughout this section, $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ denote probability spaces unless otherwise stated.

1.1 Transport maps and the Monge problem

We first study a notion of transport induced by a map. Recall that any measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ induces a probability measure $T_{\#}\mu$ on $(\mathcal{Y}, \mathcal{B})$, which we call the pushforward measure of μ by T. Roughly speaking, the pushforward measure is obtained by transporting mass consisting of μ via T. As a simple example, let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$, $\mathcal{A} = \mathcal{B} = \mathscr{B}(\mathbb{R}^2)$, and $\mu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$. Then, $T_{\#}\mu = \frac{1}{m}\sum_{i=1}^m \delta_{T(x_i)}$, which amounts to transporting the mass $\frac{1}{m}$ at each location x_i to another site $T(x_i)$ on the plane. In summary, a map gives rise to a notion of transport via pushforward measures. Using this, we define a transport map between two probability measures as follows.

Definition 1.1. A measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ is called a transport map from μ to ν if $T_{\#}\mu = \nu$. The collection of all transport maps from μ to ν is denoted as $\mathcal{T}(\mu, \nu)$.

Next, we define a notion of cost associated with transport maps. Recall from the aforementioned example with $\mu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ on the plane, that a map T transports the mass $\frac{1}{m}$ at location x_i to $T(x_i)$. We may price such transport by defining a unit cost, say the distance $\|x_i - T(x_i)\|_2$, and multiplying it by the mass $\frac{1}{m}$; then, the total cost is $\frac{1}{m}\sum_{i=1}^m \|x_i - T(x_i)\|_2$. To apply this idea to the abstract setting, we view the unit cost incurred by $x \mapsto T(x)$ as any quantity depending on the source location x and the destination $T(x)$ of the transport. For this reason, we consider a function c define on $\mathcal{X} \times \mathcal{Y}$ and define the unit cost associated with $x \mapsto T(x)$ as $c(x, T(x))$.

Definition 1.2. We call $c\colon \mathcal{X} \times \mathcal{Y} \to (-\infty, \infty]$ a cost function if it is bounded from below and measurable with respect to the product σ-algebra $\mathcal{A} \otimes \mathcal{B}$.[1]

Given a cost function c, by integrating the unit cost with respect to the source measure μ, we obtain the total transport cost incurred by a map T as follows:

$$\operatorname{cost}(T) = \int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x).$$

The optimal transport problem consists in finding a transport map that achieves the smallest transport cost, i.e., minimizing $\operatorname{cost}(T)$ over $T \in \mathcal{T}(\mu, \nu)$. This problem is attributed to Gaspard Monge who considered such a formulation to find the most economical way to transport the soil from the ground to several construction sites [Mon81].

Definition 1.3 (Monge Problem). Given a cost function c, suppose we associate each transport map $T \in \mathcal{T}(\mu, \nu)$ with the cost

$$\int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x).$$

The Monge problem seeks a transport map incurring the smallest cost; any element in

$$\operatorname*{arg\,min}_{T \in \mathcal{T}(\mu,\nu)} \int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x)$$

is called an optimal transport map. The optimal transport cost of the Monge problem is

$$\mathbb{M}_c(\mu, \nu) = \inf_{T \in \mathcal{T}(\mu,\nu)} \int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x) \in (-\infty, \infty].$$

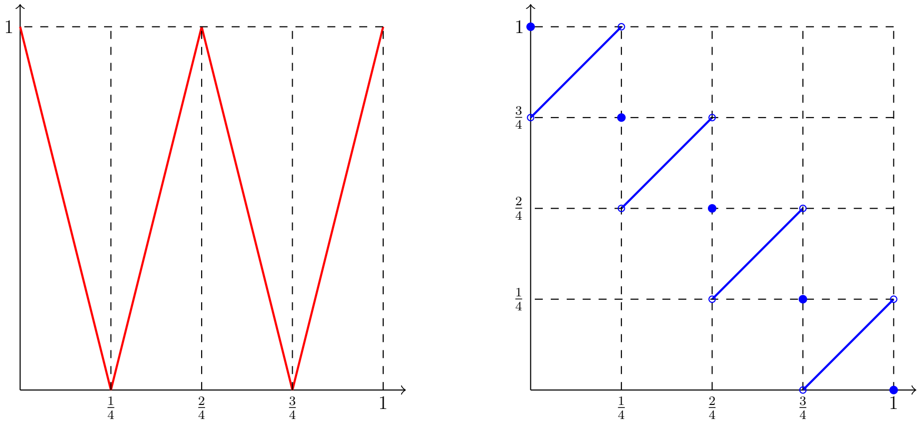
From a purely mathematical perspective, the Monge problem is simply a minimization of some function over $\mathcal{T}(\mu, \nu)$ and the so-called 'optimality' is used to refer to the minima. Depending on the situation, such optimality may be related to a desirable property.

Example 1.1. Suppose $\mathcal{X} = \mathcal{Y} = [0, 1]$ and $\mathcal{A} = \mathcal{B} = \mathscr{B}([0, 1])$. Let $\mu = \nu$ be the Lebesgue measure on $[0, 1]$. Clearly, the identity map Id from $[0, 1]$ to $[0, 1]$ is a transport map from μ to ν. In fact, there are many other transport maps that are highly nontrivial and complicated. For instance, define $T\colon [0, 1] \to [0, 1]$ as $T(x) = |2x - 1|$, let $T_1 = T$, and recursively define $T_{n+1} = T \circ T_n$ for $n \in \mathbb{N}$. Then, $T_n \in \mathcal{T}(\mu, \nu)$ for every $n \in \mathbb{N}$. Also, for each $n \in \mathbb{N}$, define $S_n\colon [0, 1] \to [0, 1]$ as follows:

$$S_n(x) = \begin{cases} \left(x - \frac{k-1}{2^n}\right) + \left(1 - \frac{k}{2^n}\right) & x \in \left(\frac{k-1}{2^n}, \frac{k}{2^n}\right) \text{ for } k = 1, \dots, 2^n, \\ 1 - x & x \in \left\{0, \frac{1}{2^n}, \dots, \frac{2^n - 1}{2^n}, 1\right\}. \end{cases}$$

[1] We require c to be bounded from below to prevent any integrability issues.

Then, one can verify that $(S_n)_{\#}\mu = \nu$ holds; hence, $S_n \in \mathcal{T}(\mu, \nu)$ for all $n \in \mathbb{N}$. Now, let $c(x, y) = |x - y|$. Then, the identity map Id incurs the zero transport cost, whereas T_n and S_n incur the positive transport cost. In other words, Id is an optimal transport map, while T_n and S_n are not. As such, solving the Monge problem leads to the most intuitive transport map Id between two probability measures.



(a) T_2

(b) S_2

Figure 1: T_2 and S_2 of Example 1.1.

Last but not least, we mention that the Monge problem may be infeasible, that is, $\mathcal{T}(\mu, \nu) = \emptyset$. In other words, there is no transport map between μ and ν. In this case, by convention, we often say the optimal transport cost of the Monge problem is ∞.

Example 1.2. Suppose $\mu = \delta_x$ and $\nu = \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$ for some $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$, assuming $\{x\} \in \mathcal{A}$ and $\{y_1\}, \{y_2\} \in \mathcal{B}$. For any measurable map $T \colon \mathcal{X} \to \mathcal{Y}$, the pushforward measure $T_{\#}\mu = \delta_{T(x)}$ is supported on a singleton $\{T(x)\}$, while the support of ν consists of two points. Hence, $T_{\#}\mu \neq \nu$ for any T.

1.2 Transport plans and the Kantorovich problem

As shown in Example 1.2, transport maps cannot split mass; if $\mu = \delta_x$, the total mass 1 at $x \in \mathcal{X}$ is transported to $T(x)$, resulting in the pushforward measure $T_{\#}\mu = \delta_{T(x)}$ that also has the total mass placed at one site $T(x)$. This is because a map T always maps each location $x \in \mathcal{X}$ to one location $T(x) \in \mathcal{Y}$, thereby prohibiting mass at x from splitting into multiple destinations.

We introduce a notion of transport that permits multiple destinations. The key object is a transport plan which records the amount of mass to be transported from x to y for any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. First, consider a simple case where $\mu = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j}$ for some $x_1, \dots, x_m \in \mathcal{X}$ and $y_1, \dots, y_n \in \mathcal{Y}$ with $\{x_1\}, \dots, \{x_n\} \in \mathcal{A}$ and $\{y_1\}, \dots, \{y_m\} \in \mathcal{B}$. Suppose we transport mass from μ to ν by specifying the amount of mass to transport from x_i to y_j, say $P_{ij} \geq 0$, for any pair $(i, j) \in [m] \times [n]$. Then, any admissible transport plan is represented by the following constraint on P_{ij}'s:

$$\sum_{j=1}^{n} P_{ij} = \frac{1}{m} \quad \forall i \in [m] \quad \text{and} \quad \sum_{i=1}^{m} P_{ij} = \frac{1}{n} \quad \forall j \in [n].$$

In other words, we can plan transport by simply determining a quantity (P_{ij}) of mass to transport from x_i to y_j for any pair $(i, j) \in [m] \times [n]$. Now, let us imagine $m, n \to \infty$ to generalize this concept to general μ and ν. As $m, n \to \infty$, the quantities (P_{ij}) become approximations of the amount of mass to transport from any $x \in \mathcal{X}$ to any $y \in \mathcal{Y}$. Accordingly, we can specify the amount of total mass to transport from a local area $\mathrm{d}x \in \mathcal{A}$ to another local area $\mathrm{d}y \in \mathcal{B}$ via a measure γ on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$, i.e., $\gamma(\mathrm{d}x \times \mathrm{d}y)$ denotes the total mass to transport from $\mathrm{d}x$ to $\mathrm{d}y$. In other words, now the transport plan is a measure recording the amount of mass to be transported from any local area $\mathrm{d}x \in \mathcal{A}$ to another local area $\mathrm{d}y \in \mathcal{B}$. As discussed earlier, any admissible transport plan should satisfy some constraint; in this case, $\gamma(\mathrm{d}x \times \mathcal{Y})$ is the total mass transported from a local area $\mathrm{d}x$, which must be $\mu(\mathrm{d}x)$, and similarly $\gamma(\mathcal{X} \times \mathrm{d}y) = \nu(\mathrm{d}y)$.

Definition 1.4. A probability measure γ on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ is called a transport plan from μ to ν if

$$\gamma(A \times \mathcal{Y}) = \mu(A) \quad \forall A \in \mathcal{A} \quad \text{and} \quad \gamma(\mathcal{X} \times B) = \nu(B) \quad \forall B \in \mathcal{B}.$$

The collection of all transport plans from μ to ν is denoted as $\Pi(\mu, \nu)$.

Definition 1.5 (**Kantorovich Problem**)**.** Given a cost function c, suppose we associate each transport plan $\gamma \in \Pi(\mu, \nu)$ with the cost

$$\int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\gamma.$$

The Kantorovich problem finds a transport plan incurring the smallest cost; any element in

$$\Pi_c(\mu, \nu) := \operatorname*{arg\,min}_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\gamma$$

is called an optimal transport plan from μ to ν. The optimal transport cost of the Kantorovich problem is

$$\mathbb{K}_c(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\gamma.$$

Again, $\mathbb{K}_c(\mu,\nu)=\infty$ might happen. In this case, it happens if and only if the cost of every transport plan is ∞. Also, notice that the Kantorovich problem is an infinite-dimensional linear program, which generalizes the following discrete case to abstract settings.

Example 1.3. Suppose that both probability spaces are discrete, that is, $\mathcal{X}=\{x_1,\dots,x_m\}$ and $\mathcal{Y}=\{y_1,\dots,y_n\}$, where $\mathcal{A}$ and $\mathcal{B}$ are their discrete σ-algebras. Then, we may represent μ and ν as

$$\mu=\sum_{i=1}^{m} a_i\delta_{x_i} \quad \text{and} \quad \nu=\sum_{j=1}^{n} b_j\delta_{y_j},$$

where $a_1,\dots,a_m,b_1,\dots,b_n\in\mathbb{R}_+$ and $\sum_{i=1}^m a_i=\sum_{j=1}^n b_j=1$. A cost function c can be represented as a matrix by specifying all possible values, that is, $C_{ij}:=c(x_i,y_j)\in\mathbb{R}$. Also, each transport plan $\gamma\in\Pi(\mu,\nu)$ takes the following form:

$$\gamma=\sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij}\delta_{(x_i,y_j)},$$

where $P_{ij}\in\mathbb{R}_+$ and

$$\sum_{j=1}^{n} P_{ij}=a_i \quad \text{and} \quad \sum_{i=1}^{m} P_{ij}=b_j.$$

In this case, γ incurs the cost $\sum_{i=1}^m\sum_{j=1}^n C_{ij}P_{ij}$. Accordingly, we may write the Kantorovich problem as follows:

$$\begin{aligned} &\text{minimize} \quad \langle \mathbf{C},\mathbf{P}\rangle \\ &\text{subject to} \quad \mathbf{P}\in\mathbb{R}_+^{m\times n}, \mathbf{P}1_n=\mathbf{a}, \mathbf{P}^\top 1_m=\mathbf{b}, \end{aligned} \tag{1.1}$$

where $\mathbf{a}=(a_1,\dots,a_m)^\top\in\mathbb{R}_+^m$, $\mathbf{b}=(b_1,\dots,b_n)^\top\in\mathbb{R}_+^n$, $\mathbf{C}=(C_{ij})\in\mathbb{R}^{m\times n}$, and $1_m,1_n$ are the vectors of all ones in $\mathbb{R}^m,\mathbb{R}^n$, respectively. In other words, the Kantorovich problem between discrete probability spaces is a linear program; the variable $\mathbf{P}$ is a matrix, the objective function is linear in $\mathbf{P}$, and the constraint set is a convex polytope, an intersection of a hyperspace $\mathbf{P}\in\mathbb{R}_+^{m\times n}$ and two hyperplanes $\mathbf{P}1_n=\mathbf{a}$ and $\mathbf{P}^\top 1_m=\mathbf{b}$.

1.3 Connections between the two optimal transport problems

We prove that the Kantorovich problem is a relaxed version of the Monge problem. The idea is that any transport map T induces a transport plan $(\mathrm{Id},T)_{\#}\mu$, and they incur the same transport cost given any cost function c.

Lemma 1.1. *For any $T\colon\mathcal{X}\to\mathcal{Y}$, let $(\mathrm{Id},T)\colon\mathcal{X}\to\mathcal{X}\times\mathcal{Y}$ denote a function that maps each $x\in\mathcal{X}$ to $(x,T(x))$. If $T\colon\mathcal{X}\to\mathcal{Y}$ is measurable, so is (Id,T).*

Proof. Fix a measurable map $T\colon \mathcal{X} \to \mathcal{Y}$. As $\mathcal{A}\otimes\mathcal{B}$ is generated by $\{A\times B : A \in \mathcal{A}, B \in \mathcal{B}\}$, it suffices to check $(\mathrm{Id}, T)^{-1}(A \times B) \in \mathcal{A}$ for any $A \in \mathcal{A}$ and $B \in \mathcal{B}$. This is true because measurability of T implies $(\mathrm{Id}, T)^{-1}(A \times B) = A \cap T^{-1}(B) \in \mathcal{A}$. □

Proposition 1.1. *$T \in \mathcal{T}(\mu, \nu)$ implies $(\mathrm{Id}, T)_{\#}\mu \in \Pi(\mu, \nu)$. Moreover, for any cost function c, the cost by a transport map T in the Monge problem coincides with the cost by a transport plan $(\mathrm{Id}, T)_{\#}\mu$ in the Kantorovich problem. Accordingly, $\mathbb{K}_c(\mu, \nu) \le \mathbb{M}_c(\mu, \nu)$.*

Proof. Fix $T \in \mathcal{T}(\mu, \nu)$. By Lemma 1.1, (Id, T) is measurable; hence, $\gamma_T = (\mathrm{Id}, T)_{\#}\mu$ is well-defined. Then, $\gamma_T(A \times \mathcal{Y}) = \mu(\mathrm{Id}^{-1}(A)) = \mu(A)$ for any $A \in \mathcal{A}$ and $\gamma_T(\mathcal{X} \times B) = \mu(T^{-1}(B)) = \nu(B)$ for any $B \in \mathcal{B}$. Hence, $\gamma_T \in \Pi(\mu, \nu)$. Also, for any cost function c,

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\,\mathrm{d}\gamma_T(x, y) = \int_{\mathcal{X}} (c \circ (\mathrm{Id}, T))(x)\,\mathrm{d}\mu(x) = \int_{\mathcal{X}} c(x, T(x))\,\mathrm{d}\mu(x).$$

Consequently, for any $T \in \mathcal{T}(\mu, \nu)$,

$$\int_{\mathcal{X}} c(x, T(x))\,\mathrm{d}\mu(x) \ge \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\,\mathrm{d}\gamma_T(x, y) \ge \inf_{\gamma\in\Pi(\mu,\nu)} c(x, y)\,\mathrm{d}\gamma(x, y) = \mathbb{K}_c(\mu, \nu).$$

Therefore, we have $\mathbb{K}_c(\mu, \nu) \le \mathbb{M}_c(\mu, \nu)$. □

Note that we can rewrite the Monge problem as

$$\inf_{\gamma\in\Pi_{\mathcal{T}}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\,\mathrm{d}\gamma,$$

where $\Pi_{\mathcal{T}}(\mu, \nu) = \{(\mathrm{Id}, T)_{\#}\mu : T \in \mathcal{T}(\mu, \nu)\}$ is a subset of $\Pi(\mu, \nu)$ by Proposition 1.1. Hence, both Monge and Kantorovich problems are minimization of the objective function

$$\gamma \mapsto \int_{\mathcal{X}\times\mathcal{Y}} c\,\mathrm{d}\gamma,$$

where the only difference is that the Kantorovich problem minimizes it over a constraint set $\Pi(\mu, \nu)$, which is larger than that of the Monge problem, that is, $\Pi_{\mathcal{T}}(\mu, \nu)$. In other words, we may view the Kantorovich problem as a relaxed version of the Monge problem obtained by relaxing the constraint $\Pi_{\mathcal{T}}(\mu, \nu)$ to $\Pi(\mu, \nu)$.

We will later see that optimal transport plans exists, i.e., given a cost function c, under mild assumptions, we can find $\gamma^\star \in \Pi(\mu, \nu)$ such that

$$\gamma^\star \in \operatorname*{arg\,min}_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\,\mathrm{d}\gamma \quad \Leftrightarrow \quad \mathbb{K}_c(\mu, \nu) = \int_{\mathcal{X}\times\mathcal{Y}} c\,\mathrm{d}\gamma^\star.$$

From the aforementioned connection between the Monge and the Kantorovich problems, we can deduce that $\mathbb{M}_c(\mu, \nu) = \mathbb{K}_c(\mu, \nu)$ if $\gamma^\star \in \Pi_{\mathcal{T}}(\mu, \nu)$, or equivalently, there is $T^\star \in \mathcal{T}(\mu, \nu)$ such that $\gamma^\star = (\mathrm{Id}, T^\star)_{\#}\mu$. In other words, both problems have the same optimal cost and $T^\star$ is an optimal transport map. The following proposition summarizes this simple observation.

Proposition 1.2. *If $\gamma^\star$ is an optimal transport plan such that $\gamma^\star = (\mathrm{Id}, T^\star)_\# \mu$ for some transport map $T^\star \in \mathcal{T}(\mu, \nu)$, we have $\mathbb{M}_c(\mu, \nu) = \mathbb{K}_c(\mu, \nu)$ and $T^\star$ is an optimal transport map.*

To utilize Proposition 1.2, we need to verify whether an optimal transport plan is induced by a transport map. As stated below, this depends on whether a transport plan is concentrated on the graph of some map.

Proposition 1.3. *Suppose* $\operatorname{graph}(T) \in \mathcal{A} \otimes \mathcal{B}$ *for any measurable $T \colon \mathcal{X} \to \mathcal{Y}$.*

(i) For any $T \in \mathcal{T}(\mu, \nu)$, the induced transport plan $\gamma_T := (\mathrm{Id}, T)_\# \mu$ is concentrated on the graph of T, i.e., $\gamma_T(\operatorname{graph}(T)) = 1$.

(ii) If $\gamma \in \Pi(\mu, \nu)$ is concentrated on $\operatorname{graph}(T)$ *for some measurable map $T \colon \mathcal{X} \to \mathcal{Y}$, then $\gamma = (\mathrm{Id}, T)_\# \mu$ and $T \in \mathcal{T}(\mu, \nu)$.*

Proof. (i) Note that $(\mathrm{Id}, T)^{-1}(\operatorname{graph}(T)) = \mathcal{X}$; hence, $\gamma_T(\operatorname{graph}(T)) = \mu(\mathcal{X}) = 1$.
(ii) For any $A \in \mathcal{A}$ and $B \in \mathcal{B}$, note that

$$(A \times B) \cap \operatorname{graph}(T) = \{(x, T(x)) : x \in A, T(x) \in B\} = \big((A \cap T^{-1}(B)) \times \mathcal{Y}\big) \cap \operatorname{graph}(T).$$

Hence,

$$\begin{aligned}
\gamma(A \times B) &= \gamma((A \times B) \cap \operatorname{graph}(T)) \quad (\because \gamma(\operatorname{graph}(T)) = 1) \\
&= \gamma\left(((A \cap T^{-1}(B)) \times \mathcal{Y}) \cap \operatorname{graph}(T)\right) \\
&= \gamma\left((A \cap T^{-1}(B)) \times \mathcal{Y}\right) \quad (\because \gamma(\operatorname{graph}(T)) = 1) \\
&= \mu(A \cap T^{-1}(B)) \quad (\because \gamma \in \Pi(\mu, \nu)) \\
&= (\mathrm{Id}, T)_\# \mu(A \times B).
\end{aligned}$$

Therefore, we conclude $\gamma = (\mathrm{Id}, T)_\# \mu$ by the π-λ theorem. The marginal of $\gamma = (\mathrm{Id}, T)_\# \mu$ on $\mathcal{Y}$ is $T_\# \mu$, which should be ν by definition. □

In summary, $\gamma \in \Pi(\mu, \nu)$ belongs to $\Pi_{\mathcal{T}}(\mu, \nu) = \{(\mathrm{Id}, T)_\# \mu : T \in \mathcal{T}(\mu, \nu)\}$ if and only if γ is concentrated on the graph of some measurable map. Lastly, we note that (ii) of Proposition 1.3 implies the following: for any measurable map $T \colon \mathcal{X} \to \mathcal{Y}$, there can be only one transport plan concentrated on $\operatorname{graph}(T)$, which is $(\mathrm{Id}, T)_\# \mu$. This will later play a crucial role in proving uniqueness of optimal transport plans, which is also related to uniqueness of optimal transport maps; see Proposition 1.8.

Remark 1.1 (**Measurability of Graphs**)**.** In Proposition 1.3, note that we have assumed measurability of $\operatorname{graph}(T)$. Without any conditions on $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, there might exists a measurable map $T \colon \mathcal{X} \to \mathcal{Y}$ such that $\operatorname{graph}(T) \notin \mathcal{A} \otimes \mathcal{B}$. If $\mathcal{X}$ and $\mathcal{Y}$ are separable metrizable spaces equipped with their Borel σ-algebras, $\operatorname{graph}(T)$ is always measurable if T is measurable; see Proposition 1.7.

1.4 Probabilistic interpretation of transport

As we have seen in the previous sections, optimal transport problems are formulated and handled by measure theory. However, purely measure-theoretic thinking is often difficult to follow. It turns out that we can alleviate such a difficulty by understanding optimal transport problems via probability theory. To this end, we introduce a probabilistic interpretation of optimal transport problems.

Transport plans are closely related to the coupling in probability theory. Coupling refers to constructing two random variables X and Y on some probability space such that the laws of X and Y are μ and ν, respectively; the joint law of (X, Y) is also referred to as a coupling of (μ, ν). Hence, couplings of (μ, ν) are exactly transport plans from μ to ν. Accordingly, we can rewrite the Kantorovich problem as follows:

$$\inf_{\substack{X\sim\mu\\ Y\sim\nu}} \mathbb{E}\, c(X, Y),$$

where $X \sim \mu$ and $Y \sim \nu$ stand for constructing $\mathcal{X}$-valued random variable X whose law is μ and $\mathcal{Y}$-valued random variable Y whose law is ν, respectively, and $\mathbb{E}$ denotes the expectation with respect to X and Y.

Proposition 1.4. *Define a map* $(\mathrm{Id}, \mathrm{Id})\colon \mathcal{X} \to \mathcal{X} \times \mathcal{X}$ *by mapping each* $x \in \mathcal{X}$ *to* $(x, x) \in \mathcal{X} \times \mathcal{X}$*. Then,* $(\mathrm{Id}, \mathrm{Id})$ *is measurable, and* $(\mathrm{Id}, \mathrm{Id})_{\#}\mu$ *belongs to* $\Pi(\mu, \mu)$*.*

Remark 1.2. A probabilistic interpretation of Proposition 1.4 is as follows: if there is a $\mathcal{X}$-valued random variable X whose law is μ, the law of a $(\mathcal{X} \times \mathcal{X})$-valued random variable (X, X) is simply $(\mathrm{Id}, \mathrm{Id})_{\#}\mu$. Clearly, the law of (X, X) is marginally μ and thus belongs to $\Pi(\mu, \mu)$.

Meanwhile, a coupling of (μ, ν) is said to be deterministic if there exists a measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ such that $T(X)$ and Y have the same law. Hence, such a map is exactly a transport map from μ to ν. Therefore, we can rewrite the Monge problem as follows:

$$\inf_{\substack{X\sim\mu\\ T(X)\sim\nu}} \mathbb{E}\, c(X, T(X)),$$

where $T(X) \sim \nu$ means that the law of $T(X)$ is ν.

We apply such an interpretation to the case where $(\mathcal{X}, \mathcal{A}) = (\mathcal{Y}, \mathcal{B}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, i.e., $\mu, \nu \in \mathscr{P}(\mathbb{R})$. It turns out that we can always find a transport map from $\lambda \in \mathscr{P}([0, 1])$—the Lebesgue measure restricted on $[0, 1]$—to any member of $\mathscr{P}(\mathbb{R})$. In the language of probability theory, this is represented as $F_\mu^{-1}(U) \sim \mu$, where U is the uniform random variable on $[0, 1]$ and F_μ^{-1} is the quantile function of $\mu \in \mathscr{P}(\mu)$. This is also known as the inverse transform sampling, implying that we can sample from μ by transforming samples from the uniform distribution on $[0, 1]$.

Lemma 1.2. *For any $\mu \in \mathscr{P}(\mathbb{R})$, let $F_\mu \colon \mathbb{R} \to [0,1]$ denote its distribution function, i.e., $F_\mu(x) = \mu(-\infty, x]$. Also, let $F_\mu^{-1} \colon (0,1) \to \mathbb{R}$ denotes its quantile function, i.e.,*

$$F_\mu^{-1}(u) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq u\}. \tag{1.2}$$

For simplicity, we always let $F_\mu^{-1}(0) = F_\mu^{-1}(1) = 0$ so that $F_\mu^{-1} \colon [0,1] \to \mathbb{R}$.[2]

(i) $F_\mu^{-1}(u) \leq x$ if and only if $u \leq F_\mu(x)$ for any $u \in (0,1)$ and $x \in \mathbb{R}$.

(ii) $(F_\mu^{-1})_\# \lambda = \mu$.

(iii) If F_μ is continuous, $(F_\mu)_\# \mu = \lambda$.

Proof. (i) Fix $u \in (0,1)$ and $x \in \mathbb{R}$. Then, $u \leq F_\mu(x)$ implies $F_\mu^{-1}(u) \leq x$ by definition. Conversely, assume $F_\mu^{-1}(u) \leq x$. By definition, we can find a sequence $(x_n)_{n \in \mathbb{N}}$ in $\mathbb{R}$ converging to $F_\mu^{-1}(u)$ such that $u \leq F_\mu(x_n)$ for all $n \in \mathbb{N}$. As F_μ is nondecreasing and right-continuous,

$$u \leq \lim_{n \to \infty} F(x_n) = F_\mu(F_\mu^{-1}(u)) \leq F_\mu(x).$$

(ii) It suffices to prove $(F_\mu^{-1})_\# \lambda(-\infty, x] = F_\mu(x)$ for all $x \in \mathbb{R}$. By definition,

$$(F_\mu^{-1})_\# \lambda(-\infty, x] = \lambda\{u \in [0,1] : F_\mu^{-1}(u) \leq x\} = \lambda\{u \in (0,1) : F_\mu^{-1}(u) \leq x\}.$$

Due to (i),

$$(F_\mu^{-1})_\# \lambda(-\infty, x] = \lambda\{u \in (0,1) : u \leq F_\mu(x)\} = \lambda(0, F_\mu(x)] = F_\mu(x).$$

(iii) It suffices to prove $(F_\mu)_\# \mu(0,u] = u$ for all $u \in (0,1)$. By definition,

$$(F_\mu)_\# \mu(0,u] = \mu\{x \in \mathbb{R} : F_\mu(x) \leq u\}.$$

As F_μ is continuous, $\{x \in \mathbb{R} : F_\mu(x) \leq u\} = (-\infty, x_u]$, where

$$x_u = \sup\{x \in \mathbb{R} : F_\mu(x) = u\}.$$

Continuity of F_μ implies $F_\mu(x_u) = u$; hence, $(F_\mu)_\# \mu(0,u] = \mu(-\infty, x_u] = F_\mu(x_u) = u$. In (ii) and (iii), measurability of F_μ and F_μ^{-1} is implied by their monotonicity. □

Proposition 1.5. *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\mathcal{A} = \mathcal{B} = \mathscr{B}(\mathbb{R})$; accordingly, $\mu, \nu \in \mathscr{P}(\mathbb{R})$ and $\Pi(\mu,\nu) \subset \mathscr{P}(\mathbb{R}^2)$.*[3] *For any $\gamma \in \mathscr{P}(\mathbb{R}^2)$, let $F_\gamma \colon \mathbb{R}^2 \to [0,1]$ denote its distribution function, i.e., $F_\gamma(x,y) = \gamma((-\infty, x] \times (-\infty, y])$.*

[2] As the values of F_μ^{-1} at 0 and 1 are inconsequential, we set them arbitrarily to extend its domain to $[0,1]$ without changing its range $\mathbb{R}$.

[3] By definition, transport plans are probability measures on $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R}))$; as $\mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R}) = \mathscr{B}(\mathbb{R}^2)$, transport plans are indeed Borel probability measures on $\mathbb{R}^2$.

(i) $\gamma \in \Pi(\mu,\nu)$ *if and only if*

$$F_\mu(x) + F_\nu(y) - 1 \le F_\gamma(x,y) \le \min\{F_\mu(x), F_\nu(y)\} \quad \forall (x,y) \in \mathbb{R}^2. \tag{1.3}$$

(ii) $\gamma = (F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ *if and only if* $F_\gamma(x,y) = \min\{F_\mu(x), F_\nu(y)\}$ *for all* $(x,y) \in \mathbb{R}^2$.

(iii) If F_μ *is continuous,* $F_\nu^{-1} \circ F_\mu$ *is a transport map from* μ *to* ν *and*

$$(F_\mu^{-1}, F_\nu^{-1})_\# \lambda = (\mathrm{Id}, F_\nu^{-1} \circ F_\mu)_\# \mu.$$

Proof. (i) Assume $\gamma \in \Pi(\mu,\nu)$. Since $(-\infty,x] \times (-\infty,y] = ((-\infty,x] \times \mathbb{R}) \cap (\mathbb{R} \times (-\infty,y])$,

$$\begin{aligned} F_\gamma(x,y) &= \gamma((-\infty,x] \times \mathbb{R}) + \gamma(\mathbb{R} \times (-\infty,y]) - \gamma(((-\infty,x] \times \mathbb{R}) \cup (\mathbb{R} \times (-\infty,y])) \\ &\ge \gamma((-\infty,x] \times \mathbb{R}) + \gamma(\mathbb{R} \times (-\infty,y]) - 1 \\ &= F_\mu(x) + F_\nu(y) - 1. \end{aligned}$$

Also, $F_\gamma(x,y) \le \min\{\gamma((-\infty,x] \times \mathbb{R}), \gamma(\mathbb{R} \times (-\infty,y])\} = \min\{F_\mu(x), F_\nu(y)\}$. Conversely, suppose γ satisfies (1.3). Letting $y \to \infty$ shows $\gamma((-\infty,x] \times \mathbb{R}) = \mu(-\infty,x]$ for all $x \in \mathbb{R}$; similarly, $\gamma(\mathbb{R} \times (-\infty,y]) = \nu(-\infty,y]$ for all $y \in \mathbb{R}$. Hence, $\gamma \in \Pi(\mu,\nu)$.

(ii) Since a Borel probability measure is completely determined by its distribution function, it suffices to prove the "only if" part. For $\gamma = (F_\mu^{-1}, F_\nu^{-1})_\# \lambda$, as in the proof of Lemma 1.2,

$$\begin{aligned} F_\gamma(x,y) &= (F_\mu^{-1}, F_\nu^{-1})_\# \lambda((-\infty,x] \times (-\infty,y]) \\ &= \lambda\{u \in [0,1] : F_\mu^{-1}(u) \le x \text{ and } F_\nu^{-1}(u) \le y\} \\ &= \lambda\{u \in (0,1) : F_\mu^{-1}(u) \le x \text{ and } F_\nu^{-1}(u) \le y\} \\ &= \lambda\{u \in (0,1) : u \le \min\{F_\mu(x), F_\nu(y)\}\} \\ &= \min\{F_\mu(x), F_\nu(y)\}. \end{aligned}$$

(iii) Due to Lemma 1.2, $(F_\nu^{-1} \circ F_\mu)_\# \mu = ((F_\nu^{-1})_\# (F_\mu)_\# \mu) = (F_\nu^{-1})_\# \lambda = \nu$. Similarly,

$$(F_\mu^{-1}, F_\nu^{-1})_\# \lambda = (F_\mu^{-1}, F_\nu^{-1})_\# ((F_\mu)_\# \mu) = (F_\mu^{-1} \circ F_\mu, F_\nu^{-1} \circ F_\mu)_\# \mu.$$

By Lemma 1.6, it suffices to show $\mu\{x \in \mathbb{R} : F_\mu^{-1} \circ F_\mu(x) \neq x\} = 0$. One can verify that

$$\{x \in \mathbb{R} : F_\mu^{-1} \circ F_\mu(x) \neq x\} \subset \{x \in \mathbb{R} : F_\mu(x) \in D\},$$

where D is the set of points of discontinuity of F_μ^{-1}, which is countable as F_μ^{-1} is monotone. Hence,

$$\mu\{x \in \mathbb{R} : F_\mu^{-1} \circ F_\mu(x) \neq x\} \le \sum_{d \in D} \mu\{x \in \mathbb{R} : F_\mu(x) = d\} = 0,$$

where $\mu\{x \in \mathbb{R} : F_\mu(x) = d\} = 0$ holds for any $d \in [0,1]$ as F_μ is continuous. □

1.5 Disintegration and gluing

We introduce two important techniques for analyzing transport plans: disintegration and gluing. Disintegration is a method for decomposing a probability measure on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ along with its marginal. More precisely, any probability measure γ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ can be decomposed along with the values in $\mathcal{X}$, namely, for each $x \in \mathcal{X}$, we can find a probability measure γ_x on $(\mathcal{Y}, \mathcal{B})$ such that $\gamma_x(B) = \gamma(\mathrm{d}x \times B)$, meaning that $\gamma_x(B)$ is roughly $\gamma(\{x\} \times B)$ for every $B \in \mathcal{B}$. In this case, γ can be represented as an integration of $x \mapsto \gamma_x$ with respect to the marginal of γ on $\mathcal{X}$. Such a result is known as disintegration theorem which holds if $\mathcal{Y}$ is a Polish space and $\mathcal{B} = \mathscr{B}(\mathcal{Y})$.

Theorem 1.1 (**Disintegration**). *Let $\mathcal{Y}$ be a Polish space and $\mathcal{B} = \mathscr{B}(\mathcal{Y})$. For any probability measure γ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ that is marginally μ on $\mathcal{X}$, there exists a μ-almost unique collection $\{\gamma_x : x \in \mathcal{X}\}$ of Borel probability measures on $\mathcal{Y}$ such that a function $x \mapsto \gamma_x(B)$ is measurable for any $B \in \mathcal{B}$ and*

$$\gamma(S) = \int_{\mathcal{X}} \int_{\mathcal{Y}} 1_{\{(x,y) \in S\}} \, \mathrm{d}\gamma_x(y) \mathrm{d}\mu(x) \quad \forall S \in \mathcal{A} \otimes \mathcal{B}.$$

More generally, for any measurable function $h \colon \mathcal{X} \times \mathcal{Y} \to [0, \infty]$,

$$\int_{\mathcal{X} \times \mathcal{Y}} h \, \mathrm{d}\gamma = \int_{\mathcal{X}} \int_{\mathcal{Y}} h(x, y) \, \mathrm{d}\gamma_x(y) \mathrm{d}\mu(x).$$

Remark 1.3. In Theorem 1.1, we call $\{\gamma_x : x \in \mathcal{X}\}$ a collection of conditional probability measures of γ with respect to its marginal μ on $\mathcal{X}$; also, μ-almost uniqueness means that if there are two such collections $\{\gamma_x : x \in \mathcal{X}\}$ and $\{\tilde{\gamma}_x : x \in \mathcal{X}\}$ of conditional probability measures of γ, there must exist $A \in \mathcal{A}$ such that $\mu(A) = 1$ and $\gamma_x = \tilde{\gamma}_x$, i.e., they are the same probability measure on $\mathcal{Y}$, for all $x \in A$.

Remark 1.4. In the setting of Theorem 1.1, suppose $T \colon \mathcal{X} \to \mathcal{Y}$ is a transport map from μ to ν and let γ be the transport plan induced by T (recall Proposition 1.1). Then, one can verify that $\gamma_x = \delta_{T(x)}$ since for any measurable function $h \colon \mathcal{X} \times \mathcal{Y} \to [0, \infty]$,

$$\int_{\mathcal{X} \times \mathcal{Y}} h \, \mathrm{d}\gamma = \int_{\mathcal{X}} h(x, T(x)) \, \mathrm{d}\mu(x) = \int_{\mathcal{X}} \int_{\mathcal{Y}} h(x, y) \, \mathrm{d}\delta_{T(x)}(y) \mathrm{d}\mu(x).$$

Disintegration theorem is a measure-theoretic version of the regular conditional distribution. Let X and Y be $\mathcal{X}$-valued and $\mathcal{Y}$-valued random variables, respectively. Letting γ be the joint law of (X, Y), Theorem 1.1 is essentially the same as obtaining the following conditional distribution: for each $x \in \mathcal{X}$, find a probability measure on $(\mathcal{Y}, \mathcal{B})$ given by

$$B \mapsto \mathbb{P}(Y \in B \,|\, X = x) =: \gamma_x(B) \quad \forall B \in \mathcal{B}$$

such that

$$\mathbb{E}\, h(X,Y) = \mathbb{E} \int_{\mathcal{Y}} h(X,y) \,\mathrm{d}\gamma_X(y),$$

where γ_X represents a $\mathscr{P}(\mathcal{Y})$-valued random variable, i.e., a random probability measure. In summary, a conditional probability measure γ_x of Theorem 1.1 is exactly the regular conditional distribution of Y given $X = x$. See Chapter IV of [Ç11] and Chapter 5 of [Kal97] for comprehensive treatment on the regular condition distribution.

Lemma 1.3 (Gluing). *Let $(\mathcal{X}_1, \mathcal{A}_1, \mu_1)$, $(\mathcal{X}_2, \mathcal{A}_2, \mu_2)$, and $(\mathcal{X}_3, \mathcal{A}_3, \mu_3)$ be probability spaces, where $\mathcal{X}_1$ and $\mathcal{X}_3$ are Polish spaces, $\mathcal{A}_1 = \mathscr{B}(\mathcal{X}_1)$, and $\mathcal{A}_3 = \mathscr{B}(\mathcal{X}_3)$. For $\gamma_{12} \in \Pi(\mu_1, \mu_2)$ and $\gamma_{23} \in \Pi(\mu_2, \mu_3)$, there exists a probability measure Γ on $(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3, \mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \mathcal{A}_3)$ such that $(P_{12})_\#\Gamma = \gamma_{12}$ and $(P_{23})_\#\Gamma = \gamma_{23}$, where $P_{ij}(x_1, x_2, x_3) = (x_i, x_j)$ for all $i, j \in \{1, 2, 3\}$ such that $i \neq j$ and all $(x_1, x_2, x_3) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$.*

Proof. By Theorem 1.1, we can find two collections $\{\gamma^{(1)}_{x_2} : x_2 \in \mathcal{X}_2\}$ and $\{\gamma^{(3)}_{x_2} : x_2 \in \mathcal{X}_2\}$ of Borel probability measures on $\mathcal{X}_1$ and $\mathcal{X}_3$, respectively, such that

$$\begin{aligned}
\gamma_{12}(S_{12}) &= \int_{\mathcal{X}_2} \int_{\mathcal{X}_1} 1_{\{(x_1,x_2)\in S_{12}\}} \,\mathrm{d}\gamma^{(1)}_{x_2}(x_1) \mathrm{d}\mu_2(x_2) \quad \forall S_{12} \in \mathcal{A}_1 \otimes \mathcal{A}_2,\\
\gamma_{23}(S_{23}) &= \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} 1_{\{(x_2,x_3)\in S_{23}\}} \,\mathrm{d}\gamma^{(3)}_{x_2}(x_3) \mathrm{d}\mu_2(x_2) \quad \forall S_{23} \in \mathcal{A}_2 \otimes \mathcal{A}_3.
\end{aligned}$$

Define $\Gamma \in \mathscr{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$ by

$$\Gamma(S) = \int_{\mathcal{X}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_3} 1_{\{(x_1,x_2,x_3)\in S\}} \,\mathrm{d}\gamma^{(1)}_{x_2}(x_1) \mathrm{d}\gamma^{(3)}_{x_2}(x_3) \mathrm{d}\mu_2(x_2) \quad \forall S \in \mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \mathcal{A}_3.$$

Then, $(P_{12})_\#\Gamma = \gamma_{12}$ and $(P_{23})_\#\Gamma = \gamma_{23}$; note that this also implies $(P_{13})_\#\Gamma \in \Pi(\mu_1, \mu_3)$. □

1.6 Supplementary results

We have defined transport maps and plans using set-theoretic definitions. These can always be rewritten in terms of integration.

Lemma 1.4. *Let $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ be two measure spaces. For a measurable map $T \colon \mathcal{X} \to \mathcal{Y}$, the following are equivalent.*

(i) $T_\#\mu = \nu$.

(ii) For any measurable function $\psi \colon \mathcal{Y} \to [0, \infty]$,

$$\int_{\mathcal{Y}} \psi \,\mathrm{d}\nu = \int_{\mathcal{X}} \psi \circ T \,\mathrm{d}\mu.$$

(iii) For any measurable function $\psi \colon \mathcal{Y} \to [-\infty, \infty]$ such that $\int_{\mathcal{Y}} \psi \,\mathrm{d}\nu$ is well-defined,

$$\int_{\mathcal{Y}} \psi \,\mathrm{d}\nu = \int_{\mathcal{X}} \psi \circ T \,\mathrm{d}\mu.$$

We can rewrite the marginal constraints in terms of integration.

Lemma 1.5. *The following are equivalent.*

(i) $\gamma \in \Pi(\mu, \nu)$.

(ii) For any measurable functions $\varphi \colon \mathcal{X} \to [0, \infty]$ and $\psi \colon \mathcal{Y} \to [0, \infty]$,

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) \,\mathrm{d}\gamma(x, y) = \int_{\mathcal{X}} \varphi(x) \,\mathrm{d}\mu(x) \quad \text{and} \quad \int_{\mathcal{X}\times\mathcal{Y}} \psi(y) \,\mathrm{d}\gamma(x, y) = \int_{\mathcal{Y}} \psi(y) \,\mathrm{d}\nu(y).$$

(iii) For any measurable functions $\varphi \colon \mathcal{X} \to [-\infty, \infty]$ and $\psi \colon \mathcal{Y} \to [-\infty, \infty]$ such that $\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu$ and $\int_{\mathcal{Y}} \psi \,\mathrm{d}\nu$ are well-defined,

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) \,\mathrm{d}\gamma(x, y) = \int_{\mathcal{X}} \varphi(x) \,\mathrm{d}\mu(x) \quad \text{and} \quad \int_{\mathcal{X}\times\mathcal{Y}} \psi(y) \,\mathrm{d}\gamma(x, y) = \int_{\mathcal{Y}} \psi(y) \,\mathrm{d}\nu(y).$$

Proposition 1.6. *Let $\mu = \delta_{x_0}$ and $\nu = \delta_{y_0}$ for some $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$. Suppose $\{x_0\} \in \mathcal{A}$ and $\{y_0\} \in \mathcal{B}$. Then, $\Pi(\mu, \nu) = \{\delta_{(x_0, y_0)}\}$.*

Proposition 1.7. *Suppose $\{(y, y) : y \in \mathcal{Y}\} \in \mathcal{B} \otimes \mathcal{B}$. Then, $\operatorname{graph}(T) \in \mathcal{A} \otimes \mathcal{B}$ for any measurable $T \colon \mathcal{X} \to \mathcal{Y}$. Also, $\{x \in \mathcal{X} : T_1(x) = T_2(x)\} \in \mathcal{A}$ for any measurable maps $T_1 \colon \mathcal{X} \to \mathcal{Y}$ and $T_2 \colon \mathcal{X} \to \mathcal{Y}$.*

Proof. Fix a measurable map $T \colon \mathcal{X} \to \mathcal{Y}$ and let $f_T \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} \times \mathcal{Y}$ be a map such that $f_T(x, y) = (T(x), y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We verify measurability of f_T. As $\mathcal{B} \otimes \mathcal{B}$ is generated by $\{B_1 \times B_2 : B_1, B_2 \in \mathcal{B}\}$, it suffices to check $f_T^{-1}(B_1 \times B_2) \in \mathcal{A} \otimes \mathcal{B}$. This is true because measurability of T implies $f_T^{-1}(B_1 \times B_2) = T^{-1}(B_1) \times B_2 \in \mathcal{A} \otimes \mathcal{B}$. Therefore, letting $\Delta = \{(y, y) : y \in \mathcal{Y}\} \in \mathcal{B} \otimes \mathcal{B}$, we have $\operatorname{graph}(T) = f_T^{-1}(\Delta) \in \mathcal{A} \otimes \mathcal{B}$. Lastly, for measurable maps T_1 and T_2, notice that

$$\{x \in \mathcal{X} : T_1(x) = T_2(x)\} = (\mathrm{Id}, T_1)^{-1}(\operatorname{graph}(T_1) \cap \operatorname{graph}(T_2)) \in \mathcal{A}.$$

□

Remark 1.5. The assumption $\{(y, y) : y \in \mathcal{Y}\} \in \mathcal{B} \otimes \mathcal{B}$ of Proposition 1.7 is satisfied if $\mathcal{Y}$ is a separable metrizable space and $\mathcal{B} = \mathscr{B}(\mathcal{Y})$, i.e., its Borel σ-algebra. To see this, verify that $\{(y, y) : y \in \mathcal{Y}\}$ is a closed set in $\mathcal{Y} \times \mathcal{Y}$ using metrizability, which guarantees $\{(y, y) : y \in \mathcal{Y}\} \in \mathscr{B}(\mathcal{Y} \times \mathcal{Y})$. Lastly, $\mathscr{B}(\mathcal{Y} \times \mathcal{Y}) = \mathscr{B}(\mathcal{Y}) \otimes \mathscr{B}(\mathcal{Y})$ due to separability.

Lemma 1.6. *Suppose that two measurable maps $T_1 \colon \mathcal{X} \to \mathcal{Y}$ and $T_2 \colon \mathcal{X} \to \mathcal{Y}$ coincide μ-almost everywhere, that is, $\mu\{x \in \mathcal{X} : T_1(x) \neq T_2(x)\} = 0$. Then, $(T_1)_{\#}\mu = (T_2)_{\#}\mu$.*

Proof. Let $A = \{x \in \mathcal{X} : T_1(x) \neq T_2(x)\}$ (measurable by assumption). For $B \in \mathcal{B}$,

$$\begin{aligned}\mu\{x \in \mathcal{X} : T_2(x) \in B\} &= \mu\{x \in A : T_2(x) \in B\} + \mu\{x \in \mathcal{X}\backslash A : T_2(x) \in B\}\\ &= \mu\{x \in \mathcal{X}\backslash A : T_1(x) \in B\}\\ &\leq \mu\{x \in \mathcal{X} : T_1(x) \in B\},\end{aligned}$$

hence $(T_2)_{\#}\mu(B) \leq (T_1)_{\#}\mu(B)$. By symmetry, $(T_2)_{\#}\mu(B) \geq (T_1)_{\#}\mu(B)$. □

The converse of Lemma 1.6 is not true in general, i.e., $(T_1)_{\#}\mu = (T_2)_{\#}\mu$ does not imply $T_1 = T_2$. Instead, the following holds.

Lemma 1.7. *Suppose* $\mathrm{graph}(T) \in \mathcal{A}\otimes\mathcal{B}$ *for any measurable $T \colon \mathcal{X} \to \mathcal{Y}$. For two measurable maps $T_1 \colon \mathcal{X} \to \mathcal{Y}$ and $T_2 \colon \mathcal{X} \to \mathcal{Y}$, suppose* $(\mathrm{Id}, T_1)_{\#}\mu = (\mathrm{Id}, T_2)_{\#}\mu$. *Then, $T_1 \colon \mathcal{X} \to \mathcal{Y}$ and $T_2 \colon \mathcal{X} \to \mathcal{Y}$ coincide μ-almost everywhere.*

Proof. Let $\gamma = (\mathrm{Id}, T_1)_{\#}\mu = (\mathrm{Id}, T_2)_{\#}\mu$, $G_1 = \mathrm{graph}(T_1)$, and $G_2 = \mathrm{graph}(T_2)$. Then, $\gamma(G_1) = \gamma(G_2) = 1$, which is true because $\mathcal{X} = (\mathrm{Id}, T_1)^{-1}(G_1) = (\mathrm{Id}, T_2)^{-1}(G_2)$; this does not need $\gamma \in \Pi(\mu, \nu)$. Accordingly, $\gamma(G_1 \cap G_2) = 1$. Also, as $G_1 \cap G_2 \in \mathcal{A} \otimes \mathcal{B}$,

$$(\mathrm{Id}, T_1)^{-1}(G_1 \cap G_2) = \{x \in \mathcal{X} : T_1(x) = T_2(x)\} \in \mathcal{A}.$$

Therefore,

$$\mu\{x \in \mathcal{X} : T_1(x) = T_2(x)\} = \mu((\mathrm{Id}, T_1)^{-1}(G_1 \cap G_2)) = \gamma(G_1 \cap G_2) = 1.$$

□

In Proposition 1.2, we have seen that existence of optimal transport maps is implied by an optimal transport plan induced by some transport map. The next proposition establishes uniqueness of optimal transport maps. If there is a unique optimal transport plan and it is induced by a transport map, then such a transport map is also μ-almost everywhere unique.

Proposition 1.8. *Suppose* $\mathrm{graph}(T) \in \mathcal{A}\otimes\mathcal{B}$ *for any measurable $T \colon \mathcal{X} \to \mathcal{Y}$. Suppose $\gamma^\star$ is the unique optimal transport plan and is induced by some $T^\star \in \mathcal{T}(\mu, \nu)$, i.e., $\gamma =$* $(\mathrm{Id}, T^\star)_{\#}\mu$. *Then, $T^\star$ is a μ-almost everywhere unique optimal transport map.*

Proof. Since $\gamma = (\mathrm{Id}, T^\star)_{\#}\mu$ is the unique optimal transport plan, one can verify from Proposition 1.1 that $\mathbb{K}_c(\mu, \nu) = \mathbb{M}_c(\mu, \nu)$, which shows that $T^\star$ is an optimal transport map. Suppose $T \in \mathcal{T}(\mu, \nu)$ is another optimal transport map. Then, $(\mathrm{Id}, T)_{\#}\mu$ must be an optimal transport plan, meaning that $(\mathrm{Id}, T^\star)_{\#}\mu = (\mathrm{Id}, T)_{\#}\mu$. By Lemma 1.7, we conclude that T and $T^\star$ coincide μ-almost everywhere. □

2 Existence of Optimal Transport Plans

We will prove that the Kantorovich problem admits a minimizer under mild conditions. The key idea is (semi-)continuity of the objective function $\gamma \mapsto \int c \, \mathrm{d}\gamma$ and compactness of $\Pi(\mu, \nu)$, which necessitates a suitable topology on $\Pi(\mu, \nu)$. To this end, we utilize the weak topology on $\mathscr{P}(\mathcal{X} \times \mathcal{Y})$, assuming $\mathcal{X}$ and $\mathcal{Y}$ are separable metrizable spaces.

Settings Throughout this section, $\mathcal{X}$ and $\mathcal{Y}$ are separable metrizable spaces. We always formulate optimal transport problems between $(\mathcal{X}, \mathscr{B}(\mathcal{X}), \mu)$ and $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}), \nu)$. Accordingly, we consider a cost function $c \colon \mathcal{X} \times \mathcal{Y} \to (-\infty, \infty]$, where $\mathcal{X} \times \mathcal{Y}$ is equipped with the product σ-algebra $\mathscr{B}(\mathcal{X}) \otimes \mathscr{B}(\mathcal{Y})$. Since $\mathcal{X}$ and $\mathcal{Y}$ are separable, so is $\mathcal{X} \times \mathcal{Y}$, which implies $\mathscr{B}(\mathcal{X}) \otimes \mathscr{B}(\mathcal{Y}) = \mathscr{B}(\mathcal{X} \times \mathcal{Y})$. Hence, measurability of c is with respect to the Borel σ-algebra $\mathscr{B}(\mathcal{X} \times \mathcal{Y})$. For the same reason, $\Pi(\mu, \nu) \subset \mathscr{P}(\mathcal{X} \times \mathcal{Y})$, that is, transport plans are Borel probability measures on $\mathcal{X} \times \mathcal{Y}$.

Remark 2.1 (**On Metrizability**)**.** Though $\mathcal{X}$ and $\mathcal{Y}$ are metrizable spaces, we will usually avoid specifying metrics that metrize their topologies. The reason is that most of the upcoming results originate from topological properties, not metric-dependent properties.

Remark 2.2 (**On Separability**)**.** Separability of $\mathcal{X}$ and $\mathcal{Y}$ is not only a mild assumption, but also an inevitable setting. The most natural/essential assumption on the cost function is (semi)-continuity. Continuous functions, however, might not be measurable with respect to the product σ-algebra $\mathscr{B}(\mathcal{X}) \otimes \mathscr{B}(\mathcal{Y})$ because it can be strictly smaller than the Borel σ-algebra $\mathscr{B}(\mathcal{X} \times \mathcal{Y})$, which turns out to be the smallest σ-algebra on $\mathcal{X} \times \mathcal{Y}$ that makes all continuous functions measurable as $\mathcal{X} \times \mathcal{Y}$ is metrizable; see Lemma 4.65 of [AB06]. Hence, ensuring $\mathscr{B}(\mathcal{X}) \otimes \mathscr{B}(\mathcal{Y}) = \mathscr{B}(\mathcal{X} \times \mathcal{Y})$ via separability is indispensable for considering (semi)-continuous cost functions. Also, as separability ensures $\Pi(\mu, \nu) \subset \mathscr{P}(\mathcal{X} \times \mathcal{Y})$, we can study $\Pi(\mu, \nu)$ by means of the weak topology $\mathscr{P}(\mathcal{X} \times \mathcal{Y})$. Without separability, we cannot say that transport plans are Borel probability measures as they are defined over $\mathscr{B}(\mathcal{X}) \otimes \mathscr{B}(\mathcal{Y})$ which can be strictly smaller than $\mathscr{B}(\mathcal{X} \times \mathcal{Y})$.

2.1 Weak topology

The main results of this section come from various properties of the weak topology on $\mathscr{P}(\mathcal{X} \times \mathcal{Y})$. We briefly study the weak topology on $\mathscr{P}(\mathcal{Z})$ for a general metrizable space $\mathcal{Z}$; see Chapter 15 of [AB06], Chapter 11 of [Dud02], Section 1 of [Bil99].

In general, given a set and a family of functions defined on the set, the weak topology generated by that family is the smallest (weakest) topology that makes all functions in that family continuous; see Section 2.13 of [AB06]. In the case of the space of Borel probability measures, the weak topology means the one generated by a family of functionals associated with bounded and continuous functions.

Definition 2.1. Let $\mathcal{Z}$ be a metrizable space. For each $f \in C_b(\mathcal{Z})$, define a functional L_f on $\mathscr{P}(\mathcal{Z})$ as follows:

$$L_f(\gamma) = \int_{\mathcal{Z}} f \,\mathrm{d}\gamma.$$

The weak topology on $\mathscr{P}(\mathcal{Z})$ generated by $\{L_f : f \in C_b(\mathcal{Z})\}$ is simply called the weak topology on $\mathscr{P}(\mathcal{Z})$. The convergence in this weak topology is called the weak convergence; a sequence $(\gamma_n)_{n\in\mathbb{N}}$ in $\mathscr{P}(\mathcal{Z})$ is said to converge weakly to $\gamma \in \mathscr{P}(\mathcal{Z})$ if

$$\lim_{n\to\infty} \int_{\mathcal{Z}} f \,\mathrm{d}\gamma_n = \int_{\mathcal{Z}} f \,\mathrm{d}\gamma \quad \forall f \in C_b(\mathcal{Z}). \tag{2.1}$$

Of course, there is an obvious reason for choosing these functionals to define the weak topology. As stated in the following lemma, $\gamma \mapsto \{L_f(\gamma) : f \in C_b(\mathcal{Z})\}$ is injective. In other words, $\{L_f : f \in C_b(\mathcal{Z})\}$ distinguishes elements of $\mathscr{P}(\mathcal{Z})$.

Lemma 2.1. *Given a metrizable space $\mathcal{Z}$, two elements γ_1 and γ_2 of $\mathscr{P}(\mathcal{Z})$ coincide if and only if*

$$\int_{\mathcal{Z}} f \,\mathrm{d}\gamma_1 = \int_{\mathcal{Z}} f \,\mathrm{d}\gamma_2 \quad \forall f \in C_b(\mathcal{Z}).$$

By definition of weak convergence, we can see that the objective function of the Kantorovich problem is continuous provided the cost function is continuous and bounded. In other words, if $c \in C_b(\mathcal{X} \times \mathcal{Y})$, for a sequence $(\gamma_n)_{n\in\mathbb{N}}$ in $\Pi(\mu, \nu)$ converging weakly to $\gamma \in \Pi(\mu, \nu)$,

$$\lim_{n\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c \,\mathrm{d}\gamma_n = \int_{\mathcal{X}\times\mathcal{Y}} c \,\mathrm{d}\gamma.$$

However, the assumption $c \in C_b(\mathcal{X} \times \mathcal{Y})$ is often too restrictive. Using the following lemma, we will show that lower semi-continuity of c leads to

$$\liminf_{n\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c \,\mathrm{d}\gamma_n \geq \int_{\mathcal{X}\times\mathcal{Y}} c \,\mathrm{d}\gamma,$$

which turns out to be sufficient for existence of optimal transport plans.

Lemma 2.2. *For a nonnegative lower semi-continuous function f defined on a metric space $(\mathcal{Z}, \rho)$, there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of bounded Lipschitz functions on $\mathcal{Z}$ converging pointwise to f such that $0 \leq f_n \leq f_{n+1} \leq f$ for all $n \in \mathbb{N}$.*

Proof. For each $n \in \mathbb{N}$, define $f_n(z) = \inf_{y\in\mathcal{Z}} \left(f(y) \wedge n + n\rho(z, y)\right)$ for all $z \in \mathcal{Z}$. Observe that f_n is bounded and n-Lipschitz and $0 \leq f_n \leq f_{n+1} \leq f$ for all $n \in \mathbb{N}$. Fix $z \in \mathcal{Z}$

and prove $\lim_{n\to\infty} f_n(z) = f(z)$. Fix $a \in \mathbb{R}$ such that $a < f(z)$. Then, due to lower semi-continuity of f, there exists $\delta > 0$ such that $a < f(y)$ for all $y \in \mathcal{Z}$ satisfying $\rho(y, z) < \delta$. For $n \in \mathbb{N}$ such that $n > a$,

$$\inf_{\substack{y\in\mathcal{Z}\\ \rho(y,z)<\delta}} (f(y) \wedge n + n\rho(z, y)) \geq \inf_{\substack{y\in\mathcal{Z}\\ \rho(y,z)<\delta}} f(y) \wedge n \geq a.$$

For $y \in \mathcal{Z}$ satisfying $\rho(y, z) \geq \delta$, provided $n > a/\delta$,

$$f(y) \wedge n + n\rho(z, y) \geq 0 + n\delta > a.$$

In summary, $f_n(z) \geq a$ for $n \in \mathbb{N}$ such that $n > a \vee a/\delta$, which means $\lim_{n\to\infty} f_n(z) \geq a$. As this is true for any $a \in \mathbb{R}$ satisfying $a < f(z)$, we conclude that f_n converges to f pointwise. □

In other words, any lower semi-continuous that is bounded from below is represented as a pointwise limit of a nondecreasing sequence of bounded Lipschitz functions.

Proposition 2.1. *Given a metrizable space $\mathcal{Z}$, suppose that a sequence $(\gamma_n)_{n\in\mathbb{N}}$ in $\mathscr{P}(\mathcal{Z})$ converges weakly to $\gamma \in \mathscr{P}(\mathcal{Z})$. Then, for any lower semi-continuous f that is bounded from below,*

$$\int_{\mathcal{Z}} f \, \mathrm{d}\gamma \leq \liminf_{n\to\infty} \int_{\mathcal{Z}} f \, \mathrm{d}\gamma_n.$$

Proof. Without loss of generality, assume $f \geq 0$. By Lemma 2.2, we can find a sequence $(f_k)_{k\in\mathbb{N}}$ of bounded Lipschitz functions—under some compatible metric on $\mathcal{Z}$—converging pointwise to f. As $0 \leq f_k \leq f_{k+1} \leq f$ for all $k \in \mathbb{N}$, the monotone convergence theorem implies

$$\int_{\mathcal{Z}} f \, \mathrm{d}\gamma = \sup_{k\in\mathbb{N}} \int_{\mathcal{Z}} f_k \, \mathrm{d}\gamma = \sup_{k\in\mathbb{N}} \left(\lim_{n\to\infty} \int_{\mathcal{Z}} f_k \, \mathrm{d}\gamma_n \right) \leq \liminf_{n\to\infty} \left(\sup_{k\in\mathbb{N}} \int_{\mathcal{Z}} f_k \, \mathrm{d}\gamma_n \right).$$

Applying the monotone convergence theorem again,

$$\int_{\mathcal{Z}} f \, \mathrm{d}\gamma \leq \liminf_{n\to\infty} \left(\sup_{k\in\mathbb{N}} \int_{\mathcal{Z}} f_k \, \mathrm{d}\gamma_n \right) = \liminf_{n\to\infty} \int_{\mathcal{Z}} f \, \mathrm{d}\gamma_n.$$

□

Remark 2.3. If the weak topology on $\mathscr{P}(\mathcal{Z})$ is metrizable, Proposition 2.1 essentially shows that the functional $\gamma \mapsto \int_{\mathcal{Z}} f \, \mathrm{d}\gamma$ is lower semi-continuous. The weak topology on $\mathscr{P}(\mathcal{Z})$ is indeed metrizable (and separable) provided $\mathcal{Z}$ is separable; see Theorem 15.12 of [AB06].

The essence of Lemma 2.2 and Proposition 2.1 is the approximation technique based on the collection of all bounded Lipschitz functions. This collection plays an important role in weak convergence as well, namely, we may replace $C_b(\mathcal{Z})$ with this collection in (2.1). In fact, there are many alternative definitions of weak convergence, which is summarized as follows; see Theorem 11.1.1 of [Dud02] for the proof.

Theorem 2.1 (Portmanteau Theorem). *Let $\mathcal{Z}$ be a metrizable space, $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $\mathscr{P}(\mathcal{Z})$, and $\gamma \in \mathscr{P}(\mathcal{Z})$. The following are equivalent.*

(i) $(\gamma_n)_{n\in\mathbb{N}}$ *converges weakly to* γ.

(ii) Given any compatible metric ρ *on* $\mathcal{Z}$,

$$\lim_{n\to\infty} \int_{\mathcal{Z}} f \,\mathrm{d}\gamma_n = \int_{\mathcal{Z}} f \,\mathrm{d}\gamma \quad \forall f \in BL(\mathcal{Z}, \rho).$$

(ii) $\liminf_{n\to\infty} \gamma_n(G) \geq \gamma(G)$ *for every open set* $G \subset \mathcal{Z}$.

(iii) $\limsup_{n\to\infty} \gamma_n(F) \leq \gamma(F)$ *for every closed set* $F \subset \mathcal{Z}$.

(iv) $\lim_{n\to\infty} \gamma_n(B) = \gamma_n(B)$ *for every set* $B \subset \mathcal{Z}$ *such that* $\gamma(\partial B) = 0$.

Now, we shift our interest to compactness in the weak topology. It turns out that the following regularity called tightness plays a crucial role.

Definition 2.2. Let $\mathcal{Z}$ be a metrizable space. We say $\gamma \in \mathscr{P}(\mathcal{Z})$ is tight if for any $\varepsilon > 0$, there exists a compact set K such that $\gamma(\mathcal{Z}\backslash K) < \varepsilon$. We say a collection $\mathcal{P} \subset \mathscr{P}(\mathcal{Z})$ is tight if for any $\varepsilon > 0$, there exists a compact set K such that $\gamma(\mathcal{Z}\backslash K) < \varepsilon$ for all $\gamma \in \mathcal{P}$.

Remark 2.4. If $\mathcal{Z}$ is a Polish space, any element of $\mathscr{P}(\mathcal{Z})$ is tight, which is referred to as Ulam's theorem; see Theorem 7.1.4 of [Dud02].

We introduce Prokhorov's theorem, the most fundamental result establishing compactness in the weak topology. Essentially, it relates tightness of $\mathcal{P} \subset \mathscr{P}(\mathcal{Z})$ with relative compactness; see Theorems 5.1 and 5.2 of [Bil99] or Lemma 15.21 and Theorem 15.22 of [AB06] for the proof.

Theorem 2.2 (Prokhorov's Theorem). *Let $\mathcal{Z}$ be a metrizable space and $\mathcal{P} \subset \mathscr{P}(\mathcal{Z})$.*

(i) $\mathcal{P}$ *is tight.*

(ii) Any sequence in $\mathcal{P}$ *has a weakly convergent subsequence (its limit may not be in* $\mathcal{P}$*).*

(iii) $\mathcal{P}$ *is relatively compact.*

Then, (i) implies (ii). If $\mathcal{Z}$ is separable, (ii) and (iii) are equivalent. If $\mathcal{Z}$ is a Polish space, (ii) = (iii) implies (i).

Remark 2.5. In Theorem 2.2, (ii) is often referred to as relative sequential compactness of $\mathcal{P}$. In general, relative compactness and relative sequential compactness are equivalent in a metrizable space. Hence, (ii) and (iii) are equivalent provided the weak topology on $\mathscr{P}(\mathcal{Z})$ is metrizable, which is true if $\mathcal{Z}$ is separable (Remark 2.3).

2.2 Existence of optimal transport plans

First, we give an alternative characterization of a transport plan that is useful in metrizable space cases (compare with Lemma 1.5).

Lemma 2.3. $\gamma \in \Pi(\mu,\nu)$ *if and only if*

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{X}} \varphi(x)\,\mathrm{d}\mu(x) \quad \text{and} \quad \int_{\mathcal{X}\times\mathcal{Y}} \psi(y)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{Y}} \psi(y)\,\mathrm{d}\nu(y).$$

for all $(\varphi,\psi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y})$.

Proof. Define a map $P_{\mathcal{X}} \colon \mathcal{X}\times\mathcal{Y} \to \mathcal{X}$ as $P_{\mathcal{X}}(x,y) = x$. Note that, $P_{\mathcal{X}}$ is measurable with respect to $(\mathcal{X}\times\mathcal{Y}, \mathscr{B}(\mathcal{X})\otimes\mathscr{B}(\mathcal{Y}))$ and $(\mathcal{X},\mathscr{B}(\mathcal{X}))$. It suffices to show that $(P_{\mathcal{X}})_{\#}\gamma = \mu$ if and only if

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{X}} \varphi(x)\,\mathrm{d}\mu(x)$$

for all $\varphi \in C_b(\mathcal{X})$. Also, $(P_{\mathcal{X}})_{\#}\gamma = \mu$ if and only if

$$\int_{\mathcal{X}} \varphi\,\mathrm{d}(P_{\mathcal{X}})_{\#}\gamma = \int_{\mathcal{X}} \varphi\,\mathrm{d}\mu,$$

for all $\varphi \in C_b(\mathcal{X})$ by Lemma 2.1. Since the change of variables formula yields

$$\int_{\mathcal{X}} \varphi\,\mathrm{d}(P_{\mathcal{X}})_{\#}\gamma = \int_{\mathcal{X}\times\mathcal{Y}} \varphi\,\mathrm{d}\gamma$$

for all $\varphi \in C_b(\mathcal{X})$, we prove $(P_{\mathcal{X}})_{\#}\gamma = \mu$ if and only if

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi\,\mathrm{d}\gamma = \int_{\mathcal{X}} \varphi\,\mathrm{d}\mu.$$

□

In this case, $\Pi(\mu,\nu)$ is closed in the weak topology on $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$.

Proposition 2.2. $\Pi(\mu,\nu)$ *is closed in* $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$.

Proof. Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $\Pi(\mu,\nu)$ converging weakly to some $\gamma \in \mathscr{P}(\mathcal{X}\times\mathcal{Y})$. Then for each $\varphi \in C_b(\mathcal{X})$ since $(x,y) \mapsto \varphi(x)$ is in $C_b(\mathcal{X}\times\mathcal{Y})$,

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x)\,\mathrm{d}\gamma(x,y) = \lim_{n\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} \varphi(x)\,\mathrm{d}\gamma_n(x,y) = \int_{\mathcal{X}} \varphi(x)\,\mathrm{d}\mu(x).$$

Repeating the same process with $\psi \in C_b(\mathcal{Y})$ yields $\gamma \in \Pi(\mu,\nu)$ due to the previous part of the lemma. □

Now, we present the main theorem. As mentioned earlier, the key is to prove $\Pi(\mu,\nu)$ is compact in the weak topology on $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$.

Theorem 2.3 (**Existence of Optimal Transport Plans**). *If $\mu\in\mathscr{P}(\mathcal{X})$ and $\nu\in\mathscr{P}(\mathcal{Y})$ are tight, $\Pi(\mu,\nu)$ is compact in $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$. Moreover, if the cost function c is lower semi-continuous, the Kantorovich problem admits a minimizer, i.e., there exists an optimal transport plan.*

Proof. Since $\Pi(\mu,\nu)$ is closed, it suffices to show that it is relatively compact to prove compactness; due to the separability of $\mathcal{X}\times\mathcal{Y}$, tightness implies relative compactness in the weak topology of $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$ (see Theorem 2.2). Since μ and ν are tight, for $\varepsilon>0$, there exist compact sets K and L such that $\mu(\mathcal{X}\backslash K)<\varepsilon$ and $\nu(\mathcal{Y}\backslash L)\le\varepsilon$. Then, for any $\gamma\in\Pi(\mu,\nu)$,

$$\gamma\left((\mathcal{X}\times\mathcal{Y})\backslash(K\times L)\right)\le\gamma(\mathcal{X}\times(\mathcal{Y}\backslash L))+\gamma((\mathcal{X}\backslash K)\times\mathcal{Y})=\nu(\mathcal{Y}\backslash L)+\mu(\mathcal{X}\backslash K)\le 2\varepsilon.$$

Hence, $\Pi(\mu,\nu)$ is tight. As mentioned earlier, this leads to compactness. Now, choose a sequence $(\gamma_n)_{n\in\mathbb{N}}$ in $\Pi(\mu,\nu)$ such that $\int c\,\mathrm{d}\gamma_n\to\mathbb{K}_c(\mu,\nu)$. Since $\Pi(\mu,\nu)$ is compact, by taking a subsequence if necessary, we may assume that $(\gamma_n)_{n\in\mathbb{N}}$ converges weakly to some $\gamma^\star\in\Pi(\mu,\nu)$. Since c is lower semi-continuous and bounded from below, Proposition 2.1 implies

$$\int_{\mathcal{X}\times\mathcal{Y}}c\,\mathrm{d}\gamma^\star\le\liminf_{n\to\infty}\int_{\mathcal{X}\times\mathcal{Y}}c\,\mathrm{d}\gamma_n=\mathbb{K}_c(\mu,\nu).$$

Hence, $\gamma^\star$ is an optimal transport plan. □

Remark 2.6. Notice that the compactness of $\Pi(\mu,\nu)$ comes from the tightness of μ and ν. If $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces, as mentioned in Remark 2.4, all the elements of $\mathscr{P}(\mathcal{X})$ and $\mathscr{P}(\mathcal{Y})$ are tight; in this case, we can always find an optimal transport plan between μ and ν provided c is lower semi-continuous.

2.3 Supplementary results

Proposition 2.3. *$\mu\in\mathscr{P}(\mathcal{X})$ and $\nu\in\mathscr{P}(\mathcal{Y})$ are tight if and only if $\Pi(\mu,\nu)$ is tight.*

Proof. We have already shown $(\Rightarrow)$. To prove $(\Leftarrow)$, assume $\Pi(\mu,\nu)$ is tight. Fix $\varepsilon>0$. Then, we can find a compact set $K\subset\mathcal{X}\times\mathcal{Y}$ such that $\gamma((\mathcal{X}\times\mathcal{Y})\backslash K)<\varepsilon$ for all $\gamma\in\Pi(\mu,\nu)$. Now, we define $K_{\mathcal{X}}\subset\mathcal{X}$ such that $x\in K_{\mathcal{X}}$ if and only if there exists $y\in\mathcal{Y}$ satisfying $(x,y)\in K$. Note that for any $\gamma\in\Pi(\mu,\nu)$,

$$\mu(\mathcal{X}\backslash K_{\mathcal{X}})=\gamma((\mathcal{X}\backslash K_{\mathcal{X}})\times\mathcal{Y})=\gamma((\mathcal{X}\times\mathcal{Y})\backslash(K_{\mathcal{X}}\times\mathcal{Y}))\le\gamma((\mathcal{X}\times\mathcal{Y})\backslash K)<\varepsilon,$$

where the first inequality is due to $K\subset K_{\mathcal{X}}\times\mathcal{Y}$. Now, it suffices to prove that $K_{\mathcal{X}}$ is a compact subset of $\mathcal{X}$. If $(U_i)_{i\in I}$ is an open cover of $K_{\mathcal{X}}$, then $(U_i\times\mathcal{Y})_{i\in I}$ is an open cover of

K. Compactness of K implies that there exists a finite subcover, say $(U_i \times \mathcal{Y})_{1\le i\le N}$. Verify that $(U_i)_{1\le i\le N}$ covers $K_{\mathcal{X}}$. Hence, μ is tight. Similarly, ν is tight. □

Lemma 2.4 (**Tightness of Transport Plans**)**.** *Let $\mathcal{P}$ and $\mathcal{Q}$ be tight subsets of $\mathscr{P}(\mathcal{X})$ and $\mathscr{P}(\mathcal{Y})$, respectively, and $\Pi(\mathcal{P},\mathcal{Q})$ be the collection of all Borel probability measures on $\mathcal{X}\times\mathcal{Y}$ whose marginals lie in $\mathcal{P}$ and $\mathcal{Q}$, respectively, i.e., $(P_{\mathcal{X}})_{\#}\gamma\in\mathcal{P}$ and $(P_{\mathcal{Y}})_{\#}\gamma\in\mathcal{Q}$. Then, $\Pi(\mathcal{P},\mathcal{Q})$ is a tight subset of $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$.*

Proof. For $\varepsilon>0$, there exist compact sets K and L such that $\mu(\mathcal{X}\backslash K)<\varepsilon$ for all $\mu\in\mathscr{P}(\mathcal{X})$ and $\nu(\mathcal{Y}\backslash L)\le\varepsilon$ for all $\nu\in\mathscr{P}(\mathcal{Y})$. Hence, for any $\gamma\in\Pi(\mathcal{P},\mathcal{Q})$,

$$\begin{aligned}\gamma\left((\mathcal{X}\times\mathcal{Y})\backslash(K\times L)\right) &\le \gamma(\mathcal{X}\times(\mathcal{Y}\backslash L))+\gamma((\mathcal{X}\backslash K)\times\mathcal{Y})\\ &= (P_{\mathcal{Y}})_{\#}\gamma(\mathcal{Y}\backslash L)+(P_{\mathcal{X}})_{\#}\gamma(\mathcal{X}\backslash K)\\ &\le 2\varepsilon.\end{aligned}$$

Therefore, $\Pi(\mathcal{P},\mathcal{Q})$ is tight. □

Proposition 2.4 (**Stability of Transport Plans**)**.** *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces. Let $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ be sequences converging weakly to μ in $\mathscr{P}(\mathcal{X})$ and ν in $\mathscr{P}(\mathcal{Y})$, respectively. Also, let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $\mathscr{P}(\mathcal{X}\times\mathcal{Y})$ such that $\gamma_n\in\Pi(\mu_n,\nu_n)$ for all $n\in\mathbb{N}$. Then, $(\gamma_n)_{n\in\mathbb{N}}$ has a subsequence converging weakly to some $\gamma\in\Pi(\mu,\nu)$.*

Proof. Since $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces, $\mathcal{P}=\{\mu_n:n\in\mathbb{N}\}$ and $\mathcal{Q}=\{\nu_n:n\in\mathbb{N}\}$ are sequentially compact, which is equivalent to compactness as the weak topologies on $\mathscr{P}(\mathcal{X})$ and $\mathscr{P}(\mathcal{Y})$ are metrizable (Remark 2.3). Accordingly, $\mathcal{P}$ and $\mathcal{Q}$ are tight by Theorem 2.2. Due to Lemma 2.4, $\Pi(\mathcal{P},\mathcal{Q})$ is tight; hence, so is $\{\gamma_n:n\in\mathbb{N}\}\subset\Pi(\mathcal{P},\mathcal{Q})$. Theorem 2.2 tells that $(\gamma_n)_{n\in\mathbb{N}}$ has a subsequence, say $(\gamma_{n(k)})_{k\in\mathbb{N}}$, converging weakly to some $\gamma\in\mathscr{P}(\mathcal{X}\times\mathcal{Y})$. Note that

$$\int_{\mathcal{X}\times\mathcal{Y}}\varphi(x)\,\mathrm{d}\gamma(x,y)=\lim_{k\to\infty}\int_{\mathcal{X}\times\mathcal{Y}}\varphi(x)\,\mathrm{d}\gamma_{n(k)}(x,y)=\lim_{k\to\infty}\int_{\mathcal{X}}\varphi(x)\,\mathrm{d}\mu_{n(k)}(x)=\int_{\mathcal{X}}\varphi(x)\,\mathrm{d}\mu(x)$$

for all $\varphi\in C_b(\mathcal{X})$. Hence, $\gamma\in\Pi(\mu,\nu)$ follows due to Lemma 2.3. □

3 Optimality and Duality in the Kantorovich Problem

This section studies a necessary and sufficient condition for optimal transport plans, which will serve as the most fundamental result in optimal transport theory. In addition, we show that the Kantorovich problem is associated with its dual problem, generalizing the duality theory of finite-dimensional linear programming to the infinite-dimensional case.

Settings Throughout this section, $\mathcal{X}$ and $\mathcal{Y}$ are separable metrizable spaces unless otherwise stated; we consider the Kantorovich problem between $(\mathcal{X}, \mathscr{B}(\mathcal{X}), \mu)$ and $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}), \nu)$ with a cost function c.

3.1 Overview

We first briefly discuss high-level ideas of optimality and duality.

Optimality It turns out that the support of a transport plan determines optimality; such a property is called the c-cyclical monotonicity. More precisely, if $\gamma \in \Pi(\mu, \nu)$ is an optimal transport plan,

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

for any $n \in \mathbb{N}$, $(x_1, y_1), \ldots, (x_n, y_n) \in \operatorname{supp}(\gamma)$, and any permutation $\sigma \in \operatorname{Perm}(n)$. To see why optimality is related to such a condition, let us consider a simple example, where $\mu = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}$ and $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$. Suppose that $\gamma = \frac{1}{2}\delta_{(x_1,y_1)} + \frac{1}{2}\delta_{(x_2,y_2)}$ is an optimal transport plan. Then, $c(x_1, y_1) + c(x_2, y_2) \le c(x_1, y_2) + c(x_2, y_1)$ must hold; otherwise, $\gamma^\star := \frac{1}{2}\delta_{(x_1,y_2)} + \frac{1}{2}\delta_{(x_2,y_1)}$ incurs the smaller cost than γ, contradicting γ is optimal.

Duality Besides the optimality result, we will derive the duality result of the Kantorovich problem. The dual problem takes the following form:

$$\begin{aligned} \text{maximize} \quad & \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \, \mathrm{d}\nu, \\ \text{subject to} \quad & (\varphi, \psi) \in L^1(\mu) \times L^1(\nu), \\ & \varphi(x) + \psi(y) \le c(x, y) \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \end{aligned}$$

By Lemma 1.5, note that the dual objective function is

$$D(\varphi, \psi) := \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \, \mathrm{d}\nu = \int_{\mathcal{X} \times \mathcal{Y}} (\varphi(x) + \psi(y)) \, \mathrm{d}\gamma(x, y) \quad \forall \gamma \in \Pi(\mu, \nu).$$

Due to the constraint of the dual problem, $D(\varphi, \psi) \le \mathbb{K}_c(\mu, \nu)$ for any dual variable (φ, ψ), which implies that the supremum of the dual problem $\le \mathbb{K}_c(\mu, \nu)$. Under mild conditions, we will see that this inequality becomes equality, showing the duality.

Semi-Duality The objective function D of the dual problem increases if we replace the dual variable (φ, ψ) with (φ, φ^c), where $\varphi^c(y) = \inf_{x \in \mathcal{X}} (c(x, y) - \varphi(x))$, because $\psi \leq \varphi^c$. In other words, φ^c—called the c-transform of φ—is the largest possible function ψ such that $\varphi(x) + \psi(y) \leq c(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Accordingly, maximizing the following, called the semi-dual problem, is equivalent to the dual problem:

$$\varphi \mapsto S(\varphi) := \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c \, \mathrm{d}\nu$$

The semi-duality, i.e., the supremum of S over a suitable collection coincides with $\mathbb{K}_c(\mu, \nu)$, holds under mild assumptions. Also, the semi-dual problem admits a maximizer, say φ_o, satisfying a property called c-concavity. Importantly, for any optimal transport plan γ,

$$\int_{\mathcal{X} \times \mathcal{Y}} (\varphi_o(x) + \varphi_o^c(y)) \, \mathrm{d}\gamma(x, y) = S(\varphi_o) = \mathbb{K}_c(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\gamma,$$

which implies that γ is concentrated on the following set called the c-superdifferential of φ:

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

This set will play an important role in characterizing optimal transport plans.

Definition 3.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets and $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

(i) A subset Π of $\mathcal{X} \times \mathcal{Y}$ is said to be c-cyclically monotone if

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

for any $n \in \mathbb{N}$, $(x_1, y_1), \ldots, (x_n, y_n) \in \Pi$, and any permutation $\sigma \in \mathrm{Perm}(n)$.

(ii) The c-transform of $\varphi\colon \mathcal{X} \to [-\infty, \infty]$ is a function $\varphi^c\colon \mathcal{Y} \to [-\infty, \infty]$ defined by

$$\varphi^c(y) = \inf_{x \in \mathcal{X}} \left(c(x, y) - \varphi(x)\right).$$

Similarly, the c-transform of $\psi\colon \mathcal{Y} \to [-\infty, \infty]$ is $\psi^c\colon \mathcal{X} \to [-\infty, \infty]$ defined by

$$\psi^c(x) = \inf_{y \in \mathcal{Y}} \left(c(x, y) - \psi(y)\right).$$

(iii) A function $\varphi\colon \mathcal{X} \to [-\infty, \infty]$ is c-concave if $\varphi = \psi^c$ for some $\psi\colon \mathcal{Y} \to [-\infty, \infty]$. Similarly, $\psi\colon \mathcal{Y} \to [-\infty, \infty]$ is c-concave if $\psi = \varphi^c$ for some $\varphi\colon \mathcal{X} \to [-\infty, \infty]$.

(iv) The c-superdifferential of a function $\varphi\colon \mathcal{X} \to [-\infty, \infty]$ is defined by

$$\partial_c \varphi := \left\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \varphi(x) + \varphi^c(y) = c(x, y)\right\}.$$

In fact, all the aforementioned optimality, duality, and semi-duality results are closely intertwined due to connections among c-cyclical monotonicity, c-transform, c-concavity, and c-superdifferential. The main result is that every c-cyclically monotone set is a subset of the c-superdifferential of some proper c-concave function.

Theorem 3.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets and $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. If $\Pi \subset \mathcal{X} \times \mathcal{Y}$ is c-cyclically monotone, there exists a proper c-concave function $\varphi\colon \mathcal{X} \to [-\infty, \infty)$ such that*

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

Remark 3.1. Let us consider the case where $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x,y) = \frac{1}{2}\|x-y\|_2^2$. Then, all the ingredients in Definition 3.1 as well as Theorem 3.1 boil down to well-known convex analysis results; we see this in Section 5.2. Indeed, Theorem 3.1 is a generalization of Rockafellar's result on cyclical monotonicity (Theorem 5.2), which plays a role in characterizing the subdifferential of a convex function. We defer the proof of Theorem 3.1 to Section 3.4.

Also, c-concavity plays a crucial role in the semi-dual problem. Recall that the dual variable (φ, ψ) is associated with the dual objective function

$$D(\varphi, \psi) = \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \, \mathrm{d}\nu.$$

We have seen that $D(\varphi, \psi) \leq D(\varphi, \varphi^c)$, i.e., we can increase the dual objective function by replacing the (φ, ψ) with (φ, φ^c), which results in the semi-dual problem. In fact, we can apply this reasoning again to deduce that $D(\varphi, \varphi^c) \leq D(\varphi^{cc}, \varphi^c) \leq D(\varphi^{cc}, \varphi^{ccc})$. Can we increase D endlessly by repeating this process? It turns out that $\varphi^c = \varphi^{ccc}$, meaning that this process causes no increase after $(\varphi^{cc}, \varphi^c)$. The crucial fact $\varphi^c = \varphi^{ccc}$ is indeed a key result in characterizing c-concavity. We will show that any c-concave function remains unchanged after taking the c-transform twice.

All the discussions so far will be restated in the subsequent sections, rigorously checking technical details. To this end, we prepare basic properties the c-transform and c-concavity.

Proposition 3.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets and $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Fix $\varphi\colon \mathcal{X} \to [-\infty, \infty]$.*

(i) $\varphi^c \equiv -\infty$ *if* $\varphi(x) = \infty$ *for some* $x \in \mathcal{X}$.

(ii) $\varphi^c \equiv \infty$ *if* $\varphi \equiv -\infty$.

Now, suppose $\varphi\colon \mathcal{X} \to [-\infty, \infty)$ and φ is proper.

(iii) $\varphi^c\colon \mathcal{Y} \to [-\infty, \infty)$.

(v) $\varphi(x) + \varphi^c(y) \leq c(x,y)$ *for all* $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

Remark 3.2. Note that φ^c in (iii) of Proposition 3.1 might not be proper, i.e., $\varphi^c \equiv -\infty$ can happen. For instance, consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $c(x, y) \equiv 0$, and $\varphi(x) = x^2$.

Next, we characterize c-concave functions. As mentioned earlier, the key is that any c-concave function remains unchanged after taking the c-transform twice.

Proposition 3.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets and $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Fix $\varphi\colon \mathcal{X} \to [-\infty, \infty]$.*

(i) $\varphi^{cc} \geq \varphi$.

(ii) $\varphi = \varphi^{cc}$ *if and only if* φ *is c-concave; in this case, only one of the following is true:*

(1) $\varphi \equiv \infty$ *and* $\varphi^c \equiv -\infty$.

(2) $\varphi \equiv -\infty$ *and* $\varphi^c \equiv \infty$.

(3) $\varphi\colon \mathcal{X} \to [-\infty, \infty)$ *and* $\varphi^c\colon \mathcal{Y} \to [-\infty, \infty)$*, where both* φ *and* φ^c *are proper.*

Remark 3.3. As in Remark 3.1, if $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x, y) = \frac{1}{2}\|x - y\|_2^2$, Proposition 3.2 boils down to the standard result on conjugate of convex functions.

Lastly, we derive topological properties of c-concave functions.

Proposition 3.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets and $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.*

(i) If c is bounded, every proper c-concave function is bounded.

Let $\mathcal{X}$ and $\mathcal{Y}$ be topological spaces and suppose c is continuous.

(ii) Every c-concave function is upper semi-continuous.

(iii) $\partial_c\varphi$ *is closed for every c-concave function.*

Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces.

(iv) If c is uniformly continuous, every proper c-concave function is uniformly continuous.

(v) If c is L-Lipschitz for some $L > 0$*, every proper c-concave function is L-Lipschitz.*

We defer the proofs of Propositions 3.2 and 3.3 to Section 3.4.

3.2 Optimality

Now, we are ready to establish the optimality result. We first start with the following necessary condition for optimal transport plans, stating that every optimal transport plan has a c-cyclically monotone support.

Proposition 3.4. *Suppose $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and $\mathbb{K}_c(\mu, \nu) < \infty$. Then, the support of any optimal transport plan is c-cyclically monotone.*

The formal proof of Proposition 3.4 is defer to Section 3.4. Instead, consider a simple case where both μ and ν are finitely supported. Suppose γ is an optimal transport plan whose support is not c-cyclically monotone. By definition, we can find $(x_1, y_1), \ldots, (x_n, y_n) \in \operatorname{supp}(\gamma)$ and a permutation $\sigma \in \operatorname{Perm}(n)$ such that

$$C := \sum_{i=1}^{n} c(x_i, y_i) - \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) > 0.$$

Note that γ is also finitely supported; hence, $(x_i, y_i) \in \operatorname{supp}(\gamma)$ implies $w_i := \gamma(\{(x_i, y_i)\}) > 0$ for all $i \in [n]$. Take $\eta > 0$ such that $n\eta < \min_{i\in[n]} w_i$ and let

$$\gamma^\star = \gamma - \eta \sum_{i=1}^{n} \delta_{(x_i,y_i)} + \eta \sum_{i=1}^{n} \delta_{(x_i,\sigma(i))}.$$

By definition,

$$\gamma^\star(S) \geq \gamma(S) - \eta \sum_{i=1}^{n} \delta_{(x_i,y_i)}(S) \geq \sum_{i=1}^{n} (\gamma(S) - w_i \delta_{(x_i,y_i)}(S)) \quad \forall S \in \mathscr{B}(\mathcal{X} \times \mathcal{Y}).$$

Hence, $\gamma^\star \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$. Also, $\gamma^\star \in \Pi(\mu, \nu)$ follows as

$$\gamma^\star(A \times \mathcal{Y}) = \gamma(A \times \mathcal{Y}) - \eta \sum_{i=1}^{n} \delta_{x_i}(A) + \eta \sum_{i=1}^{n} \delta_{x_i}(A) = \mu(A) \quad \forall A \in \mathscr{B}(\mathcal{X}),$$
$$\gamma^\star(\mathcal{X} \times B) = \gamma(\mathcal{X} \times B) - \eta \sum_{i=1}^{n} \delta_{y_i}(B) + \eta \sum_{i=1}^{n} \delta_{y_{\sigma(i)}}(B) = \nu(B) \quad \forall B \in \mathscr{B}(\mathcal{Y}).$$

Therefore,

$$\int c \, \mathrm{d}\gamma^\star = \int c \, \mathrm{d}\gamma - \eta \sum_{i=1}^{n} c(x_i, y_i) + \eta \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) < \int c \, \mathrm{d}\gamma,$$

where the last inequality holds because $C > 0$.[4] This contradicts that γ is optimal. Therefore, $\operatorname{supp}(\gamma)$ must be c-cyclically monotone.

[4] Note that the strict inequality holds as $\int c \, \mathrm{d}\gamma < \infty$ which follows from $\mathbb{K}_c(\mu, \nu) < \infty$.

In summary, if the support is not c-cyclically monotone, we can modify a transport plan by reallocating the mass at such points $(x_1, y_1), \dots, (x_n, y_n)$ to $(x_1, y_{\sigma(1)}), \dots, (x_n, y_{\sigma(n)})$. In the case where μ and ν are not necessarily finitely supported, we can still apply this idea by taking a small open rectangular $U_i \times V_i$ containing (x_i, y_i) for each $i \in [n]$ and reallocating the mass on $U_1 \times V_1, \dots, U_n \times V_n$ to $U_1 \times V_{\sigma(1)}, \dots, U_n \times V_{\sigma(n)}$.

We show that the converse is true under mild assumptions, thereby proving that c-cyclically monotone support is the necessary and sufficient condition for optimality.

Theorem 3.2 (Optimality). *Suppose $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and $\mathbb{K}_c(\mu, \nu) < \infty$. Assume*

$$\int_{\mathcal{X}} c(x, y)\, \mathrm{d}\mu(x) < \infty \quad \forall y \in \mathcal{Y} \quad \textit{and} \quad \int_{\mathcal{Y}} c(x, y)\, \mathrm{d}\nu(y) < \infty \quad \forall x \in \mathcal{X}. \tag{MC}$$

For $\gamma \in \Pi(\mu, \nu)$, the following are equivalent.

(i) γ is an optimal transport plan.

(ii) $\mathrm{supp}(\gamma)$ is c-cyclically monotone.

(iii) There exists a proper c-concave function $\varphi_o\colon \mathcal{X} \to [-\infty, \infty)$ such that $\mathrm{supp}(\gamma) \subset \partial_c \varphi_o$.

Proof. We have already proved (i) ⇒ (ii) by Proposition 3.4. Also, (ii) ⇒ (iii) holds by Theorem 3.1. Suppose (iii) holds. We show that $\varphi_o^+ \in L^1(\mu)$ and $(\varphi_o^c)^+ \in L^1(\nu)$. First, φ_o and φ_o^c are measurable as they are upper semi-continuous by Proposition 3.3. As φ_o is proper, $\varphi_o(x) + \varphi_o^c(y) \le c(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ by Proposition 3.1. Hence, for any $y_0 \in \mathcal{Y}$,

$$\varphi_o(x) \le c(x, y_0) - \varphi_o^c(y_0) \quad \forall x \in \mathcal{X}. \tag{3.1}$$

As $\partial_c \varphi_o$ contains $\mathrm{supp}(\gamma)$ which is nonempty, φ_o^c is proper. Hence, we may pick $y_0 \in \mathcal{Y}$ such that $\varphi_o^c(y_0) \in \mathbb{R}$. Then, (MC) and (3.1) imply $\varphi_o^+ \in L^1(\mu)$; similarly, $(\varphi_o^c)^+ \in L^1(\nu)$. Therefore, for any $\gamma' \in \Pi(\mu, \nu)$,

$$\int_{\mathcal{X}} \varphi_o \, \mathrm{d}\mu + \int_{\mathcal{Y}} \varphi_o^c \, \mathrm{d}\nu = \int_{\mathcal{X} \times \mathcal{Y}} (\varphi_o(x) + \varphi_o^c(y))\, \mathrm{d}\gamma'(x, y) \le \int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\gamma', \tag{3.2}$$

where the equality is due to Lemma 1.5 and the inequality holds as $\varphi_o(x) + \varphi_o^c(y) \le c(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ by Proposition 3.1. Meanwhile, note that

$$\int_{\mathcal{X}} \varphi_o \, \mathrm{d}\mu + \int_{\mathcal{Y}} \varphi_o^c \, \mathrm{d}\nu = \int_{\mathcal{X} \times \mathcal{Y}} (\varphi_o(x) + \varphi_o^c(y))\, \mathrm{d}\gamma(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\gamma, \tag{3.3}$$

where the second equality is due to $\mathrm{supp}(\gamma) \subset \partial_c \varphi_o$. Therefore, by comparing (3.2) and (3.3), we conclude that γ is an optimal transport plan. □

Remark 3.4 (**On Moment Condition**)**.** As we have seen in the proof of Theorem 3.2, the Moment Condition (MC) guarantees the following: for any proper c-concave function $\varphi\colon \mathcal{X} \to [-\infty, \infty)$, we have $\varphi^+ \in L^1(\mu)$ and $(\varphi^c)^+ \in L^1(\nu)$. This enabled us to obtain (3.3) for φ_o in (iii), which—together with the assumption $\mathbb{K}_c(\mu, \nu) < \infty$—leads to $\varphi_o \in L^1(\mu)$ and $\varphi_o^c \in L^1(\nu)$. In fact, (MC) is a mild assumption which is satisfied in many situations. First, it is obvious when c is bounded; in this case, $\mathbb{K}_c(\mu, \nu) < \infty$ is also guaranteed. As we will see later, another common situation is where we can find $a \in L^1(\mu)$ and $b \in L^1(\nu)$ such that

$$c(x, y) \le a(x) + b(y) \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \tag{3.4}$$

which also ensures $\mathbb{K}_c(\mu, \nu) < \infty$.

Remark 3.5. In the proof of Theorem 3.2, note that (iii) $\Rightarrow$ (i) still holds even if we replace c-concavity of φ_o with measurability of φ_o. In other words, Theorem 3.2 is still true even if we weaken the condition (iii) as follows: there exists a proper measurable function $\varphi_o\colon \mathcal{X} \to [-\infty, \infty)$ such that $\operatorname{supp}(\gamma) \subset \partial_c \varphi_o$.

One important implication of Theorem 3.2 is that (iii) implies that the support of any optimal transport plan is contained in $\partial_c \varphi_o$.

Corollary 3.1. *In Theorem 3.2, if (iii) holds for some $\gamma \in \Pi(\mu, \nu)$, the support of any optimal transport plan (including γ) is contained in $\partial_c \varphi_o$.*

Proof. We have already proved that γ is optimal from (iii) $\Rightarrow$ (i) of Theorem 3.2. Let γ' be any optimal transport plan. As in the proof of (iii) $\Rightarrow$ (i), we have

$$\int_{\mathcal{X} \times \mathcal{Y}} c \,\mathrm{d}\gamma' = \int_{\mathcal{X} \times \mathcal{Y}} c \,\mathrm{d}\gamma = \int_{\mathcal{X}} \varphi_o \,\mathrm{d}\mu + \int_{\mathcal{Y}} \varphi_o^c \,\mathrm{d}\nu = \int_{\mathcal{X} \times \mathcal{Y}} (\varphi_o + \varphi_o^c) \,\mathrm{d}\gamma',$$

where the first equality is due to optimality of γ and γ' and the other two equalities are from (3.2) and (3.3). As $\varphi_o(x) + \varphi_o^c(y) \le c(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ by Proposition 3.1, we have $\gamma'(\partial_c \varphi_o) = 1$. Since $\partial_c \varphi_o$ is a closed set by Proposition 3.3, we conclude $\operatorname{supp}(\gamma') \subset \partial_c \varphi_o$. □

3.3 Duality and semi-duality

For φ_o in (iii) of Theorem 3.2, we have indeed shown that

$$\mathbb{K}_c(\mu, \nu) = \int_{\mathcal{X}} \varphi_o \,\mathrm{d}\mu + \int_{\mathcal{Y}} \varphi_o^c \,\mathrm{d}\nu.$$

This is the semi-duality mentioned earlier. It is worth noting that optimality and semi-duality come together; plus, this also means that the dual problem admits a maximizer. Before stating the details, we first formally define the dual and semi-dual problems.

Definition 3.2. Consider the Kantorovich problem between two probability spaces $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ with a cost function c. We write $\varphi \oplus \psi \leq c$ if $\varphi\colon \mathcal{X} \to [-\infty, \infty]$ and $\psi\colon \mathcal{Y} \to [-\infty, \infty]$ are measurable functions such that $\varphi(x) + \psi(y) \leq c(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The dual problem of the Kantorovich problem is defined as follows:

$$\begin{aligned} \text{maximize} \quad & \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \, \mathrm{d}\nu, \\ \text{subject to} \quad & (\varphi, \psi) \in L^1(\mu) \times L^1(\nu), \\ & \varphi \oplus \psi \leq c. \end{aligned}$$

The semi-dual problem of the Kantorovich problem is defined as follows:

$$\begin{aligned} \text{maximize} \quad & \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c \, \mathrm{d}\nu, \\ \text{subject to} \quad & (\varphi, \varphi^c) \in L^1(\mu) \times L^1(\nu). \end{aligned}$$

By definition, one can easily verify that the supremum of the dual problem or the semi-dual problem is bounded above by $\mathbb{K}_c(\mu, \nu)$, i.e.,

$$\sup_{\substack{(\varphi,\psi) \in L^1(\mu) \times L^1(\nu) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \, \mathrm{d}\nu \right), \sup_{(\varphi,\varphi^c) \in L^1(\mu) \times L^1(\nu)} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c \, \mathrm{d}\nu \right) \leq \mathbb{K}_c(\mu, \nu).$$

Under mild assumptions, the dual and semi-dual problems have the same supremum.

Lemma 3.1. *Suppose $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and $\mathbb{K}_c(\mu, \nu) < \infty$. Then,*

$$\sup_{\substack{(\varphi,\psi) \in L^1(\mu) \times L^1(\nu) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \, \mathrm{d}\nu \right) = \sup_{(\varphi,\varphi^c) \in L^1(\mu) \times L^1(\nu)} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c \, \mathrm{d}\nu \right).$$

Proof. Let D be the supremum of the dual problem and S be the supremum of the semi-dual problem. First, note that $(\varphi, \varphi^c) \in L^1(\mu) \times L^1(\nu)$ implies $\varphi\colon \mathcal{X} \to [-\infty, \infty)$ must be proper; hence, $\varphi \oplus \varphi^c \leq c$. Therefore, $D \geq S$ holds. For any $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ such that $\varphi \oplus \psi \leq c$, note that $\psi(y) \leq c(x, y) - \varphi(x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ as $c < \infty$ and $\varphi\colon \mathcal{X} \to [-\infty, \infty)$ must be proper. Therefore, $\psi \leq \varphi^c$ holds, which also implies $(\varphi^c)^- \leq \psi^-$ and thus $(\varphi^c)^- \in L^1(\nu)$. Accordingly,

$$\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c \, \mathrm{d}\nu = \int_{\mathcal{X} \times \mathcal{Y}} (\varphi(x) + \varphi^c(y)) \, \mathrm{d}\gamma(x, y) \quad \forall \gamma \in \Pi(\mu, \nu),$$

where the first equality is due to Lemma 1.5. As $\mathbb{K}_c(\mu, \nu) < \infty$, this shows $\int_{\mathcal{Y}} \varphi^c \, \mathrm{d}\nu < \infty$, proving $\varphi^c \in L^1(\nu)$. This shows that $D \leq S$, and thus $D = S$ holds. □

Now, we finally derive the following result from Theorem 3.2.

Corollary 3.2 (**Semi-Duality**). *Suppose $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and $\mathbb{K}_c(\mu,\nu) < \infty$. Assume*

$$\int_{\mathcal{X}} c(x,y)\,\mathrm{d}\mu(x) < \infty \quad \forall y \in \mathcal{Y} \quad \textit{and} \quad \int_{\mathcal{Y}} c(x,y)\,\mathrm{d}\nu(y) < \infty \quad \forall x \in \mathcal{X}. \tag{MC}$$

If μ and ν are tight, there exists a proper c-concave function $\varphi_o\colon \mathcal{X} \to [-\infty,\infty)$ such that $(\varphi_o, \varphi_o^c) \in L^1(\mu) \times L^1(\nu)$ and

$$\min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\,\mathrm{d}\gamma = \int_{\mathcal{X}} \varphi_o\,\mathrm{d}\mu + \int_{\mathcal{Y}} \varphi_o^c\,\mathrm{d}\nu. \tag{3.5}$$

Accordingly, the following versions of semi-duality hold:

$$\begin{aligned}\mathbb{K}_c(\mu,\nu) &= \max_{(\varphi,\varphi^c)\in L^1(\mu)\times L^1(\nu)} \left(\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c\,\mathrm{d}\nu\right) \\ &= \max_{\substack{\varphi\colon \mathcal{X}\to[-\infty,\infty) \\ \textit{proper and c-concave}}} \left(\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c\,\mathrm{d}\nu\right),\end{aligned} \tag{3.6}$$

where φ_o is a maximizer of both semi-dual problems. Also, the following duality holds:

$$\mathbb{K}_c(\mu,\nu) = \max_{\substack{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu) \\ \varphi\oplus\psi\le c}} \left(\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi\,\mathrm{d}\nu\right), \tag{3.7}$$

where (φ_o, φ_o^c) is a maximizer of the dual problem.

Proof. Tightness of μ and ν guarantees existence of optimal transport plans by Theorem 2.3. Hence, we can invoke the function in (iii) of Theorem 3.2; denote it as φ_o. Then, we have already seen that (3.5) holds. We have also mentioned $(\varphi_o, \varphi_o^c) \in L^1(\mu) \times L^1(\nu)$ in Remark 3.4. Hence, we can see that the first equality of (3.6) holds; for the same reason, we have (3.7). To verify the second equality of (3.6), notice that for any $\varphi\colon \mathcal{X} \to [-\infty,\infty)$ that is proper and c-concave, we have $\varphi^+ \in L^1(\mu)$ and $(\varphi^c)^+ \in L^1(\nu)$ as in Remark 3.4. Hence, as in the proof of Theorem 3.2, we have

$$\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c\,\mathrm{d}\nu \le \mathbb{K}_c(\mu,\nu).$$

This shows the second equality of (3.6). Due to (3.5), we can see that the right-hand sides of (3.6) and (3.7) admit a maximizer: φ_o or (φ_o, φ_o^c). □

Remark 3.6. In Corollary 3.2, we can weaken the constraint of the semi-dual problem $(\varphi, \varphi^c) \in L^1(\mu) \times L^1(\nu)$ by only requiring $\varphi\colon \mathcal{X} \to [-\infty,\infty)$ is proper and measurable such that $\varphi^c\colon \mathcal{Y} \to [-\infty,\infty)$ is also proper. The latter implies that $\varphi^+ \in L^1(\mu)$ and $(\varphi^c)^+ \in L^1(\nu)$ as in the proof of Theorem 3.2 due to (MC). Hence, we still have

$$\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu + \int_{\mathcal{Y}} \varphi^c\,\mathrm{d}\nu \le \mathbb{K}_c(\mu,\nu).$$

Remark 3.7. The c-concave function φ_o in Corollary 3.2 enjoys regularity properties that inherit from c. For instance, boundedness of c implies that φ_o and φ_o^c are bounded. Next, suppose we have equipped $\mathcal{X}$ and $\mathcal{Y}$ with compatible metrics, respectively. Then, uniform continuity of c implies that both φ_o and φ_o^c are uniformly continuous by Proposition 3.3. Similarly, L-Lipschitzness of c implies that both φ_o and φ_o^c are L-Lipschitz. Consequently, we can replace the constraint $L^1(\mu) \times L^1(\nu)$ of the dual problem with a smaller class without decreasing the maximum as long as that class contains (φ_o, φ_o^c) as shown in the following lemmas.

Lemma 3.2. *Equipping $\mathcal{X}$ and $\mathcal{Y}$ with their compatible metrics, suppose $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is uniformly continuous and bounded. If μ and ν are tight, the following duality holds:*

$$\mathbb{K}_c(\mu, \nu) = \max_{\substack{(\varphi,\psi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu \right),$$

where the right-hand side admits a maximizer.

Proof. We can apply Corollary 3.2; boundedness of c implies $\mathbb{K}_c(\mu, \nu) < \infty$ and (MC). Then, $\varphi_o^c \in C_b(\mathcal{X})$ and $\varphi_o^c \in C_b(\mathcal{Y})$ are guaranteed by Proposition 3.3. As discussed in Remark 3.7,

$$\max_{\substack{(\varphi,\psi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu \right) = \max_{\substack{(\varphi,\psi) \in L^1(\mu) \times L^1(\nu) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu \right)$$

as $(\varphi_o, \varphi_o^c) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \subset L^1(\mu) \times L^1(\nu)$ is a maximizer of the dual problem. □

Remark 3.8. Notice that 3.2 is applicable if $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a continuous cost function and $\mathcal{X}, \mathcal{Y}$ are compact metrizable spaces.

Lemma 3.3. *Equipping $\mathcal{X}$ and $\mathcal{Y}$ with their compatible metrics, say $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$, respectively, suppose $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is L-Lipschitz and bounded. If μ and ν are tight, the following duality holds:*

$$\mathbb{K}_c(\mu, \nu) = \max_{\substack{(\varphi,\psi) \in BL(\mathcal{X},\rho_{\mathcal{X}}) \times BL(\mathcal{Y},\rho_{\mathcal{Y}}) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu \right),$$

where the right-hand side admits a maximizer.

Proof. We can apply Corollary 3.2; boundedness of c implies $\mathbb{K}_c(\mu, \nu) < \infty$ and (MC). Then, $\varphi_o^c \in BL(\mathcal{X}, \rho_{\mathcal{X}})$ and $\varphi_o^c \in BL(\mathcal{Y}, \rho_{\mathcal{Y}})$ are guaranteed by Proposition 3.3; in fact, φ_o and φ_o^c are L-Lipschitz. As discussed in Remark 3.7,

$$\max_{\substack{(\varphi,\psi) \in BL(\mathcal{X},\rho_{\mathcal{X}}) \times BL(\mathcal{Y},\rho_{\mathcal{Y}}) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu \right) = \max_{\substack{(\varphi,\psi) \in L^1(\mu) \times L^1(\nu) \\ \varphi \oplus \psi \leq c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu \right)$$

as $(\varphi_o, \varphi_o^c) \in BL(\mathcal{X}, \rho_{\mathcal{X}}) \times BL(\mathcal{Y}, \rho_{\mathcal{Y}}) \subset L^1(\mu) \times L^1(\nu)$ is a maximizer of the dual problem. □

Lastly, we extend the duality result to the case where c is lower semi-continuity. The key idea is to approximate c with a sequence of bounded Lipschitz—hence uniformly continuous—functions by Lemma 2.2.

Theorem 3.3 (**Kantorovich Duality**). *Suppose c is lower semi-continuous. If μ and ν are tight, the Kantorovich duality holds:*

$$\mathbb{K}_c(\mu,\nu) = \sup_{\substack{(\varphi,\psi)\in C_b(\mathcal{X})\times C_b(\mathcal{Y})\\ \varphi\oplus\psi\le c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu\right). \tag{3.8}$$

Proof. It suffices prove $\le$ instead of $=$ in (3.8). As c is lower semi-continuous and bounded from below, we can find a sequence $(c_n)_{n\in\mathbb{N}}$ of uniformly continuous and bounded functions such that $c_n \uparrow c$ by Lemma 2.2. Let γ_n be an optimal transport plan with respect to the cost function c_n; the existence is guaranteed by Theorem 2.3. The same theorem states the compactness of $\Pi(\mu,\nu)$. Hence, by taking a subsequence if necessary, we may assume that $(\gamma_n)_{n\in\mathbb{N}}$ converges weakly to some $\gamma\in\Pi(\mu,\nu)$. Then the duality follows since

$$\begin{aligned}\mathbb{K}_c(\mu,\nu) &\le \int c\,\mathrm{d}\gamma\\ &= \lim_{m\to\infty}\int c_m\,\mathrm{d}\gamma \quad (\because\ \text{monotone convergence theorem})\\ &= \lim_{m\to\infty}\left(\lim_{n\to\infty}\int c_m\,\mathrm{d}\gamma_n\right) \quad (\because\ \gamma_n\to\gamma \text{ weakly})\\ &\le \limsup_{n\to\infty}\int c_n\,\mathrm{d}\gamma_n \quad (\because\ c_m\le c_n \text{ for } n\ge m).\end{aligned}$$

Now, we apply Lemma 3.2 to each c_n, which is possible since c_n is continuous and bounded and thus $\Pi_{c_n}(\mu,\nu)$ is nonempty. Hence,

$$\int c_n\,\mathrm{d}\gamma_n = \max_{\substack{(\varphi,\psi)\in C_b(\mathcal{X})\times C_b(\mathcal{Y})\\ \varphi\oplus\psi\le c_n}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu\right) \le \sup_{\substack{(\varphi,\psi)\in C_b(\mathcal{X})\times C_b(\mathcal{Y})\\ \varphi\oplus\psi\le c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu\right),$$

where the last inequality holds since $c_n \le c$. Therefore, we have

$$\mathbb{K}_c(\mu,\nu) \le \sup_{\substack{(\varphi,\psi)\in C_b(\mathcal{X})\times C_b(\mathcal{Y})\\ \varphi\oplus\psi\le c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu\right).$$

□

Remark 3.9. In Theorem 3.3, one can show the following:

$$\mathbb{K}_c(\mu,\nu) = \sup_{\substack{(\varphi,\psi)\in BL(\mathcal{X},\rho_{\mathcal{X}})\times BL(\mathcal{Y},\rho_{\mathcal{Y}})\\ \varphi\oplus\psi\le c}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu\right),$$

where $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ are compatible metrics of $\mathcal{X}$ and $\mathcal{Y}$, respectively; simply use Lemma 3.3 instead of Lemma 3.2 in the proof of Theorem 3.3.

3.4 Omitted proofs

Proof of Proposition 3.2. (i) By definition, we have $\varphi^c(y) \leq c(x,y) - \varphi(x)$, which implies $\varphi(x) - c(x,y) \leq -\varphi^c(y)$. As c is real-valued, by adding $c(x,y)$, we have $\varphi(x) \leq c(x,y) - \varphi^c(y)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Therefore, for all $x \in \mathcal{X}$,

$$\varphi^{cc}(x) = \inf_{y \in \mathcal{Y}} (c(x,y) - \varphi^c(y)) \geq \varphi(x).$$

(ii) Clearly, $\varphi = (\varphi^c)^c$ implies that φ is c-concave by definition. Conversely, if φ is c-concave, i.e., $\varphi = \psi^c$ for some $\psi \colon \mathcal{Y} \to [-\infty, \infty]$, by applying (i) to ψ based on symmetric, we have $\varphi^c = \psi^{cc} \geq \psi$. From $\varphi^c \geq \psi$, we have $\varphi^{cc} \leq \psi^c = \varphi$. Hence, $\varphi = \varphi^{cc}$. □

Proof of Proposition 3.3. (i) Let $\varphi \colon \mathcal{X} \to [-\infty, \infty]$ be a c-concave function and $|c| \leq M$ for some $M > 0$. Then, $\varphi^c(y) \leq c(x,y) - \varphi(x) \leq M - \varphi(x)$ implies that φ^c is bounded above by some $L > 0$. This also means that $\varphi = (\varphi^c)^c$ is bounded above by symmetry and $\varphi = \varphi^{cc} \geq -M - L$, i.e., φ is bounded from below.
(ii) Let $\varphi \colon \mathcal{X} \to [-\infty, \infty]$ be a c-concave function so that $\varphi = (\varphi^c)^c$. Then for each $y \in \mathcal{Y}$ a function $x \mapsto c(x,y) - \varphi^c(y)$ is continuous as c is continuous, thus φ is upper semi-continuous since it is defined by the infimum of a collection of upper semi-continuous functions.
(iii) Nothing to prove if $\partial_c \varphi = \emptyset$. Assume $\partial_c \varphi \neq \emptyset$, equivalently, $\varphi \colon \mathcal{X} \to [-\infty, \infty)$ is proper. Suppose a sequence (x_n, y_n) in $\partial_c \varphi$ converges to $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Then,

$$\begin{aligned} c(x,y) &= \lim_{n \to \infty} c(x_n, y_n) \quad (\because \text{ continuity of } c) \\ &\leq \limsup_{n \to \infty} \varphi(x_n) + \limsup_{n \to \infty} \varphi^c(y_n) \quad (\because \ (x_n, y_n) \in \partial_c \varphi) \\ &\leq \varphi(x) + \varphi^c(y) \quad (\because \text{ upper semi-continuity}) \\ &\leq c(x,y) \quad (\because \text{ Proposition 3.1}). \end{aligned}$$

(iv) Let ρ_1 and ρ_2 be metrics on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Fix a proper c-concave function $\varphi \colon \mathcal{X} \to [-\infty, \infty)$ be a c-concave function. Uniform continuity of c implies that for $\varepsilon > 0$, there exists $\delta > 0$ such that $\rho_1(x,x') + \rho_2(y,y') \leq \delta$ implies $|c(x,y) - c(x',y')| \leq \varepsilon$. Therefore, whenever $\rho_1(x,x') \leq \delta$, we have $|c(x,y) - c(x',y)| \leq \varepsilon$ for all $y \in \mathcal{Y}$. Hence,

$$\varphi(x) = \inf_{y \in \mathcal{Y}} (c(x,y) - \varphi^c(y)) \leq \inf_{y \in \mathcal{Y}} (c(x',y) - \varphi^c(y)) + \varepsilon = \varphi(x') + \varepsilon,$$

and by symmetry $\varphi(x') \leq \varphi(x) + \varepsilon$. This means φ must be real-valued and $|\varphi(x) - \varphi(x')| \leq \varepsilon$. Hence, φ is uniformly continuous.
(v) Mimic the proof of (iv): $|c(x,y) - c(x',y)| \leq L\rho_1(x,x')$ implies $|\varphi(x) - \varphi(x')| \leq L\rho_1(x,x')$. □

Proof of Theorem 3.1. Fix $(x_0, y_0) \in \Pi$. For each $x \in \mathcal{X}$, let

$$\varphi(x) = \inf \left\{ \sum_{i=1}^{n} \left(c(x_i, y_{i-1}) - c(x_{i-1}, y_{i-1})\right) + c(x, y_n) - c(x_n, y_n) : \forall n \in \mathbb{N}, (x_i, y_i) \in \Pi \right\}.$$

Clearly, $\varphi \colon \mathcal{X} \to [-\infty, \infty)$ since $\varphi(x) \leq c(x, y_0) - c(x_0, y_0)$. Also, due to c-cyclical monotonicity of Π, we have $\varphi(x_0) \geq 0$ and $\varphi(x) + c(x_0, y) - c(x, y) \geq 0$ and hence $\varphi(x) > -\infty$ for any $(x, y) \in \Pi$, which proves φ is proper. For each $y \in \mathcal{Y}$, let

$$-\psi(y) = \inf \left\{ \sum_{i=1}^{n} \left(c(x_i, y_{i-1}) - c(x_{i-1}, y_{i-1})\right) - c(x_n, y_n) : \forall n \in \mathbb{N}, (x_i, y_i) \in \Pi, y_n = y \right\}.$$

By definition, $-\psi(y) < \infty$, equivalently $\psi(y) > -\infty$ if and only if $(x, y) \in \Pi$ for some $x \in \mathcal{X}$. One can verify that

$$\varphi(x) = \inf_{y \in \mathcal{Y}} \left(c(x, y) - \psi(y)\right),$$

which proves $\varphi = \psi^c$ and hence φ is c-concave. For each $(x, y) \in \Pi$, we claim that $\varphi(x) + \varphi^c(y) = c(x, y)$. Clearly, $\varphi(x) + \varphi^c(y) \leq c(x, y)$ holds since φ is proper. Since $\varphi(x) > -\infty$, for each $\varepsilon > 0$ we can find $y_\varepsilon \in \mathcal{Y}$ such that

$$\varphi(x) \leq c(x, y_\varepsilon) - \psi(y_\varepsilon) < \varphi(x) + \varepsilon.$$

Then

$$-\psi(y) \leq -\psi(y_\varepsilon) + c(x, y_\varepsilon) - c(x, y) < \varphi(x) - c(x, y) + \varepsilon,$$

where the first inequality holds by the definition of $-\psi$. This proves $-\psi(y) \leq \varphi(x) - c(x, y)$ and hence $c(x, y) \leq \varphi(x) + \psi(y)$. □

Proof of Proposition 3.4. Let γ be an optimal transport plan and suppose $\operatorname{supp}(\gamma)$ is not c-cyclically monotone. By definition, we can find $(x_1, y_1), \ldots, (x_n, y_n) \in \operatorname{supp}(\gamma)$ and a permutation $\sigma \in \operatorname{Perm}(n)$ such that

$$C := \sum_{i=1}^{n} c(x_i, y_i) - \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) > 0.$$

As c is continuous, for each $i \in [n]$, we can find open neighborhoods $U_i \subset \mathcal{X}$ and $V_i \subset \mathcal{Y}$ of x_i and y_i, respectively, such that

$$\begin{aligned} c(x, y) &> c(x_i, y_i) - \varepsilon \quad \forall (x, y) \in U_i \times V_i, \\ c(x, y) &< c(x_i, y_{\sigma(i)}) + \varepsilon \quad \forall (x, y) \in U_i \times V_{\sigma(i)}, \end{aligned}$$

where $\varepsilon > 0$ is a constant such that $\varepsilon < \frac{C}{2n}$. For each $i \in [n]$, let $S_i = U_i \times V_i$; then, $(x_i, y_i) \in \operatorname{supp}(\gamma)$ implies $\gamma(S_i) > 0$; hence, we can define $\gamma_i(A) := \gamma(A \cap S_i)/\gamma(S_i)$ for all

$A \in \mathscr{B}(\mathcal{X} \times \mathcal{Y})$; also, let μ_i and ν_i be the marginals of γ_i on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Take $\eta > 0$ such that $n\eta < \min_{i\in[n]} \gamma(S_i)$ and let

$$\gamma^\star = \gamma - \eta \sum_{i=1}^n \gamma_i + \eta \sum_{i=1}^n \mu_i \otimes \nu_{\sigma(i)}.$$

By definition,

$$\gamma(S) - \eta \sum_{i=1}^n \gamma_i(S) \geq \gamma(S) - \frac{1}{n} \sum_{i=1}^n \gamma(S \cap S_i) \geq 0 \quad \forall S \in \mathscr{B}(\mathcal{X} \times \mathcal{Y}).$$

Hence, $\gamma^\star \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$. Also, $\gamma^\star \in \Pi(\mu, \nu)$ follows as

$$\begin{aligned} \gamma^\star(A \times \mathcal{Y}) &= \gamma(A \times \mathcal{Y}) - \eta \sum_{i=1}^n \mu_i(A) + \eta \sum_{i=1}^n \mu_i(A) = \mu(A) \quad \forall A \in \mathscr{B}(\mathcal{X}), \\ \gamma^\star(\mathcal{X} \times B) &= \gamma(\mathcal{X} \times B) - \eta \sum_{i=1}^n \nu_i(B) + \eta \sum_{i=1}^n \nu_{\sigma(i)}(B) = \nu(B) \quad \forall B \in \mathscr{B}(\mathcal{Y}). \end{aligned}$$

For each $i \in [n]$, as γ_i and $\mu_i \otimes \nu_{\sigma(i)}$ are concentrated on S_i and $U_i \times V_{\sigma(i)}$, respectively,

$$\int c \, \mathrm{d}\gamma_i \geq c(x_i, y_i) - \varepsilon \quad \text{and} \quad \int c \, \mathrm{d}\mu_i \otimes \nu_{\sigma(i)} \leq c(x_i, y_{\sigma(i)}) + \varepsilon.$$

Therefore,

$$\begin{aligned} \int c \, \mathrm{d}\gamma^\star &= \int c \, \mathrm{d}\gamma - \eta \sum_{i=1}^n \int c \, \mathrm{d}\gamma_i + \eta \sum_{i=1}^n \int c \, \mathrm{d}\mu_i \otimes \nu_{\sigma(i)} \\ &\leq \int c \, \mathrm{d}\gamma - \eta \sum_{i=1}^n (c(x_i, y_i) - \varepsilon) + \eta \sum_{i=1}^n \big(c(x_i, y_{\sigma(i)}) + \varepsilon\big) \\ &< \int c \, \mathrm{d}\gamma, \end{aligned}$$

where the last strict inequality holds since the assumption $\mathbb{K}_c(\mu, \nu) < \infty$ implies $\int c \, \mathrm{d}\gamma < \infty$. This contradicts that γ is optimal. Hence, $\mathrm{supp}(\gamma)$ must be c-cyclically monotone. □

4 Applications of Optimality and Duality

Settings As in Section 3, $\mathcal{X}$ and $\mathcal{Y}$ are separable metrizable spaces unless otherwise stated; we consider the Kantorovich problem between $(\mathcal{X}, \mathscr{B}(\mathcal{X}), \mu)$ and $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}), \nu)$ with a cost function c.

4.1 Stability of optimal transport plans

Given sequences $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ in $\mathscr{P}(\mathcal{X})$ and $\mathscr{P}(\mathcal{Y})$, respectively, consider a sequence $(\gamma_n)_{n\in\mathbb{N}}$ in $\mathscr{P}(\mathcal{X})$ such that γ_n is an optimal transport plan from μ_n to ν_n, i.e., $\gamma_n \in \Pi_c(\mu_n, \nu_n)$, for all $n \in \mathbb{N}$. If $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ converge weakly to $\mu \in \mathscr{P}(\mathcal{X})$ and $\nu \in \mathscr{P}(\mathcal{Y})$, respectively, we show that $(\gamma_n)_{n\in\mathbb{N}}$ has a subsequence converges weakly to some optimal transport plan from μ and ν, i.e., the weak limit is contained in $\Pi_c(\mu, \nu)$. If $\Pi_c(\mu, \nu)$ is a singleton, i.e., there exists a unique optimal transport plan from μ to ν, then the whole $(\gamma_n)_{n\in\mathbb{N}}$ sequence converges weakly to that unique optimal transport plan.

Lemma 4.1. *Let $\mathcal{Z}$ be a separable metric space. Fix $N \in \mathbb{N}$.*

(i) For $\gamma \in \mathscr{P}(\mathcal{Z})$, let $\gamma^{\otimes N} \in \mathscr{P}(\mathcal{Z}^N)$ denote the product of N copies of γ. Then, $\operatorname{supp}(\gamma^{\otimes N}) = \operatorname{supp}(\gamma)^N$.

(ii) Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $\mathscr{P}(\mathcal{Z})$ converging weakly to some $\gamma \in \mathscr{P}(\mathcal{Z})$. Then, $(\gamma_n^{\otimes N})_{n\in\mathbb{N}}$ converges weakly to $\gamma^{\otimes N}$ in $\mathscr{P}(\mathcal{Z}^N)$.

Proof. We prove for $N = 2$. Let $S = \operatorname{supp}(\gamma)$. Note that $S^2 = S \times S \subset \mathcal{Z} \times \mathcal{Z}$ is closed and $\gamma^{\otimes 2}(S^2) = \gamma(S)^2 = 1$. Hence, $\operatorname{supp}(\gamma^{\otimes 2}) \subset S^2$. Now, suppose $(z_1, z_2) \in \mathcal{Z}^2 \backslash \operatorname{supp}(\gamma^{\otimes 2})$; by definition, we can find an open neighborhood U of (z_1, z_2) such that $\gamma^{\otimes 2}(U) = 0$. Further assume $(z_1, z_2) \in S^2$, i.e., $z_1, z_2 \in S$. For any $r > 0$, let $B_r(z) \subset \mathcal{Z}$ denote the open ball of radius r centered at $z \in \mathcal{Z}$. We can find $r > 0$ such that $B_r(z_1) \times B_r(z_2) \subset U$. Then, $\gamma^{\otimes 2}(B_r(z_1) \times B_r(z_2)) = 0$ as $\gamma^{\otimes 2}(U) = 0$, which implies $\gamma(B_r(z_1)) = 0$ or $\gamma(B_r(z_2)) = 0$. This contradicts $z_1, z_2 \in S$. Therefore, $\mathcal{Z}^2 \backslash \operatorname{supp}(\gamma^{\otimes 2}) \subset \mathcal{Z}^2 \backslash S^2$. Hence, $\operatorname{supp}(\gamma^{\otimes 2}) = S^2$. For (ii), refer to Theorem 2.8 of [Bil99]. □

Lemma 4.2. *Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $\mathscr{P}(\mathcal{X} \times \mathcal{Y})$ converging weakly to $\gamma \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$. Suppose $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and* $\operatorname{supp}(\gamma_n)$ *is c-cyclically monotone for all $n \in \mathbb{N}$. Then,* $\operatorname{supp}(\gamma)$ *is c-cyclically monotone.*

Proof. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For each $N \in \mathbb{N}$, define

$$A_N = \left\{ ((x_1, y_1), \ldots, (x_N, y_N)) \in \mathcal{Z}^N : \sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}) \ \forall \sigma \in \operatorname{Perm}(N) \right\}.$$

Note that c-cyclical monotonicity implies $\operatorname{supp}(\gamma_n)^N \subset A_N$. Hence, using Lemma 4.1, we have

$$\gamma_n^{\otimes N}(A_N) \geq \gamma_n^{\otimes N}(\operatorname{supp}(\gamma_n)^N) = \gamma_n^{\otimes N}(\operatorname{supp}(\gamma_n^{\otimes N})) = 1.$$

As c is continuous, A_N is a closed set. By Theorem 2.1, we have

$$\gamma^{\otimes N}(A_N) \geq \limsup_{n\to\infty} \gamma_n^{\otimes N}(A_N) = 1.$$

Hence, we conclude $\operatorname{supp}(\gamma^{\otimes N}) \subset A_N$. As $\operatorname{supp}(\gamma^{\otimes N}) = \operatorname{supp}(\gamma)^N$, we have $\operatorname{supp}(\gamma)^N \subset A_N$. Since this holds for all $N \in \mathbb{N}$, we conclude that $\operatorname{supp}(\gamma)$ is c-cyclically monotone. □

Theorem 4.1 **(Stability of Optimal Transport Plans).** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Suppose $c\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is continuous and $\mathbb{K}_c(\mu,\nu) < \infty$. Assume*

$$\int_{\mathcal{X}} c(x,y)\,\mathrm{d}\mu(x) < \infty \quad \forall y \in \mathcal{Y} \quad \textit{and} \quad \int_{\mathcal{Y}} c(x,y)\,\mathrm{d}\nu(y) < \infty \quad \forall x \in \mathcal{X}. \tag{MC}$$

Let $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ be sequences converging weakly to μ in $\mathscr{P}(\mathcal{X})$ and ν in $\mathscr{P}(\mathcal{Y})$, respectively. For each $n \in \mathbb{N}$, assume $\mathbb{K}_c(\mu_n,\nu_n) < \infty$ and pick $\gamma_n \in \Pi_c(\mu_n,\nu_n)$.

(i) $(\gamma_n)_{n\in\mathbb{N}}$ has a subsequence that converges weakly to some $\gamma \in \Pi_c(\mu,\nu)$.

(ii) $(\gamma_n)_{n\in\mathbb{N}}$ converges weakly to $\gamma^\star$ provided $\Pi_c(\mu,\nu) = \{\gamma^\star\}$.

Proof. Recall that $\operatorname{supp}(\gamma_n)$ is c-cyclically monotone for all $n \in \mathbb{N}$ by Proposition 3.4. Also, $(\gamma_n)_{n\in\mathbb{N}}$ has a subsequence that converges weakly to some $\gamma \in \Pi(\mu,\nu)$ by Proposition 2.4. By Lemma 4.2, $\operatorname{supp}(\gamma)$ is c-cyclically monotone, which implies $\gamma \in \Pi_c(\mu,\nu)$ by Theorem 3.2. Now, suppose $\Pi_c(\mu,\nu) = \{\gamma^\star\}$. By Theorem 2.2, any subsequence, say □, of $(\gamma_n)_{n\in\mathbb{N}}$ must converge weakly. Applying (i) to the sequence □, we conclude that □ has a further subsequence converging to $\gamma^\star$, which means the weak limit of □ must be $\gamma^\star$. In summary, any subsequence of $(\gamma_n)_{n\in\mathbb{N}}$ converges weakly to $\gamma^\star$, which proves that the whole sequence $(\gamma_n)_{n\in\mathbb{N}}$ must converge weakly to $\gamma^\star$. □

4.2 Kantorovich-Rubinstein theorem

We study the case where c is a metric on $\mathcal{X} = \mathcal{Y}$. In this case, c-concavity and c-transforms become very simple as follows.

Lemma 4.3. *Given a set $\mathcal{X}$, suppose $c\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ is a metric on $\mathcal{X}$.*[5] *A proper function $\varphi\colon \mathcal{X} \to [-\infty,\infty)$ is c-concave if and only if φ is real-valued and*

$$|\varphi(x) - \varphi(y)| \leq c(x,y) \quad \forall x,y \in \mathcal{X}.$$

In this case, $\varphi^c = -\varphi$.

[5] We may assume c is a pseudometric.

Proof. If φ is c-concave, $\varphi = \varphi^{cc}$ implies that for any $x_1, x_2 \in \mathcal{X}$,

$$\varphi(x_1) = \inf_{y \in \mathcal{X}} (c(x_1, y) - \varphi^c(y)) \leq \inf_{y \in \mathcal{X}} (c(x_2, y) - \varphi^c(y)) + c(x_1, x_2) = \varphi(x_2) + c(x_1, x_2).$$

This implies that $\varphi \equiv -\infty$ if $\varphi(x) = -\infty$ for some $x \in \mathcal{X}$. As φ is proper, we conclude that φ be real-valued and $\varphi(x_1) - \varphi(x_2) \leq c(x_1, x_2)$ holds; by symmetry, $|\varphi(x_1) - \varphi(x_2)| \leq c(x_1, x_2)$. Conversely, $|\varphi(x) - \varphi(y)| \leq c(x, y)$ implies $-\varphi(x) \leq c(x, y) - \varphi(y)$ for all $x, y \in \mathcal{X}$. Hence,

$$-\varphi(x) = \inf_{y \in \mathcal{X}} (c(x, y) - \varphi(y)),$$

where the equality is due to $c(x, x) = 0$. Hence, $\varphi^c = -\varphi$. □

Theorem 4.2 (Kantorovich-Rubinstein). *Let c be a metric on $\mathcal{X} = \mathcal{Y}$ that is continuous with respect to the product topology of $\mathcal{X} \times \mathcal{X}$. Assume*

$$\int_{\mathcal{X}} c(x, x_0) \, \mathrm{d}\mu(x) + \int_{\mathcal{X}} c(x, x_0) \, \mathrm{d}\nu(x) < \infty \quad \exists x_0 \in \mathcal{X}.$$

If μ and ν are tight,

$$\inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c \, \mathrm{d}\gamma = \sup_{\substack{\varphi \colon \mathcal{X} \to \mathbb{R} \\ \|\varphi\|_c \leq 1}} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu - \int_{\mathcal{X}} \varphi \, \mathrm{d}\nu \right), \tag{4.1}$$

where we define for any $\varphi \colon \mathcal{X} \to \mathbb{R}$,

$$\|\varphi\|_c := \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{c(x, y)}.$$

Moreover, the right-hand side of (4.1) admits a maximizer.

Proof. Notice that $\mathbb{K}_c(\mu, \nu) < \infty$ since

$$\int_{\mathcal{X} \times \mathcal{X}} c(x, y) \, \mathrm{d}\mu \otimes \nu \leq \int_{\mathcal{X}} c(x, x_0) \, \mathrm{d}\mu(x) + \int_{\mathcal{X}} c(x, x_0) \, \mathrm{d}\nu(x) < \infty.$$

We use the semi-duality (3.6) of Corollary 3.2:

$$\mathbb{K}_c(\mu, \nu) = \sup_{\substack{\varphi \colon \mathcal{X} \to [-\infty, \infty) \\ \text{proper and } c\text{-concave}}} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{X}} \varphi^c \, \mathrm{d}\nu \right).$$

Due to the Lemma 4.3,

$$\sup_{\substack{\varphi \colon \mathcal{X} \to [-\infty, \infty) \\ \text{proper and } c\text{-concave}}} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{X}} \varphi^c \, \mathrm{d}\nu \right) = \sup_{\substack{\varphi \colon \mathcal{X} \to \mathbb{R} \\ \|\varphi\|_c \leq 1}} \left(\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu - \int_{\mathcal{X}} \varphi \, \mathrm{d}\nu \right).$$

□

Remark 4.1. Remember that tightness of μ and ν—which ensures $\Pi_c(\mu, \nu) \neq \emptyset$—is required to invoke Corollary 3.2. Theorem 11.8.2 of [Dud02] shows that (4.1) still holds without tightness.

5 Optimal Transport in Euclidean Spaces

This section studies optimal transport problems between Euclidean spaces equipped with their Borel σ-algebras. So far, we have defined optimal transport problems between arbitrary probability spaces and derived general results (optimality, (semi-)duality, etc.). In the Euclidean space case, we can establish much more concrete results that not only provide rich theory, but also play a crucial role in a number of optimal transport applications.

Settings Throughout the section, both Monge and Kantorovich problems are considered between $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d), \mu)$ and $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d), \nu)$ with $d \in \mathbb{N}$, i.e., $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ equipped with the Borel σ-algebra on $\mathbb{R}^d$, where c denotes the cost function; we call c the quadratic cost if $c(x, y) = \frac{1}{2}\|x - y\|_2^2$ for all $x, y \in \mathbb{R}^d$. When $d = 1$, we write F_μ and F_ν to denote the distribution functions of μ and ν, respectively, as in Lemma 1.2.

Remark 5.1. As $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, there are many options for c, for instance, any distance-based function, say $c(x, y) = \|x - y\|_2$, leads to a natural notion of the unit cost to transport from $x \in \mathbb{R}^d$ to $y \in \mathbb{R}^d$. On the other hand, if we assume instead $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$ with $d_1 \neq d_2$, it is unclear how to design a function c over $\mathcal{X} \times \mathcal{Y}$ that represents intuitive cost associated with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. This is the rationale behind the setting $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$.

5.1 One-dimensional analysis

Letting $d = 1$, we consider a case where $c(x, y) = h(x - y)$ for some convex function $h \colon \mathbb{R} \to \mathbb{R}$ that is bounded from below, e.g., $c(x, y) = |x - y|^p$ for $p \geq 1$. In this setting, we can derive a closed form of an optimal transport plan thanks to Theorem 3.2, which leads to the transport plan $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ introduced in Proposition 1.5. The key idea is the following "co-monotonicity".

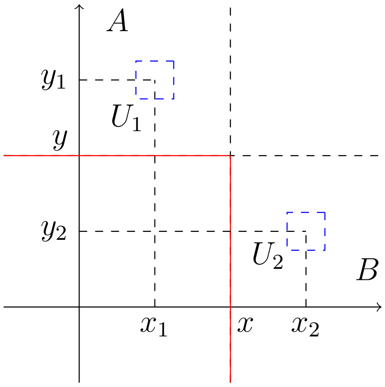


Figure 2: Proof of Lemma 5.1.

Lemma 5.1. *Letting $d = 1$, for $\gamma \in \Pi(\mu, \nu)$, the following are equivalent.*

(i) For any $(x_1, y_1), (x_2, y_2) \in \operatorname{supp}(\gamma)$ satisfying $x_1 < x_2$, we have $y_1 \leq y_2$.

(ii) $\gamma = (F_\mu^{-1}, F_\nu^{-1})_\# \lambda$.

Proof. Let F_γ be the distribution function of γ. Suppose (i). Fix $x, y \in \mathbb{R}$. If we prove that $F_\gamma(x, y) = \min\{F_\mu(x), F_\nu(y)\}$, then (ii) follows by Proposition 1.5. Let $A = (-\infty, x] \times (y, \infty)$ and $B = (x, \infty) \times (-\infty, y]$ so that

$$F_\gamma(x, y) + \gamma(A) = F_\mu(x) \quad \text{and} \quad F_\gamma(x, y) + \gamma(B) = F_\nu(y).$$

Observe that at least one of $\operatorname{supp}(\gamma) \cap A$ and $\operatorname{supp}(\gamma) \cap B$ must be empty; otherwise, we can find $(x_1, y_1) \in \operatorname{supp}(\gamma) \cap A$ and $(x_2, y_2) \in \operatorname{supp}(\gamma) \cap B$, which leads to $x_1 \leq x < x_2$ and $y_2 \leq y < y_1$, contradicting (i). Accordingly, $\gamma(A) = 0$ or $\gamma(B) = 0$ must hold, which implies $F_\gamma(x, y) = \min\{F_\mu(x), F_\nu(y)\}$. Suppose (ii). Suppose $(x_1, y_1), (x_2, y_2) \in \operatorname{supp}(\gamma)$ satisfy $x_1 < x_2$ and $y_1 > y_2$. By definition, we can find $\varepsilon > 0$ such that $x_1 + \varepsilon < x_2 - \varepsilon$, $y_1 - \varepsilon > y_2 + \varepsilon$, and $\gamma(U_1), \gamma(U_2) > 0$, where $U_1 := (x_1 - \varepsilon, x_1 + \varepsilon) \times (y_1 - \varepsilon, y_1 + \varepsilon)$ and $U_2 := (x_2 - \varepsilon, x_2 + \varepsilon) \times (y_2 - \varepsilon, y_2 + \varepsilon)$. Pick $x, y \in \mathbb{R}$ such that $x_1 + \varepsilon < x < x_2 - \varepsilon$ and $y_1 - \varepsilon > y > y_2 + \varepsilon$. Then,

$$\begin{aligned} F_\mu(x) &= F_\gamma(x, y) + \gamma((-\infty, x] \times (y, \infty)) \geq F_\gamma(x, y) + \gamma(U_1) > F_\gamma(x, y), \\ F_\nu(y) &= F_\gamma(x, y) + \gamma((x, \infty) \times (-\infty, y]) \geq F_\gamma(x, y) + \gamma(U_2) > F_\gamma(x, y), \end{aligned}$$

which contradicts $F_\gamma(x, y) = \min\{F_\mu(x), F_\nu(y)\}$. □

Proposition 5.1. *Letting $d = 1$, suppose $c(x, y) = h(x-y)$ for some strictly convex function $h\colon \mathbb{R} \to \mathbb{R}$ that is bounded from below. If the optimal transport cost is finite, $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ is the unique optimal transport plan.*

Proof. Let γ be an optimal transport plan, which must exists by Theorem 2.3. As the optimal transport cost is assumed to be finite, $\operatorname{supp}(\gamma)$ is c-cyclically monotone by Proposition 3.4. We prove that strict convexity of h implies (i) of Lemma 5.1. To this end, suppose $(x_1, y_1), (x_2, y_2) \in \operatorname{supp}(\gamma)$ satisfy $x_1 < x_2$ and $y_1 > y_2$. By construction, we have $x_1 - y_1 < x_1 - y_2, x_2 - y_1 < x_2 - y_2$, which implies, by strict convexity,

$$\begin{aligned} h(x_1 - y_2) &< \frac{(x_2 - x_1)h(x_1 - y_1) + (y_1 - y_2)h(x_2 - y_2)}{x_2 - x_1 + y_1 - y_2}, \\ h(x_2 - y_1) &< \frac{(y_1 - y_2)h(x_1 - y_1) + (x_2 - x_1)h(x_2 - y_2)}{x_2 - x_1 + y_1 - y_2}. \end{aligned}$$

Combining the two inequalities, $h(x_1 - y_2) + h(x_2 - y_1) < h(x_1 - y_1) + h(x_2 - y_2)$, which contradicts c-cyclical monotonicity of $\operatorname{supp}(\gamma)$. Therefore, (i) of Lemma 5.1 must hold. Hence, $\gamma = (F_\mu^{-1}, F_\nu^{-1})_\# \lambda$, concluding that $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ is the unique optimal transport plan; we also conclude that the support of $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ is c-cyclically monotone. □

Next, we consider a case where h is convex but may not be strictly convex; in this case, $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ is still an optimal transport plan, but uniqueness is not guaranteed. To see this, we use the following lemma to approximate h using a strictly convex function.

Lemma 5.2. *Let $h\colon \mathbb{R} \to \mathbb{R}$ be a convex function that is bounded from below. If h is not a constant function, for any $\varepsilon > 0$, there exists a strictly convex function $h_\varepsilon \colon \mathbb{R} \to \mathbb{R}$ such that $h \le h_\varepsilon \le (1+\varepsilon)h + \varepsilon$.*

Proof. Without loss of generality, suppose $h \ge 0$. As a convex function bounded below by an affine function, we can find $a, b \in \mathbb{R}$ such that $h(x) \ge ax + b$ for all $x \in \mathbb{R}$. Since h is not a constant function, we may assume $a \neq 0$. Also, $h \ge 0$ implies $h(x) \ge (ax+b)^+$. Let

$$f(x) = \frac{\sqrt{4 + (ax+b)^2} + ax + b}{2} \quad \forall x \in \mathbb{R},$$

then $0 \le f \le 1 + h$ and f is strictly convex. Hence, for $\varepsilon > 0$, define $h_\varepsilon = h + \varepsilon f$. Then, $h \le h_\varepsilon \le (1+\varepsilon)h + \varepsilon$. □

Theorem 5.1. *Letting $d = 1$, suppose $c(x,y) = h(x-y)$ for some convex function $h\colon \mathbb{R} \to \mathbb{R}$ that is bounded from below. Then, $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ is an optimal transport plan. Hence,*

$$\mathbb{K}_c(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} h(x-y)\,\mathrm{d}\gamma(x,y) = \int_0^1 h(F_\mu^{-1}(u) - F_\nu^{-1}(u))\,\mathrm{d}u. \tag{5.1}$$

Proof. First, if h is a constant function, any transport plan is optimal; then, there is nothing to prove. Assuming h is not a constant function, for fixed $\varepsilon > 0$, take a strictly convex function h_ε as in Lemma 5.2. Then, we have

$$\int_{\mathbb{R}\times\mathbb{R}} h_\varepsilon(x-y)\,\mathrm{d}(F_\mu^{-1}, F_\nu^{-1})_\# \lambda(u) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} h_\varepsilon(x-y)\,\mathrm{d}\gamma(x,y).$$

This follows by applying Proposition 5.1 to the cost function $c_\varepsilon(x,y) = h_\varepsilon(x-y)$ provided the optimal transport cost given c_ε is finite; without finiteness, this is still true as both sides are simply infinite. Hence, using $h \le h_\varepsilon \le (1+\varepsilon)h + \varepsilon$,

$$\begin{aligned}
\int_0^1 h(F_\mu^{-1}(u) - F_\nu^{-1}(u))\,\mathrm{d}u &= \int_{\mathbb{R}\times\mathbb{R}} h(x-y)\,\mathrm{d}(F_\mu^{-1}, F_\nu^{-1})_\# \lambda(u) \\
&\le \int_{\mathbb{R}\times\mathbb{R}} h_\varepsilon(x-y)\,\mathrm{d}(F_\mu^{-1}, F_\nu^{-1})_\# \lambda(u) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} h_\varepsilon(x-y)\,\mathrm{d}\gamma(x,y) \\
&\le (1+\varepsilon) \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} h(x-y)\,\mathrm{d}\gamma(x,y) + \varepsilon.
\end{aligned}$$

As this is true for any $\varepsilon > 0$, we have (5.1) and $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ is an optimal transport plan. □

Recall from Proposition 1.5 that $(F_\mu^{-1}, F_\nu^{-1})_\# \lambda$ is induced by $F_\nu^{-1} \circ F_\mu$ provided F_μ is continuous. Accordingly, not only $F_\nu^{-1} \circ F_\mu$ is an optimal transport map, but also the Monge and the Kantorovich problems have the same optimal transport cost. Moreover, by Proposition 1.8, we can establish uniqueness of optimal transport maps as follows.

Corollary 5.1. *Letting $d = 1$, suppose $c(x, y) = h(x-y)$ for some convex function $h\colon \mathbb{R} \to \mathbb{R}$ that is bounded from below. If F_μ is continuous, $F_\nu^{-1} \circ F_\mu$ is an optimal transport map and thus*

$$\mathbb{M}_c(\mu,\nu) = \inf_{T \in \mathcal{T}(\mu,\nu)} \int_{\mathbb{R}} h(x - T(x))\, \mathrm{d}\mu(x) = \int_{\mathbb{R}} h(x - F_\nu^{-1} \circ F_\mu(x))\, \mathrm{d}\mu(x) = \mathbb{K}_c(\mu,\nu).$$

If h is strictly convex and the optimal transport cost is finite, $F_\nu^{-1} \circ F_\mu$ is a μ-almost everywhere unique optimal transport map.

Example 5.1. Note that $h(x) = |x|$ is convex but not strictly convex. Let $\mu = \lambda$ and ν be the Lebesgue measure supported on $[1/2, 3/2]$ so that $F_\nu^{-1} \circ F_\mu(x) = x + \frac{1}{2}$ is an optimal transport map for the cost function $c(x, y) = |x - y|$; the optimal transport cost is $1/2$. However, there are other optimal transport maps, e.g.,

$$T(x) = \begin{cases} x + 1 & \text{if } x \le \frac{1}{2}, \\ x & \text{if } x > 1. \end{cases}$$

Then, $T_\# \mu = \nu$ holds and the cost incurred by T is also $1/2$, and thus T is optimal as well. See Figure 2.1 of [San15] for a visualization.

5.2 Quadratic cost

We study the case where c is the quadratic cost, i.e., $c(x, y) = \frac{1}{2}\|x - y\|_2^2$ for all $x, y \in \mathbb{R}^d$. In this special case, all the notions in Definition 3.1 that we used to establish optimality and (semi-)duality results boil down to well-known concepts in convex analysis. Particularly, as mentioned in Remark 3.1, if c is the quadratic cost, c-cyclical monotonicity is equivalent to cyclical monotonicity.

Definition 5.1. A subset $\Pi \in \mathbb{R}^d \times \mathbb{R}^d$ is said to be cyclically monotone if

$$\sum_{i=1}^n \langle x_i, y_i \rangle \ge \sum_{i=1}^n \langle x_i, y_{\sigma(i)} \rangle$$

for any $n \in \mathbb{N}$, $(x_1, y_1), \ldots, (x_n, y_n) \in \Pi$, and any permutation $\sigma \in \mathrm{Perm}(n)$.

Next, we show that c-transform leads to the conjugate.

Definition 5.2. The conjugate of $\phi\colon \mathbb{R}^d \to [-\infty, \infty]$ is a function $\phi^*\colon \mathbb{R}^d \to [-\infty, \infty]$ defined by

$$\phi^*(y) = \sup_{x\in\mathbb{R}^d} \left(\langle x, y\rangle - \phi(x)\right).$$

For $\varphi\colon \mathbb{R}^d \to [-\infty, \infty]$, one can verify that

$$\begin{aligned} \frac{\|y\|_2^2}{2} - \varphi^c(y) &= \frac{\|y\|_2^2}{2} - \inf_{x\in\mathbb{R}^d} \left(\frac{\|x-y\|_2^2}{2} - \varphi(x)\right) \\ &= \sup_{x\in\mathbb{R}^d} \left(\langle x, y\rangle - \left(\frac{\|x\|_2^2}{2} - \varphi(x)\right)\right). \end{aligned}$$

In other words,

$$\frac{\|\cdot\|_2^2}{2} - \varphi^c = \left(\frac{\|\cdot\|_2^2}{2} - \varphi\right)^*, \tag{5.2}$$

showing that c-transform indeed leads to the conjugate. The following proposition is an analogue to Proposition 3.1, collecting some basic properties of the conjugate.

Proposition 5.2. *Fix $\phi\colon \mathbb{R}^d \to [-\infty, \infty]$.*

(i) $\phi^* \equiv \infty$ if $\phi(x) = -\infty$ for some $x \in \mathbb{R}^d$.

(ii) $\phi^* \equiv -\infty$ if $\phi \equiv \infty$.

Now, suppose $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$ and ϕ is proper.

(iii) $\phi^*\colon \mathbb{R}^d \to (-\infty, \infty]$.

(v) $\phi(x) + \phi^*(y) \geq \langle x, y\rangle$ for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$.

Remark 5.2. Note that ϕ^* in (iii) of Proposition 5.2 might not be proper, i.e., $\phi^* \equiv \infty$ can happen. For instance, consider $\phi(x) = \log|x|$.

Next, we characterize c-concave functions. Due to the connection (5.2) of c-transform and the conjugate, one can deduce that c-concavity is related to $\phi = \phi^{**}$, which turns out to be the usual convexity plus lower semi-continuity.

Definition 5.3. $\phi\colon \mathbb{R}^d \to [-\infty, \infty]$ is convex if its epigraph

$$\mathrm{epi}(\phi) := \{(x, y) \in \mathbb{R}^d \times \mathbb{R} : y \geq \phi(x)\}$$

is a convex set.

Remark 5.3. In Definition 5.3, if we consider $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$, convexity is equivalent to

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y)$$

for any $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$.

The following proposition revisits Proposition 3.2 in the context of the conjugate.

Proposition 5.3. *Fix $\phi\colon \mathbb{R}^d \to [-\infty, \infty]$.*

(i) $\phi^{**} \leq \phi$.

(ii) $\phi = \phi^{**}$ *if and only if* ϕ *is convex and lower semi-continuous; in this case, only one of the following is true:*

(1) $\phi \equiv -\infty$ *and* $\phi^* \equiv \infty$.

(2) $\phi \equiv \infty$ *and* $\phi^* \equiv -\infty$.

(3) $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$ *and* $\phi^*\colon \mathbb{R}^d \to (-\infty, \infty]$, *where both* ϕ *and* ϕ^* *are proper.*

Now, by virtue of (5.2), one can verify that $\varphi\colon \mathbb{R}^d \to [-\infty, \infty]$ is c-concave if and only if $\frac{\|\cdot\|_2^2}{2} - \varphi$ is convex and lower semi-continuous, or equivalently

$$\frac{\|\cdot\|_2^2}{2} - \varphi = \left(\frac{\|\cdot\|_2^2}{2} - \varphi\right)^{**}.$$

Lastly, we show that c-superdifferential leads to the usual subdifferential.

Definition 5.4. The subdifferential of $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$ at $x \in \mathbb{R}^d$ is defined by

$$\partial\phi(x) = \left\{y \in \mathbb{R}^d : \phi(z) \geq \phi(x) + \langle z - x, y\rangle \quad \forall z \in \mathbb{R}^d\right\}.$$

An element of $\partial\phi(x)$ is called a subgradient of ϕ at x. The subdifferential of ϕ is defined by

$$\partial\phi = \left\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : y \in \partial\phi(x)\right\}.$$

Proposition 5.4. *If* $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$ *is proper,* $\phi(x) + \phi^*(y) = \langle x, y\rangle$ *if and only if* $y \in \partial\phi(x)$. *Therefore,*

$$\partial\phi = \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \phi(x) + \phi^*(y) = \langle x, y\rangle\}.$$

Proof. Note that $\phi(z) \geq \phi(x) + \langle z - x, y\rangle$ for all $z \in \mathbb{R}^d$ if and only if

$$\langle z, y\rangle - \phi(z) \leq \langle x, y\rangle - \phi(x) \quad \forall z \in \mathbb{R}^d \quad \Leftrightarrow \quad \phi^*(y) = \langle x, y\rangle - \phi(x).$$

Hence, $y \in \partial\phi(x)$ if and only if $\phi^*(y) = \langle x, y\rangle - \phi(x)$; as this is possible only when $\phi(x) < \infty$, i.e., $x \in \mathrm{dom}(\phi)$, this is equivalent to $\phi(x) + \phi^*(y) = \langle x, y\rangle$. □

We can now verify the connection between c-superdifferential and the subdifferential. Suppose $\varphi\colon \mathbb{R}^d \to [-\infty, \infty)$ is proper and let $\phi = \frac{\|\cdot\|_2^2}{2} - \varphi$. Then, $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$ is proper, and we have

$$\begin{aligned}\partial_c \varphi &= \left\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \varphi(x) + \varphi^c(y) = \frac{\|x-y\|_2^2}{2}\right\}\\ &= \left\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \phi(x) + \phi^*(y) = \langle x, y\rangle\right\}\\ &= \partial\phi,\end{aligned}$$

where the second equality follows from (5.2) and the last equality is from Proposition 5.4.

Lastly, we present Rockafellar's result on cyclical monotonicity, which we have mentioned in Remark 3.1; though we have already proved this in Theorem 3.1, we state this again for completeness.

Theorem 5.2 (**Rockafellar**). *If $\Pi \subset \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone, there exists a proper convex function $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$ such that $\Pi \subset \partial\phi$.*

Based on the aforementioned connections, we can derive the optimality result for the quadratic cost by simply restating Theorem 3.2 using the language of convex analysis.

Theorem 5.3 (**Knott-Smith Optimality**). *Let c be the quadratic cost and suppose*

$$\int_{\mathbb{R}^d} \|x\|_2^2 \,\mathrm{d}\mu(x) < \infty \quad \textit{and} \quad \int_{\mathbb{R}^d} \|y\|_2^2 \,\mathrm{d}\nu(y) < \infty.$$

For $\gamma \in \Pi(\mu, \nu)$, the following are equivalent.

(i) *γ is an optimal transport plan.*

(ii) $\operatorname{supp}(\gamma)$ *is cyclically monotone.*

(iii) *There exists a proper convex function $\phi_o\colon \mathbb{R}^d \to (-\infty, \infty]$ such that* $\operatorname{supp}(\gamma) \subset \partial\phi_o$.

Remark 5.4. As c-concavity corresponds to convexity plus lower semi-continuity, we should have stated (iii) of Theorem 5.3 as follows: there exists a proper, convex, and lower semi-continuous function $\phi_o\colon \mathbb{R}^d \to (-\infty, \infty]$ such that $\operatorname{supp}(\gamma) \subset \partial\phi_o$. That said, as mentioned in Remark 3.5, we may replace c-concavity with measurability, meaning that we could have stated (iii) as: there exists a proper measurable function $\phi_o\colon \mathbb{R}^d \to (-\infty, \infty]$ such that $\operatorname{supp}(\gamma) \subset \partial\phi_o$. The current (iii) is valid as convexity guarantees measurability.

As pointed out in Corollary 3.1, an important consequence of Theorem 5.3 is that the support of any optimal transport plan is contained in $\partial\phi_o$.

Corollary 5.2. *Let c be the quadratic cost and suppose*

$$\int_{\mathbb{R}^d} \|x\|_2^2 \,\mathrm{d}\mu(x) < \infty \quad \textit{and} \quad \int_{\mathbb{R}^d} \|y\|_2^2 \,\mathrm{d}\nu(y) < \infty.$$

There exists a proper, convex, and lower semi-continuous function $\phi_o \colon \mathbb{R}^d \to (-\infty, \infty]$ such that $\operatorname{supp}(\gamma) \subset \partial\phi_o$ holds for any optimal transport plan $\gamma \in \Pi(\mu, \nu)$.

Now, we are ready to prove the most important result in optimal transport theory, Brenier's theorem, which states that both the Monge and the Kantorovich problems have the same optimal transport cost under the quadratic cost, with the optimal transport plan being induced by the optimal transport map that is given as the gradient of a convex function. One can already notice that this convex function has to be ϕ_o that appears in (iii) of Theorem 5.3. It turns out that such a convex function ϕ_o must be differentiable almost everywhere, which makes its $\partial\phi_o$ almost the same as the graph of its gradient. Lemma 5.3 formally states this with the help of the following proposition on the differentiability of convex functions known as Rademacher's theorem.

Proposition 5.5. *Let $\phi \colon \mathbb{R}^d \to (-\infty, \infty]$ be a proper convex function. Then, ϕ is m_d-almost everywhere differentiable on $\operatorname{int}(\operatorname{dom}(\phi))$, i.e., we can find a Borel set $D_\phi \subset \operatorname{int}(\operatorname{dom}(\phi))$ such that ϕ is differentiable on D_ϕ and $m_d(\operatorname{int}(\operatorname{dom}(\phi))\backslash D_\phi) = 0$; also, $\partial\phi(x) = \{\nabla\phi(x)\}$ for $x \in D_\phi$.*

Lemma 5.3. *Suppose $\gamma \in \Pi(\mu, \nu)$ satisfies $\operatorname{supp}(\gamma) \subset \partial\phi$ for some proper convex function $\phi \colon \mathbb{R}^d \to (-\infty, \infty]$. If μ is absolutely continuous with respect to the Lebesgue measure m_d, then ϕ is μ-almost everywhere differentiable and $\gamma = (\mathrm{Id}, \nabla\phi)_{\#}\mu$.*

Proof. As $\operatorname{supp}(\gamma) \subset \partial\phi \subset \operatorname{dom}(\phi) \times \mathbb{R}^d$,

$$\mu(\operatorname{dom}(\phi)) = \gamma(\operatorname{dom}(\phi) \times \mathbb{R}^d) = 1.$$

Also, the boundary of $\operatorname{dom}(\phi)$ is m_d-negligible and thus is μ-negligible, which means that $\mu(\operatorname{int}(\operatorname{dom}(\phi))) = 1$. Let D_ϕ be the set of points where ϕ is differentiable. Then, Proposition 5.5 shows that $m_d(\operatorname{int}(\operatorname{dom}(\phi))\backslash D_\phi) = 0$; as μ is absolutely continuous with respect to m_d, we have $\mu(\operatorname{int}(\operatorname{dom}(\phi))\backslash D_\phi) = 0$, which means $\mu(D_\phi) = 1$. This proves that ϕ is μ-almost everywhere differentiable. Now, recall that $\operatorname{supp}(\gamma) \subset \partial\phi$ implies γ is concentrated on $\partial\phi$. Notice that

$$\partial\phi = \underbrace{(\partial\phi \cap (D_\phi \times \mathbb{R}^d))}_{A} \;\cup\; \underbrace{(\partial\phi \cap ((\mathbb{R}^d\backslash D_\phi) \times \mathbb{R}^d))}_{\partial\phi\backslash A}.$$

As $\gamma(\partial\phi\backslash A) \leq \gamma((\mathbb{R}^d\backslash D_\phi) \times \mathbb{R}^d) = \mu(\mathbb{R}^d\backslash D_\phi) = 0$, we can see that γ is concentrated on A. Meanwhile, as $\partial\phi(x) = \{\nabla\phi(x)\}$ for any $x \in D_\phi$, we have $A = \{(x, \nabla\phi(x)) : x \in D_\phi\}$.

By letting $\nabla\phi = 0$ on $\mathbb{R}^d \backslash D_\phi$, we have a well-defined measurable map $\nabla\phi\colon \mathbb{R}^d \to \mathbb{R}^d$ whose graph satisfies

$$\operatorname{graph}(\nabla\phi) = A \cup \{(x, 0) : x \in \mathbb{R}^d \backslash D_\phi\} = A \cup ((\mathbb{R}^d \backslash D_\phi) \times \{0\}).$$

Since

$$\gamma((\mathbb{R}^d \backslash D_\phi) \times \{0\}) \leq \gamma((\mathbb{R}^d \backslash D_\phi) \times \mathbb{R}^d) = \mu(\mathbb{R}^d \backslash D_\phi) = 0,$$

we conclude that γ is concentrated on $\operatorname{graph}(\nabla\phi)$ and thus $\gamma = (\mathrm{Id}, \nabla\phi)_{\#}\mu$ by (ii) of Proposition 1.3. □

Theorem 5.4 (Brenier). *Let c be the quadratic cost and suppose*

$$\int_{\mathbb{R}^d} \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} \|y\|_2^2 \, \mathrm{d}\nu(y) < \infty.$$

If μ is absolutely continuous with respect to the Lebesgue measure m_d, there exists a unique optimal transport plan γ, which satisfies $\gamma = (\mathrm{Id}, \nabla\phi_o)_{\#}\mu$ for some proper convex lower semi-continuous function $\phi\colon \mathbb{R}^d \to (-\infty, \infty]$ that is μ-almost everywhere differentiable. Moreover, $\nabla\phi_o$ is a μ-almost everywhere unique optimal transport map.

Proof. As c is continuous, optimal transport plans exist by Theorem 2.3. By Corollary 5.2, we can find a proper convex lower semi-continuous function $\phi_o\colon \mathbb{R}^d \to (-\infty, \infty]$ such that the support of any optimal transport plan is contained in $\partial\phi_o$. By Lemma 5.3, we conclude that ϕ_o is μ-almost everywhere differentiable and $(\mathrm{Id}, \nabla\phi_o)_{\#}\mu$ is the unique optimal transport plan. By Proposition 1.8, $\nabla\phi_o$ is a μ-almost everywhere unique optimal transport map. □

An immediate consequence of Brenier's theorem is that when both μ, ν are absolutely continuous with respect to the Lebesgue measure, the optimal transport map from μ to ν and the optimal transport map from ν to μ are inverses of each other almost everywhere.

Corollary 5.3. *Let c be the quadratic cost and suppose*

$$\int_{\mathbb{R}^d} \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} \|y\|_2^2 \, \mathrm{d}\nu(y) < \infty.$$

Suppose both μ and ν are absolutely continuous with respect to the Lebesgue measure m_d. Let T_μ^ν and T_ν^μ be optimal transport maps from μ to ν and from ν to μ, respectively. Then, $T_\nu^\mu \circ T_\mu^\nu = \mathrm{Id}$ holds μ-almost everywhere and $T_\mu^\nu \circ T_\nu^\mu = \mathrm{Id}$ holds ν-almost everywhere.

Proof. By Theorem 5.4, $(\mathrm{Id}, T_\mu^\nu)_{\#}\mu \in \Pi(\mu, \nu)$ is the unique optimal transport plan incurring the optimal cost $\mathbb{K}_c(\mu, \nu)$. Similarly, $(\mathrm{Id}, T_\nu^\mu)_{\#}\nu \in \Pi(\nu, \mu)$ is the unique optimal transport plan incurring the optimal cost $\mathbb{K}_c(\nu, \mu)$. Due to symmetry, $\mathbb{K}_c(\mu, \nu) = \mathbb{K}_c(\nu, \mu)$, which implies that $(T_\nu^\mu, \mathrm{Id})_\# \nu \in \Pi(\mu, \nu)$ must be an optimal transport plan. Then, uniqueness implies $\gamma := (\mathrm{Id}, T_\mu^\nu)_\# \mu = (\mathrm{Id}, T_\nu^\mu)_\# \nu$. Therefore, for any $F \in L^1(\gamma)$,

$$\int_{\mathbb{R}^d} F(x, T_\mu^\nu(x)) \,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d \times \mathbb{R}^d} F(x, y) \,\mathrm{d}\gamma(x, y) = \int_{\mathbb{R}^d} F(T_\nu^\mu(y), y) \,\mathrm{d}\nu(y).$$

Let $F(x, y) = \|x - T_\nu^\mu(y)\|_2$, then

$$\int_{\mathbb{R}^d} \|x - T_\nu^\mu \circ T_\mu^\nu(x)\|_2 \,\mathrm{d}\mu(x) = 0,$$

hence $T_\nu^\mu \circ T_\mu^\nu = \mathrm{Id}$ holds μ-almost everywhere. Similarly, $T_\mu^\nu \circ T_\nu^\mu = \mathrm{Id}$ holds ν-almost everywhere. □

A useful way to apply Brenier's theorem to find the optimal transport map is as follows: if we find any transport map that happens to be the gradient of a convex function, then it must be the optimal transport map. Corollary 5.4 states this formally.

Corollary 5.4. *Let c be the quadratic cost and suppose*

$$\int_{\mathbb{R}^d} \|x\|_2^2 \,\mathrm{d}\mu(x) < \infty \quad \textit{and} \quad \int_{\mathbb{R}^d} \|y\|_2^2 \,\mathrm{d}\nu(y) < \infty.$$

If μ is absolutely continuous with respect to the Lebesgue measure m_d, any transport map $T \in \mathcal{T}(\mu, \nu)$ such that $T = \nabla\phi$ holds μ-almost everywhere for some proper convex lower semi-continuous function $\phi \colon \mathbb{R}^d \to (-\infty, \infty]$ satisfying $\phi \in L^1(\mu)$ is an optimal transport map.

Proof. Let $\gamma = (\mathrm{Id}, T)_\# \mu$ which is a transport map by Proposition 1.1. We first prove $\phi^* \in L^1(\nu)$. Note that $\phi^*(\nabla\phi(x)) = \langle x, \nabla\phi(x)\rangle - \phi(x)$ for $x \in D_\phi$.

$$\begin{aligned}
\int_{\mathbb{R}^d} |\langle x, \nabla\phi(x)\rangle| \,\mathrm{d}\mu(x) &\le \int_{\mathbb{R}^d} \frac{\|x\|_2^2}{2} \,\mathrm{d}\mu(x) + \int_{\mathbb{R}^d} \frac{\|\nabla\phi(x)\|_2^2}{2} \,\mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \frac{\|x\|_2^2}{2} \,\mathrm{d}\mu(x) + \int_{\mathbb{R}^d} \frac{\|x\|_2^2}{2} \,\mathrm{d}\nu(x) < \infty.
\end{aligned}$$

Therefore, $x \mapsto \phi^*(\nabla\phi(x))$ is in $L^1(\mu)$, and thus

$$\int_{\mathbb{R}^d} |\phi^*(\nabla\phi(x))| \,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} |\phi^*(x)| \,\mathrm{d}\nu(x) < \infty.$$

Next, we prove that $\mathrm{supp}(\gamma) \subset \partial\phi$;this will imply that γ is optimal by (iii) of Theorem 5.3, and thus T is an optimal transport map. Note that it suffices to prove $\gamma(\partial\phi) = 1$; if this is true, closedness of $\partial\phi$ implies $\mathrm{supp}(\gamma) \subset \partial\phi$. As in the proof of Lemma 5.3, let $A = \{(x, \nabla\phi(x)) : x \in D_\phi\}$. Then, $\gamma(\partial\phi \backslash A) = 0$, and hence $\gamma(\partial\phi) = \gamma(A)$. Also, recall that

$$\gamma(\mathrm{graph}(\nabla\phi) \backslash A) = \gamma((\mathbb{R}^d \backslash D_\phi) \times \{0\}) = 0,$$

hence $\gamma(\mathrm{graph}(\nabla\phi)) = \gamma(A)$. Meanwhile, for $\Delta := \{x \in \mathbb{R}^d : T(x) = \nabla\phi(x)\}$, we have $\mu(\Delta) = 1$, which implies

$$\gamma(\mathrm{graph}(T) \cap \mathrm{graph}(\nabla\phi)) = \gamma\{(x, T(x)) : x \in \Delta\} \geq \mu(\Delta) = 1.$$

Lastly, $\gamma(\mathrm{graph}(T)) = 1$ by construction, which implies $\gamma(\mathrm{graph}(\nabla\phi)) = 1$. In summary,

$$\gamma(\partial\phi) = \gamma(A) = \gamma(\mathrm{graph}(\nabla\phi)) = 1.$$

□

Example 5.2 (Gaussian Distributions). Let μ and ν be the Gaussian distributions $N(\theta_1, \Sigma_1)$ and $N(\theta_2, \Sigma_2)$, respectively. We first consider linear transport maps, that is, $T(x) = Ax + b$ for $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$ such that $A\theta_1 + b = \theta_2$ and $A\Sigma_1 A^\top = \Sigma_2$. Now, let us find a linear transport map that incurs the smallest transport cost, which can be formulated as

$$\min_{\substack{A\in\mathbb{R}^{d\times d}\\ A\Sigma_1A^\top=\Sigma_2}} \underbrace{\int_{\mathbb{R}^d} \|A(x-\theta_1)+\theta_2-x\|_2^2 \,\mathrm{d}\mu(x)}_{=:Q(A)}.$$

Then, one can verify that

$$Q(A) = \mathrm{tr}(A\Sigma_1A^\top) + \|\theta_2\|_2^2 + \mathrm{tr}(\Sigma_1) + \|\theta_1\|_2^2 - 2(\mathrm{tr}(A\Sigma_1) + \langle\theta_1, \theta_2\rangle).$$

Therefore, the problem is equivalent to

$$\max_{\substack{A\in\mathbb{R}^{d\times d}\\ A\Sigma_1A^\top=\Sigma_2}} \mathrm{tr}(A\Sigma_1).$$

Assuming Σ_1 is invertible, the above problem is equivalent to

$$\max_{\substack{B\in\mathbb{R}^{d\times d}\\ BB^\top=\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}}} \mathrm{tr}(B),$$

where we change the variable by letting $B = \Sigma_1^{1/2}A\Sigma_1^{1/2}$. Using the spectral decomposition, one can verify that the maximum is attained by $B = (\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}$. Therefore,

$$x \mapsto \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}(x-\theta_1) + \theta_2$$

is a linear transport map that incurs the smallest transport cost, where the transport cost is

$$Q(\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}) = \|\theta_1-\theta_2\|_2^2 + \mathrm{tr}\left(\Sigma_1+\Sigma_2-2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right).$$

Importantly, this linear map is the gradient of the following convex quadratic function:

$$\phi(x) := \frac{1}{2}\langle x-\theta_1, \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}(x-\theta_1)\rangle + \langle\theta_2, x\rangle,$$

where the convexity of ϕ follows as $\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}$ is positive semidefinite. By Corollary 5.4, we conclude that this linear transport map is in fact the optimal transport map. By symmetry, if Σ_2 is invertible,

$$x \mapsto \Sigma_2^{-1/2}(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\Sigma_2^{-1/2}(x-\theta_2)+\theta_1$$

is the optimal transport map from ν to μ. Its inverse map

$$x \mapsto \Sigma_2^{1/2}(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{-1/2}\Sigma_2^{1/2}(x-\theta_1)+\theta_2$$

is the optimal transport map from μ to ν, where one can verify

$$\Sigma_2^{1/2}(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{-1/2}\Sigma_2^{1/2} = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}.$$

Now, we revisit Corollaries 3.1 and 3.2.

Corollary 5.5. *Let c be the quadratic cost and suppose*

$$M := \int_{\mathbb{R}^d} \frac{\|x\|_2^2}{2}\,\mathrm{d}\mu(x) + \int_{\mathbb{R}^d} \frac{\|y\|_2^2}{2}\,\mathrm{d}\nu(y) < \infty.$$

(i) There exists a proper, convex, and lower semi-continuous function $\phi_o\colon \mathbb{R}^d \to (-\infty,\infty]$ such that $(\phi_o,\phi_o^*) \in L^1(\mu)\times L^1(\nu)$ and

$$\inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{\|x-y\|_2^2}{2}\,\mathrm{d}\gamma(x,y) = M - \int_{\mathbb{R}^d}\phi_o\,\mathrm{d}\mu - \int_{\mathbb{R}^d}\phi_o^*\,\mathrm{d}\nu.$$

Also, $\operatorname{supp}(\gamma) \subset \partial\phi_o$ holds for any optimal transport plan $\gamma \in \Pi(\mu,\nu)$.

(ii) The following semi-duality holds:

$$\sup_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle x,y\rangle\,\mathrm{d}\gamma(x,y) = \inf_{\phi\in\mathcal{S}} \left(\int_{\mathcal{X}}\phi\,\mathrm{d}\mu + \int_{\mathcal{Y}}\phi^*\,\mathrm{d}\nu\right), \tag{5.3}$$

where $\mathcal{S}$ is the collection of all proper convex and lower semi-continuous functions $\phi\colon \mathbb{R}^d \to (-\infty,\infty]$ such that $(\phi,\phi^*) \in L^1(\mu)\times L^1(\nu)$.

(iii) The following duality holds:

$$\sup_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle x,y\rangle\,\mathrm{d}\gamma(x,y) = \inf_{(\phi,\psi)\in\mathcal{D}} \left(\int_{\mathcal{X}}\phi\,\mathrm{d}\mu + \int_{\mathcal{Y}}\psi\,\mathrm{d}\nu\right), \tag{5.4}$$

$\mathcal{D}$ is the collection of all pairs $(\phi,\psi) \in L^1(\mu)\times L^1(\nu)$ such that $\phi(x)+\psi(y) \geq \langle x,y\rangle$ for all $(x,y) \in \mathbb{R}^d\times\mathbb{R}^d$.

In particular, ϕ_o of (i) satisfies $\phi_o \in \mathcal{S}$ and $(\phi_o,\phi_o^*) \in \mathcal{D}$ which are the minimizers of the right-hand sides of (5.3) and (5.4), respectively.

Now, it should be clear that if μ is absolutely continuous with respect to the Lebesgue measure, then ϕ_o in Corollary 5.5 is the same as ϕ_o in Theorem 5.4.

6 Wasserstein Distance

One of the most important results of optimal transport theory is that optimal transport cost between two probability measures defines a distance. More concretely, given a separable metric space $(\mathcal{X}, \rho)$, we consider the Kantorovich problem between $\mu, \nu \in \mathscr{P}(\mathcal{X})$, where the cost function is $c(x, y) = \rho(x, y)^p$ for some fixed exponent $p \in [1, \infty)$. The resulting minimum of the Kantorovich problem gives rise to the Wasserstein distance. This section rigorously derives metric and topological properties of the Wasserstein distance.

Settings Unless otherwise stated, $(\mathcal{X}, \rho)$ is a separable metric space and $p \in [1, \infty)$ is a fixed exponent.

6.1 Basic properties

Definition 6.1. The Wasserstein distance of order p between $\mu, \nu \in \mathscr{P}(\mathcal{X})$ is defined by

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \rho(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}. \tag{6.1}$$

Remark 6.1 (Geometric Interpretation). As $\Pi(\delta_x, \delta_y) = \{\delta_{(x,y)}\}$, one can verify that $W_p(\delta_x, \delta_y) = \rho(x, y)$ for any $x, y \in \mathcal{X}$. Roughly speaking, this implies that W_p measures a distance between two elements of $\mathscr{P}(\mathcal{X})$ by taking into account the distance between their supports under the ground metric ρ. In other words, W_p utilizes a metric structure of the underlying space $\mathcal{X}$ to define a distance on $\mathscr{P}(\mathcal{X})$. Though this seems very natural, other distances or divergences on $\mathscr{P}(\mathcal{X})$ lack such a geometry perspective.

In practice, we mostly focus on the Euclidean space case where $\mathcal{X} = \mathbb{R}^d$ and ρ is the Euclidean distance.

Example 6.1. Recall that $h(x) = |x|^p$ is a convex function on $\mathbb{R}$ for any $p \in [1, \infty)$. Hence, Theorem 5.1 shows that

$$W_p(\mu, \nu) = \left(\int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \, \mathrm{d}u \right)^{1/p} \quad \forall \mu, \nu \in \mathscr{P}(\mathbb{R}), \tag{6.2}$$

where F_μ and F_ν denote the distribution functions of μ and ν, respectively. Using Fubini's theorem, one can verify that

$$W_1(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)| \, \mathrm{d}x.$$

In Example 6.1, note that W_p can be infinite as mentioned in the proof of Theorem 5.1. One way to ensure finiteness of W_p is to utilize a moment condition, for instance, (6.2) becomes finite provided

$$\int_0^1 |F_\mu^{-1}(u)|^p \,\mathrm{d}u = \int_{\mathbb{R}} |x|^p \,\mathrm{d}\mu(x) < \infty \quad \text{and} \quad \int_0^1 |F_\nu^{-1}(u)|^p \,\mathrm{d}u = \int_{\mathbb{R}} |x|^p \,\mathrm{d}\nu(x) < \infty.$$

In other words, if both μ and ν have finite p-th moments, $W_p(\mu,\nu) < \infty$ is guaranteed. We can extend such a moment condition to the general case as follows.

Definition 6.2. Define $\mathscr{P}_p(\mathcal{X})$ as a subset of $\mathscr{P}(\mathcal{X})$ having finite p-th moments, that is,

$$\mathscr{P}_p(\mathcal{X}) = \left\{ \mu \in \mathscr{P}(\mathcal{X}) : \int_{\mathcal{X}} \rho(x, x_0)^p \,\mathrm{d}\mu(x) < \infty \right\}.$$

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

Remark 6.2. In Definition 6.2, one can verify that $\mathscr{P}_p(\mathcal{X})$ is independent of the choice of x_0 using the triangle inequality. Also, note that

$$\int_{\mathcal{X}} \rho(x, x_0)^p = W_p^p(\mu, \delta_{x_0}).$$

Moreover, one can verify that $\mathscr{P}_p(\mathcal{X}) \supset \mathscr{P}_q(\mathcal{X})$ provided $p \le q$. In other words, having some particular moment implies existence of all the lower moments.

We have an ordering of Wasserstein distances as follows.

Proposition 6.1. *Let $p, q \in [1, \infty)$ be two exponents such that $p \le q$. Then,*

(i) $W_p \le W_q$,

(ii) $W_q \le W_p^{p/q} \mathrm{diam}(\mathcal{X})^{1-p/q}$.

Remark 6.3. Recall from Theorem 4.2 that if $\mu, \nu \in \mathscr{P}_1(\mathcal{X})$ are tight,

$$W_1(\mu,\nu) = \sup_{\substack{\varphi\colon \mathcal{X}\to\mathbb{R} \\ \|\varphi\|_{\mathrm{Lip}}\le 1}} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu - \int_{\mathcal{X}} \varphi \,\mathrm{d}\nu \right), \tag{6.3}$$

where we define for any $\varphi\colon \mathcal{X} \to \mathbb{R}$,

$$\|\varphi\|_{\mathrm{Lip}} := \sup_{x\neq y} \frac{|\varphi(x) - \varphi(y)|}{\rho(x,y)}.$$

In short, W_1 is the supremum of $\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu - \int_{\mathcal{X}} \varphi \,\mathrm{d}\nu$ over all 1-Lipschitz functions φ on $\mathcal{X}$.

6.2 Metric properties

We derive metric properties of the Wasserstein distance. First, note that W_p is symmetric by definition. Next, we show that $\mu = \nu$ if and only if $W_p(\mu,\nu) = 0$. The "if" part requires tightness of μ, ν.

Proposition 6.2. *$W_p(\mu,\mu) = 0$ for any $\mu \in \mathscr{P}(\mathcal{X})$. If $\mu, \nu \in \mathscr{P}(\mathcal{X})$ are tight, $W_p(\mu,\nu) = 0$ implies $\mu = \nu$.*

Proof. $W_p(\mu,\mu) = 0$ holds because $(\mathrm{Id},\mathrm{Id})_\#\mu \in \Pi(\mu,\mu)$ by Proposition 1.4, which gives

$$W_p^p(\mu,\mu) \le \int_{\mathcal{X}\times\mathcal{X}} \rho(x,y)^p \,\mathrm{d}(\mathrm{Id},\mathrm{Id})_\#\mu(x,y) = \int_{\mathcal{X}} \rho(x,x)^p \,\mathrm{d}\mu(x) = 0.$$

If $\mu, \nu \in \mathscr{P}(\mathcal{X})$ are tight, we can find an optimal transport plan $\gamma \in \Pi(\mu,\nu)$ with respect to the cost ρ^p by Theorem 2.3, that is, γ satisfies

$$\int_{\mathcal{X}\times\mathcal{X}} \rho^p \,\mathrm{d}\gamma = W_p^p(\mu,\nu) = 0.$$

Hence, γ is concentrated on the graph of $\mathrm{Id}\colon \mathcal{X} \to \mathcal{X}$. By Proposition 1.3, we conclude $\gamma = (\mathrm{Id},\mathrm{Id})_\#\mu$, and hence $\nu = \mathrm{Id}_\#\mu = \mu$. □

Remark 6.4. For any $\mu \in \mathscr{P}(\mathcal{X})$, one can verify that $(\mathrm{Id},\mathrm{Id})_\#\mu$ is the unique optimal transport plan from μ to itself for a cost function ρ^p. To see this, suppose $\gamma \in \Pi(\mu,\mu)$ is an optimal transport plan. As the optimal cost is $W_p^p(\mu,\mu) = 0$, we can see that γ is concentrated on $\{(x,y) \in \mathcal{X}\times\mathcal{X} : x = y\}$, the graph of a transport map $\mathrm{Id}\colon \mathcal{X} \to \mathcal{X}$ from μ to itself. By Proposition 1.3, we conclude $\gamma = (\mathrm{Id},\mathrm{Id})_\#\mu$.

Next, we prove the triangle inequality. Given three elements $\mu_1, \mu_2, \mu_3 \in \mathscr{P}(\mathcal{X})$, the main idea is to invoke the gluing technique (Lemma 1.3) to construct a probability measure $\Gamma \in \mathscr{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{X})$ such that

(i) the three marginal measures of Γ are μ_1, μ_2, μ_3 in turn,

(ii) $(P_{12})_\#\Gamma$ is an optimal transport plan from μ_1 to μ_2,

(iii) $(P_{23})_\#\Gamma$ is an optimal transport plan from μ_2 to μ_3,

where $P_{ij}(x_1,x_2,x_3) = (x_i,x_j)$ for all $i,j \in \{1,2,3\}$ such that $i \neq j$ and all $x_1,x_2,x_3 \in \mathcal{X}$. Then, the proof follows thanks to the Minkowski inequality.

Proposition 6.3. *If $(\mathcal{X},\rho)$ is a complete separable metric space,*

$$W_p(\mu_1,\mu_3) \le W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3) \quad \forall \mu_1,\mu_2,\mu_3 \in \mathscr{B}(\mathcal{X}).$$

Proof. Due to tightness, we can find $\gamma_{12} \in \Pi(\mu_1, \mu_2)$ and $\gamma_{23} \in \Pi(\mu_2, \mu_3)$ such that

$$W_p^p(\mu_1, \mu_2) = \int_{\mathcal{X}\times\mathcal{X}} \rho^p \, \mathrm{d}\gamma_{12} \quad \text{and} \quad W_p^p(\mu_2, \mu_3) = \int_{\mathcal{X}\times\mathcal{X}} \rho^p \, \mathrm{d}\gamma_{23}.$$

Using Lemma 1.3, we can find $\Gamma \in \mathscr{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{X})$ such that $\gamma_{12} = (P_{12})_\# \Gamma$ and $\gamma_{23} = (P_{23})_\# \Gamma$; note that this implies $\gamma_{13} := (P_{13})_\# \Gamma \in \Pi(\mu_1, \mu_3)$. Then,

$$\begin{aligned}
W_p(\mu_1, \mu_3) &\leq \left(\int_{\mathcal{X}\times\mathcal{X}} \rho(x_1, x_3)^p \, \mathrm{d}\gamma_{13} \right)^{1/p} \\
&= \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} \rho(x_1, x_3)^p \, \mathrm{d}\gamma \right)^{1/p} \\
&\leq \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} (\rho(x_1, x_2) + \rho(x_2, x_3))^p \, \mathrm{d}\gamma \right)^{1/p} \\
&\leq \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} \rho(x_1, x_2)^p \, \mathrm{d}\gamma \right)^{1/p} + \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} \rho(x_2, x_3)^p \, \mathrm{d}\gamma \right)^{1/p} \\
&= \left(\int_{\mathcal{X}\times\mathcal{X}} \rho(x_1, x_2)^p \, \mathrm{d}\gamma_{12} \right)^{1/p} + \left(\int_{\mathcal{X}\times\mathcal{X}} \rho(x_2, x_3)^p \, \mathrm{d}\gamma_{23} \right)^{1/p} \\
&= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3).
\end{aligned}$$

□

Therefore, if $(\mathcal{X}, \rho)$ is a complete separable metric space, W_p satisfies the triangle inequality; also, since all elements of $\mathscr{P}(\mathcal{X})$ are tight (Remark 2.4), Proposition 6.2 holds. Accordingly, by restricting W_p to $\mathscr{P}_p(\mathcal{X})$, we conclude that W_p is indeed a metric.

Corollary 6.1. *If $(\mathcal{X}, \rho)$ is a complete separable metric space, W_p is a metric on $\mathscr{P}_p(\mathcal{X})$. We call the metric space $(\mathscr{P}_p(\mathcal{X}), W_p)$ the Wasserstein space of order p.*

Example 6.2. Suppose $\mathcal{X} = \mathbb{R}^d$ is equipped with the standard Euclidean distance. As in Example 5.2, let μ and ν be the Gaussian distributions $N(\theta_1, \Sigma_1)$ and $N(\theta_2, \Sigma_2)$, respectively, where we assume Σ_1 is invertible. We have seen in Example 5.2 that

$$T(x) = \Sigma_1^{-1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \Sigma_1^{-1/2} (x - \theta_1) + \theta_2 \quad \forall x \in \mathbb{R}^d$$

is the unique optimal transport map under the quadratic cost. Therefore,

$$W_2^2(\mu, \nu) = \int_{\mathbb{R}^d} |T(x) - x|^2 \, \mathrm{d}\mu(x) = \|\theta_1 - \theta_2\|_2^2 + \underbrace{\operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \right)}_{=\beta(\Sigma_1, \Sigma_2)^2},$$

where β is often called the Bures metric. One can show that

$$\beta(\Sigma_1, \Sigma_2)^2 \leq \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_F^2,$$

where $\|\cdot\|_F$ denote the Frobenius norm. To see this, it suffices to observe that

$$\mathrm{tr}(\Sigma_1^{1/2}\Sigma_2^{1/2}) \le \mathrm{tr}((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}).$$

In fact, letting $S(x) = \Sigma_2^{1/2}\Sigma_1^{-1/2}(x-\theta_1)+\theta_2$, we can verify that S is a transport map from μ to ν and its transport cost satisfies

$$\int_{\mathbb{R}^d} |S(x)-x|^2 \,\mathrm{d}\mu(x) = \|\theta_1-\theta_2\|_2^2 + \|\Sigma_1^{1/2}-\Sigma_2^{1/2}\|_F^2.$$

6.3 Topological properties of the Wasserstein space

Now, we have two topologies on the space $\mathscr{P}_p(\mathcal{X})$: the usual weak inherited from $\mathscr{P}(\mathcal{X})$ and the Wasserstein topology induced by W_p. We show that the Wasserstein topology is stronger than the weak topology, namely, the Wasserstein topology is finer than the weak topology. It turns out that the convergence in the Wasserstein topology implies the weak convergence plus the convergence of moments.

Theorem 6.1. *If $(\mathcal{X},\rho)$ is a complete separable metric space, the following are equivalent for a sequence $(\mu_n)_{n\in\mathbb{N}}$ and μ in $\mathscr{P}_p(\mathcal{X})$:*

(i) $\lim_{n\to\infty} W_p(\mu_n,\mu)=0$.

(ii) $(\mu_n)_{n\in\mathbb{N}}$ *converges weakly to* μ *and*

$$\lim_{n\to\infty}\int \rho(x,x_0)^p \,\mathrm{d}\mu_n = \int \rho(x,x_0)^p\,\mathrm{d}\mu \quad \forall x_0\in\mathcal{X}. \tag{6.4}$$

(iii) $(\mu_n)_{n\in\mathbb{N}}$ *converges weakly to* μ *and*

$$\lim_{R\to\infty}\limsup_{n\to\infty}\int_{\rho(x,x_0)\ge R}\rho(x,x_0)^p\,\mathrm{d}\mu_n = 0 \quad \forall x_0\in\mathcal{X}. \tag{6.5}$$

Proof. Suppose (i) holds. Then, $\lim_{n\to\infty} W_1(\mu_n,\mu) = 0$ as well by Proposition 6.1. As discussed in Remark 6.3, for any 1-Lipschitz function φ on $\mathcal{X}$,

$$\left|\int_{\mathcal{X}}\varphi\,\mathrm{d}\mu_n - \int_{\mathcal{X}}\varphi\,\mathrm{d}\mu\right| \le W_1(\mu_n,\mu) \quad \forall n\in\mathbb{N},$$

which implies $\lim_{n\to\infty}\int_{\mathcal{X}}\varphi\,\mathrm{d}\mu_n = \int_{\mathcal{X}}\varphi\,\mathrm{d}\mu$. One can verify that this must hold for any bounded Lipschitz φ, which implies that $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ by Theorem 2.1. Also, for any $x_0\in\mathcal{X}$, note that

$$\lim_{n\to\infty}\int_{\mathcal{X}}\rho(x,x_0)^p\,\mathrm{d}\mu_n(x) = \lim_{n\to\infty} W_p(\mu_n,\delta_{x_0})^p = W_p(\mu,\delta_{x_0})^p = \int_{\mathcal{X}}\rho(x,x_0)^p\,\mathrm{d}\mu.$$

Suppose (ii) holds. For any $n \in \mathbb{N}$ and $R > 0$, define

$$M_n = \int_{\mathcal{X}} \rho(x, x_0)^p \, \mathrm{d}\mu_n(x) \quad \text{and} \quad M_{n,R} = \int_{\mathcal{X}} \rho(x, x_0)^p \wedge R^p \, \mathrm{d}\mu_n(x).$$

For any $R > 0$, weak convergence of $(\mu_n)_{n\in\mathbb{N}}$ to μ implies

$$\lim_{n\to\infty} M_{n,R} = \int_{\mathcal{X}} \rho(x, x_0)^p \wedge R^p \, \mathrm{d}\mu(x) =: M_R.$$

Also, letting $C_R = \{x \in \mathcal{X} : \rho(x, x_0) \geq R\}$,

$$\int_{\rho(x,x_0)\geq R} \rho(x, x_0)^p \, \mathrm{d}\mu_n(x) = M_n - M_{n,R} + R^p \mu_n(C_R).$$

Therefore, using (6.4),

$$\begin{aligned}
\limsup_{n\to\infty} \int_{\rho(x,x_0)\geq R} \rho(x, x_0)^p \, \mathrm{d}\mu_n(x) &= \int_{\mathcal{X}} \rho(x, x_0)^p \, \mathrm{d}\mu(x) - M_R + R^p \limsup_{n\to\infty} \mu_n(C_R) \\
&\leq \int_{\mathcal{X}} \rho(x, x_0)^p \, \mathrm{d}\mu(x) - M_R + R^p \mu(C_R) \\
&= \int_{\rho(x,x_0)\geq R} \rho(x, x_0)^p \, \mathrm{d}\mu(x),
\end{aligned}$$

where the inequality is due to Theorem 2.1 as C_R is a closed set. Note that the dominated convergence theorem implies

$$\lim_{R\to\infty} \int_{\rho(x,x_0)\geq R} \rho(x, x_0)^p \, \mathrm{d}\mu(x) = 0,$$

hence (6.5) holds. Suppose (iii) holds. For each $n \in \mathbb{N}$, let γ_n be an optimal transport plan between μ_n and μ. Then, $(\gamma_n)_{n\in\mathbb{N}}$ converges weakly to $\gamma = (\mathrm{Id}, \mathrm{Id})_{\#}\mu$, the unique optimal transport plan from μ to itself for a cost function ρ^p by Theorem 4.1. Now, fix $x_0 \in \mathcal{X}$ and divide $\mathcal{X} \times \mathcal{X}$ into three regions:

$$\begin{aligned}
S_1 &= \{(x, y) \in \mathcal{X} \times \mathcal{X} : \rho(x, y) < R\}, \\
S_2 &= \{(x, y) \in \mathcal{X} \times \mathcal{X} : \rho(x, y) \geq R \quad \text{and} \quad \rho(x, x_0) \geq \rho(y, x_0)\}, \\
S_3 &= \{(x, y) \in \mathcal{X} \times \mathcal{X} : \rho(x, y) \geq R \quad \text{and} \quad \rho(x, x_0) < \rho(y, x_0)\}.
\end{aligned}$$

Note that $\rho = \rho \wedge R$ on S_1. Also, $\rho(x, y) \leq 2\rho(x, x_0)$ and $\rho(x, y) \leq 2\rho(y, x_0)$ on S_2 and S_3 respectively. Hence,

$$\begin{aligned}
\rho(x, y) &\leq 2\rho(x, x_0) 1_{\{\rho(x,x_0)\geq R/2\}} \quad \forall (x, y) \in S_2, \\
\rho(x, y) &\leq 2\rho(y, x_0) 1_{\{\rho(y,x_0)\geq R/2\}} \quad \forall (x, y) \in S_3.
\end{aligned}$$

Therefore, for each $n \in \mathbb{N}$.

$$\int_{S_2} \rho^p \,\mathrm{d}\gamma_n \leq 2^p \int_{\mathcal{X}\times\mathcal{X}} \rho(x,x_0)^p 1_{\{\rho(x,x_0)\geq R/2\}} \,\mathrm{d}\gamma_n(x,y) = 2^p \int_{\rho(x,x_0)\geq R/2} \rho(x,x_0)^p \,\mathrm{d}\mu_n(x),$$
$$\int_{S_3} \rho^p \,\mathrm{d}\gamma_n \leq 2^p \int_{\mathcal{X}\times\mathcal{X}} \rho(y,x_0)^p 1_{\{\rho(y,x_0)\geq R/2\}} \,\mathrm{d}\gamma_n(x,y) = 2^p \int_{\rho(y,x_0)\geq R/2} \rho(y,x_0)^p \,\mathrm{d}\mu(y),$$

hence letting $C_{R/2} = \{x \in \mathcal{X} : \rho(x,x_0) \geq R/2\}$,

$$\begin{aligned} W_p(\mu_n,\mu)^p &= \int_{\mathcal{X}\times\mathcal{X}} \rho^p \,\mathrm{d}\gamma_n \\ &= \int_{S_1} \rho^p \,\mathrm{d}\gamma_n + \int_{S_2} \rho^p \,\mathrm{d}\gamma_n + \int_{S_3} \rho^p \,\mathrm{d}\gamma_n \\ &\leq \int_{\mathcal{X}\times\mathcal{X}} \rho^p \wedge R^p \,\mathrm{d}\gamma_n + 2^p \int_{C_{R/2}} \rho(x,x_0)^p \,\mathrm{d}\mu_n(x) + 2^p \int_{C_{R/2}} \rho(y,x_0)^p \,\mathrm{d}\mu(y). \end{aligned}$$

As $(\gamma_n)_{n\in\mathbb{N}}$ converges weakly to $\gamma = (\mathrm{Id},\mathrm{Id})_{\#}\mu$ which is concentrated on $\{(x,y) \in \mathcal{X}\times\mathcal{X} : \rho(x,y) = 0\}$,

$$\lim_{n\to\infty} \int_{\mathcal{X}\times\mathcal{X}} \rho^p \wedge R^p \,\mathrm{d}\gamma_n = \int_{\mathcal{X}\times\mathcal{X}} \rho^p \wedge R^p \,\mathrm{d}\gamma = 0.$$

Therefore,

$$\limsup_{n\to\infty} W_p(\mu_n,\mu)^p \leq 2^p \limsup_{n\to\infty} \int_{C_{R/2}} \rho(x,x_0)^p \,\mathrm{d}\mu_n(x) + 2^p \int_{C_{R/2}} \rho(y,x_0)^p \,\mathrm{d}\mu(y),$$

where the right-hand side vanishes as $R \to \infty$ due to (6.5) and the dominated convergence theorem. □

7 Useful Techniques and Their Applications

7.1 Transport plans under mappings

We introduce a useful technique to characterize transport plans under measurable mappings; see Lemma 3.12 of [AG13].

Proposition 7.1. *Let $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ be probability spaces. Given two measurable spaces $(\mathcal{Z}, \mathcal{C})$ and $(\mathcal{W}, \mathcal{D})$, suppose $f\colon \mathcal{X} \to \mathcal{Z}$ and $g\colon \mathcal{Y} \to \mathcal{W}$ are measurable. Define $(f,g)\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z} \times \mathcal{W}$ as $(f,g)(x,y) = (f(x), g(y))$.*

(i) For any transport plan $\gamma \in \Pi(\mu,\nu)$, a probability measure $(f,g)_{\#}\gamma$ on $\mathcal{Z} \times \mathcal{W}$ is a transport plan between $f_{\#}\mu$ and $g_{\#}\nu$, and hence

$$\{(f,g)_{\#}\gamma : \gamma \in \Pi(\mu,\nu)\} \subset \Pi(f_{\#}\mu, g_{\#}\nu). \tag{7.1}$$

(ii) Given any cost function c on $\mathcal{Z} \times \mathcal{W}$,

$$\inf_{\Gamma \in \Pi(f_{\#}\mu, g_{\#}\nu)} \int_{\mathcal{Z}\times\mathcal{W}} c(z,w)\, \mathrm{d}\Gamma(z,w) \leq \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(f(x), g(z))\, \mathrm{d}\gamma(x,y). \tag{7.2}$$

If $\mathcal{X}, \mathcal{Y}$ are Polish spaces, $\mathcal{A} = \mathscr{B}(\mathcal{X})$, and $\mathcal{B} = \mathscr{B}(\mathcal{Y})$, both (7.1) and (7.2) are equalities:

$$\{(f,g)_{\#}\gamma : \gamma \in \Pi(\mu,\nu)\} = \Pi(f_{\#}\mu, g_{\#}\nu)$$

and

$$\inf_{\Gamma \in \Pi(f_{\#}\mu, g_{\#}\nu)} \int_{\mathcal{Z}\times\mathcal{W}} c(z,w)\, \mathrm{d}\Gamma(z,w) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(f(x), g(z))\, \mathrm{d}\gamma(x,y).$$

Proof. One can verify (i) and (ii) easily. Now, suppose $\mathcal{X}, \mathcal{Y}$ are Polish spaces, $\mathcal{A} = \mathscr{B}(\mathcal{X})$, and $\mathcal{B} = \mathscr{B}(\mathcal{Y})$. Consider $\Gamma \in \Pi(f_{\#}\mu, g_{\#}\nu)$. We show that there exists $\gamma \in \Pi(\mu,\nu)$ such that $(f,g)_{\#}\gamma = \Gamma$. First, note that a probability measure $\gamma^{(1)} = (\mathrm{Id}, f)_{\#}\mu$ on $\mathcal{X} \times \mathcal{Z}$ satisfies $\gamma^{(1)} \in \Pi(\mu, f_{\#}\mu)$. By Theorem 1.1, we can find a collection $\{\gamma^{(1)}_z : z \in \mathcal{Z}\}$ of probability measures on $\mathcal{X}$ such that

$$\gamma^{(1)}(S_1) = \int_{\mathcal{Z}} \int_{\mathcal{X}} 1_{\{(x,z)\in S_1\}}\, \mathrm{d}\gamma^{(1)}_z(x) \mathrm{d}f_{\#}\mu(z) \quad \forall S_1 \in \mathcal{A} \otimes \mathcal{C}.$$

Similarly, for $\gamma^{(2)} := (\mathrm{Id}, g)_{\#}\nu \in \Pi(\nu, g_{\#}\nu)$ on $\mathcal{Y} \times \mathcal{W}$, we can find a collection $\{\gamma^{(2)}_w : w \in \mathcal{W}\}$ of probability measures on $\mathcal{Y}$ such that

$$\gamma^{(2)}(S_2) = \int_{\mathcal{W}} \int_{\mathcal{Y}} 1_{\{(y,w)\in S_2\}}\, \mathrm{d}\gamma^{(2)}_w(y) \mathrm{d}g_{\#}\nu(w) \quad \forall S_2 \in \mathcal{B} \otimes \mathcal{D}.$$

Now, we define γ as follows:

$$\gamma(S) = \int_{\mathcal{Z}\times\mathcal{W}} \gamma_z^{(1)} \otimes \gamma_w^{(2)}(S) \,\mathrm{d}\Gamma(z,w) \quad \forall S \in \mathcal{A}\otimes\mathcal{B},$$

or more generally, for any measurable $h\colon \mathcal{X}\times\mathcal{Y} \to [0,\infty]$,

$$\int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\,\mathrm{d}\gamma_z^{(1)}(x)\mathrm{d}\gamma_w^{(2)}(y)\,\mathrm{d}\Gamma(z,w)$$

To see $\gamma \in \Pi(\mu,\nu)$, note that for any $A \in \mathcal{A}$,

$$\gamma(A\times\mathcal{Y}) = \int_{\mathcal{Z}\times\mathcal{W}} \gamma_z^{(1)}(A)\,\mathrm{d}\Gamma(z,w) = \int_{\mathcal{Z}} \gamma_z^{(1)}(A)\,\mathrm{d}f_{\#}\mu(z) = \gamma^{(1)}(A\times\mathcal{Z}) = \mu(A).$$

Similarly, one can prove that $\gamma(\mathcal{X}\times B) = \nu(B)$ for any $B \in \mathcal{B}$. Next, we prove that $(f,g)_{\#}\gamma = \Gamma$. Note that a collection $(f_{\#}\gamma_z^{(1)})_{z\in\mathcal{Z}}$ amounts to conditional probability measures of $(f,\mathrm{Id})_{\#}\gamma^{(1)}$ given the marginal $f_{\#}\mu$ on the second coordinate. As $(f,\mathrm{Id})_{\#}\gamma^{(1)} = (f,f)_{\#}\mu$, the collection $(f_{\#}\gamma_z^{(1)})_{z\in\mathcal{Z}}$ must coincide $f_{\#}\mu$-almost everywhere with $(\delta_z)_{z\in\mathcal{Z}}$ due to uniqueness of disintegration. Therefore, for any measurable $H\colon \mathcal{Z}\times\mathcal{W}\to[0,\infty]$,

$$\begin{aligned}
\int_{\mathcal{Z}\times\mathcal{W}} H\,\mathrm{d}(f,g)_{\#}\gamma &= \int_{\mathcal{X}\times\mathcal{Y}} H(f(x),g(y))\,\mathrm{d}\gamma(x,y)\\
&= \int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}\times\mathcal{Y}} H(f(x),g(y))\,\mathrm{d}\gamma_z^{(1)}(x)\mathrm{d}\gamma_w^{(2)}(y)\,\mathrm{d}\Gamma(z,w)\\
&= \int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{Z}\times\mathcal{W}} H(z',w')\,\mathrm{d}f_{\#}\gamma_z^{(1)}(z')\mathrm{d}g_{\#}\gamma_w^{(2)}(w')\,\mathrm{d}\Gamma(z,w)\\
&= \int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{Z}\times\mathcal{W}} H(z',w')\,\mathrm{d}\delta_z(z')\mathrm{d}\delta_w(w')\,\mathrm{d}\Gamma(z,w)\\
&= \int_{\mathcal{Z}\times\mathcal{W}} H(z,w)\,\mathrm{d}\Gamma(z,w).
\end{aligned}$$

Hence, $(f,g)_{\#}\gamma = \Gamma$. The last part of the proposition follows accordingly. □

Remark 7.1. Suppose we know marginal distributions of (X,Y). Then, given two maps f and g, we know the joint distribution of $(f(X),g(Y))$. The preceding result tells that we can find a joint distribution of (X,Y) in accordance with this set of information. In particular, one possibility is to assume X and Y are independent given $f(X)$ and $g(Y)$, so that $X,Y|f(X),g(Y)$ is a tuple of two independent distributions $X|f(X)$ and $Y|g(Y)$.

Remark 7.2. We may interpret this results as follows: if we pushforward μ and ν by measurable maps, there are more transport plans between the image measures. Accordingly, the optimal transport cost from $f_{\#}\mu$ to $g_{\#}\nu$ given a cost function c is smaller than the optimal transport cost from μ to ν given a cost function $c\circ(f,g)$, i.e., $(x,y)\mapsto c(f(x),g(y))$.

Proposition 7.2. *Let $R \in SO(3)$ be a rotation matrix whose rotation angle is $\theta \in (0, \pi)$. Then, for any $\mu \in \mathscr{P}(\mathbb{R}^3)$,*

$$W_p^p(\mu, R_\# \mu) \leq (2\sin(\theta/2))^p \cdot \int_{\mathbb{R}^3} \|x\|_2^p \, \mathrm{d}\mu(x).$$

Proof. Let $P \colon \mathbb{R}^3 \to \mathbb{R}^3$ be a projection to the hyperplane orthogonal to the rotation axis. Then, for any $x \in \mathbb{R}^3$,

$$\|x - Rx\|_2 = 2\sin(\theta/2)\|Px\|_2 \leq 2\sin(\theta/2) \cdot \|x\|_2.$$

Due to Proposition 7.1 and Proposition 1.4,

$$\begin{aligned} W_p^p(\mu, R_\# \mu) &= \inf_{\gamma \in \Pi(\mu,\mu)} \int_{\mathbb{R}^3 \times \mathbb{R}^3} \|x - Ry\|_2^p \, \mathrm{d}\gamma(x, y) \\ &\leq \int_{\mathbb{R}^3} \|x - Rx\|_2^p \, \mathrm{d}\mu(x) \quad (\because (\mathrm{Id}, \mathrm{Id})_\# \mu \in \Pi(\mu, \mu)) \\ &\leq (2\sin(\theta/2))^p \cdot \int_{\mathbb{R}^3} \|x\|_2^p \, \mathrm{d}\mu(x). \end{aligned}$$

□

Proposition 7.3. *Let $(\mathcal{X}, \rho_\mathcal{X})$ and $(\mathcal{Z}, \rho_\mathcal{Z})$ be separable metric spaces. Suppose a map $f \colon \mathcal{X} \to \mathcal{Z}$ is L-Lipschitz, i.e., $\rho_\mathcal{Z}(f(x), f(y)) \leq L \cdot \rho_\mathcal{X}(x, y)$ for all $x, y \in \mathcal{X}$. Then, for any $\mu, \nu \in \mathscr{P}(\mathcal{X})$,*

$$W_p(f_\# \mu, f_\# \nu) \leq L \cdot W_p(\mu, \nu),$$

where W_p's on the left-hand side and right-hand side denote the Wasserstein distances on $\mathscr{P}(\mathcal{Z})$ and $\mathscr{P}(\mathcal{X})$, respectively.

Proof. Due to Proposition 7.1 and Lipschitzness of f,

$$\begin{aligned} W_p^p(f_\# \mu, f_\# \nu) &= \inf_{\Gamma \in \Pi(f_\# \mu, f_\# \nu)} \int_{\mathcal{Z} \times \mathcal{Z}} \rho_\mathcal{Z}(z_1, z_2)^p \, \mathrm{d}\Gamma(z_1, z_2) \\ &= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} \rho_\mathcal{Z}(f(x_1), f(x_2))^p \, \mathrm{d}\gamma(x_1, x_2) \\ &\leq L^p \cdot \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} \rho_\mathcal{X}(x_1, x_2)^p \, \mathrm{d}\gamma(x_1, x_2) \\ &= L^p \cdot W_p^p(\mu, \nu). \end{aligned}$$

□

Proposition 7.4. *Fix $k, d \in \mathbb{N}$ and $U \in \mathbb{R}^{k \times d}$. Define $L \colon \mathbb{R}^d \to \mathbb{R}^k$ as $L(x) = Ux$. Then, for any $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$,*

$$W_p(L_\# \mu, L_\# \nu) \leq \|U\|_2 \cdot W_p(\mu, \nu),$$

where $\|U\|_2$ denotes the spectral norm of U.

Proof. Since

$$\|L(x) - L(y)\|_2 = \|U^\top (x-y)\|_2 \le \|U\|_2 \cdot \|x-y\|_2 \quad \forall x, y \in \mathbb{R}^d,$$

Proposition 7.3 implies $W_p((L_U)_\# \mu, (L_U)_\# \nu) \le \|U\|_2 \cdot W_p(\mu, \nu)$. □

7.2 Total variation and transport plans

Definition 7.1. For probability measures μ, ν on a measurable space $(\mathcal{X}, \mathcal{A})$, we define the total variation distance as

$$\mathsf{TV}(\mu, \nu) := \sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|.$$

The total variation distance essentially finds a region where the two measures differ the most. In measure theory, this concept is introduced as a more general concept of the total variation norm of a signed measure: for any measures μ, ν on $(\mathcal{X}, \mathcal{A})$, the signed measure $\mu - \nu$ has a Jordan decomposition $\mu - \nu = (\mu - \nu)_+ - (\mu - \nu)_-$, where $(\mu - \nu)_+$ and $(\mu - \nu)_-$ are mutually singular. Then, $|\mu - \nu| = (\mu - \nu)_+ + (\mu - \nu)_-$ is a positive measure, and the total variation norm $\|\mu - \nu\|$ is defined as

$$\|\mu - \nu\| := |\mu - \nu|(\mathcal{X}) = (\mu - \nu)_+(\mathcal{X}) + (\mu - \nu)_-(\mathcal{X}).$$

Here, the positive part $(\mu - \nu)_+$ satisfies

$$(\mu - \nu)_+(\mathcal{X}) = \sup_{A \in \mathcal{A}} (\mu - \nu)(A) = \sup_{A \in \mathcal{A}} (\mu(A) - \nu(A)),$$

which coincides with the total variation distance $\mathsf{TV}(\mu, \nu)$ defined above when μ, ν are probability measures; in fact, it suffices to have $\mu(\mathcal{X}) = \nu(\mathcal{X})$. Hence, in this case, we have

$$\mathsf{TV}(\mu, \nu) = (\mu - \nu)_+(\mathcal{X}) = (\mu - \nu)_-(\mathcal{X}) = \frac{1}{2}\|\mu - \nu\|.$$

In Definition 7.1, note that we can write

$$\mathsf{TV}(\mu, \nu) = \sup_{A \in \mathcal{A}} \left(\int_{\mathcal{X}} 1\{x \in A\} \, \mathrm{d}\mu(x) - \int_{\mathcal{X}} 1\{y \in A\} \, \mathrm{d}\nu(y) \right).$$

For any $A \in \mathcal{A}$, note that $1\{x \in A\} - 1\{y \in A\} \le 1\{x \neq y\}$ for any $x, y \in \mathcal{X}$. Therefore, for any coupling $\gamma \in \Pi(\mu, \nu)$, we deduce that

$$\mathsf{TV}(\mu, \nu) \le \int_{\mathcal{X} \times \mathcal{X}} 1\{x \neq y\} \, \mathrm{d}\gamma(x, y),$$

which implies

$$\mathsf{TV}(\mu, \nu) \le \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} 1\{x \neq y\} \, \mathrm{d}\gamma(x, y). \tag{7.3}$$

It turns out that the inequality is actually an equality, and we can find a transport plan that achieves the infimum on the right-hand side of (7.3).

Theorem 7.1. *For any probability measures μ, ν on a measurable space $(\mathcal{X}, \mathcal{A})$, we have*

$$\mathsf{TV}(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} 1\{x \neq y\}\, \mathrm{d}\gamma(x,y),$$

where the infimum on the right-hand side is achievable.

We find a transport plan $\gamma \in \Pi(\mu,\nu)$ such that

$$\int_{\mathcal{X}\times\mathcal{X}} 1\{x \neq y\}\, \mathrm{d}\gamma(x,y) = \mathsf{TV}(\mu,\nu).$$

Here, we want to find a transport plan that puts as much mass as possible on the diagonal $\Delta = \{(x,x) : x \in \mathcal{X}\}$ while satisfying the marginal constraints μ and ν. To achieve this, we focus on the subset of $\mathcal{X}$ where μ and ν overlap the most, which we put on Δ, while the remaining mass is distributed to the complement of Δ in a way that respects the marginal constraints. The following lemma provides a concrete construction of such a transport plan.

Lemma 7.1. *Suppose μ, ν are probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. Let $\mu - \nu = (\mu-\nu)_+ - (\mu-\nu)_-$ be the Jordan decomposition of a signed measure $\mu - \nu$. Define $\mu \wedge \nu := \mu - (\mu-\nu)_+$ and $\delta = (\mu-\nu)_+(\mathcal{X}) = \mathsf{TV}(\mu,\nu)$. Then,*

$$\gamma := (\mathrm{Id},\mathrm{Id})_\#(\mu\wedge\nu) + \frac{1}{\delta}(\mu-\nu)_+ \otimes (\mu-\nu)_- \in \Pi(\mu,\nu).$$

Furthermore, suppose $\Delta := \{(x,x) : x \in \mathcal{X}\} \in \mathcal{A}\otimes\mathcal{A}$. Then, the two measures $(\mathrm{Id},\mathrm{Id})_\#(\mu\wedge\nu)$ and $(\mu-\nu)_+ \otimes (\mu-\nu)_-$ are mutually singular; more precisely, they are concentrated on Δ and $(\mathcal{X}\times\mathcal{X})\backslash\Delta$, respectively. In particular,

$$\gamma((\mathcal{X}\times\mathcal{X})\backslash\Delta) = \delta = \mathsf{TV}(\mu,\nu).$$

Proof. First, we note that γ is a probability measure because $\gamma(\mathcal{X}\times\mathcal{X}) = 1$ follows from

$$(\mathrm{Id},\mathrm{Id})_\#(\mu\wedge\nu)(\mathcal{X}\times\mathcal{X}) = (\mathrm{Id},\mathrm{Id})_\#(\mu\wedge\nu)(\Delta) = (\mu\wedge\nu)(\mathcal{X}) = 1-\delta,$$
$$\frac{1}{\delta}(\mu-\nu)_+ \otimes (\mu-\nu)_-(\mathcal{X}\times\mathcal{X}) = \frac{1}{\delta}(\mu-\nu)_+(\mathcal{X})(\mu-\nu)_-(\mathcal{X}) = \delta.$$

Then, $\gamma \in \Pi(\mu,\nu)$ follows by verifying $\gamma(A\times\mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X}\times B) = \nu(B)$. As $(\mu-\nu)_+ \perp (\mu-\nu)_-$, there are disjoint sets $E, F \in \mathcal{A}$ such that $E\cup F = \mathcal{X}$, $(\mu-\nu)_+(F) = 0$, and $(\mu-\nu)_-(E) = 0$. Hence,

$$(\mu-\nu)_+\otimes(\mu-\nu)_-(\Delta) = (\mu-\nu)_+\otimes(\mu-\nu)_-(\Delta\cap(E\times\mathcal{X})) + (\mu-\nu)_+\otimes(\mu-\nu)_-(\Delta\cap(F\times\mathcal{X})).$$

By definition,

$$\begin{aligned}(\mu-\nu)_+\otimes(\mu-\nu)_-(\Delta\cap(E\times\mathcal{X})) &\le (\mu-\nu)_+\otimes(\mu-\nu)_-\{(x,x) : x\in E\}\\ &\le (\mu-\nu)_+\otimes(\mu-\nu)_-(\mathcal{X}\times E)\\ &= (\mu-\nu)_-(E)\\ &= 0.\end{aligned}$$

Similarly, $(\mu-\nu)_+\otimes(\mu-\nu)_-(\Delta\cap(F\times\mathcal{X})) = 0$. Therefore, $(\mu-\nu)_+\otimes(\mu-\nu)_-(\Delta) = 0$. □

Having a particular transport plan is useful for deriving upper bounds on the Wasserstein distance. The following result upper bounds the Wasserstein distance W_p by the total variation distance; see Proposition 7.10 of [Vil03] for the proof.

Proposition 7.5. *Let $(\mathcal{X}, \rho)$ be a separable metric space. For $\mu, \nu \in \mathscr{P}(\mathcal{X})$,*

$$W_p^p(\mu, \nu) \le 2^{p-1} \int_{\mathcal{X}} \rho(x, x_0)^p \, \mathrm{d}|\mu - \nu|(x).$$

References

[AB06] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide.* Springer, third edition, 2006.

[ABS21] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport.* Springer, 2021.

[AG13] Luigi Ambrosio and Nicola Gigli. *A User's Guide to Optimal Transport*, pages 1–155. Springer-Verlag, 2013.

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Birkhäuser, 2005.

[Bil99] Patrick Billingsley. *Convergence of Probability Measures.* Wiley, second edition, 1999.

[Ç11] Erhan Çınlar. *Probability and Stochastics.* Springer, 2011.

[Coh13] Donald L. Cohn. *Measure Theory.* Birkhäuser, second edition, 2013.

[Dud02] R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, second edition, 2002.

[FG21] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows.* EMS Press, 2021.

[Kal97] Olav Kallenberg. *Foundations of Modern Probability.* Springer, 1997.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Birkhäuser, 2015.

[Vil03] Cédric Villani. *Topics in Optimal Transportation.* American Mathematical Society, 2003.

[Vil09] Cédric Villani. *Optimal Transport: Old and New.* Springer, 2009.