arXiv:2201.03483v1 [econ.TH] 10 Jan 2022

Simultaneous optimal transport

Ruodu Wang[*] Zhenyuan Zhang[†]

11th January 2022

Abstract

We propose a general framework of mass transport between vector-valued measures, which will be called simultaneous transport. The new framework is motivated by the need to transport resources of different types simultaneously, i.e., in single trips, from specified origins to destinations. In terms of matching, one needs to couple two groups, e.g., buyers and sellers, by equating supplies and demands of different goods at the same time. The mathematical structure of simultaneous transport is very different from the classic setting of optimal transport, leading to many new challenges. The Monge and Kantorovich formulations are contrasted and connected. Existence and uniqueness of the simultaneous transport and duality formulas are established, and a notion of Wasserstein distance in this setting is introduced. In particular, the duality theorem gives rise to a labour market equilibrium model where each worker has several types of skills and each firm seeks to employ these skills at different levels.

Contents

[*] Department of Statistics and Actuarial Science, University of Waterloo, Canada. Email: wang@uwaterloo.ca.

[†] Department of Mathematics, Stanford University, USA. Email: zzy@stanford.edu.

1 Introduction

Optimal transport theory, originally developed by Monge and Kantorovich (see Villani (2009) for a history), has wide applications in various scientific fields, including economic theory (e.g., Galichon (2016)), operations research (e.g., Blanchet and Murthy (2019)), statistics (e.g., Carlier et al. (2016)), machine learning (e.g., Peyré and Cuturi (2019)), and quantitative finance (e.g., Beiglböck et al. (2013)). For a mathematical background on optimal transport and its applications, we refer to the textbooks of Ambrosio (2003), Santambrogio (2015), and Villani (2003, 2009).

In the recent decade, optimal transport has received increasing attention in economic studies with many relevant applications, such as contract design (Ekeland (2013)), robust risk assessment (Embrechts et al. (2013)), Cournot-Nash equilibria in non-atomic games (Blanchet and Carlier (2016)), multiple-good monopoly (Daskalakis et al. (2017)), implementation problems (Nöldeke and Samuelson (2018)), and team matching (Boerma et al. (2021)). A specialized treatment of optimal transport in economics is given by Galichon (2016).

In this paper, we propose a new framework of optimal transport, which we call the *simultaneous optimal transport*. In contrast to the classic optimal transport theory, which studies transports between two measures on spaces X and Y, a simultaneous transport (either Monge or kernel; see Section 2) moves mass from d measures on X to d measures on Y *simultaneously*. As a primary example (see Example 2.1 for details), suppose that several factories need to supply d types of products to several retailers, and each factory only has one truck to transport their products to one destination. Since each product type has its own supply and demand, the objective is to make a transport plan such that all demands are met. In case $d = 1$, we speak of the classic optimal transport problem. The problem formulation and a few motivating examples are presented in Section 2. Simultaneous transport provides powerful tools for matching problems with multiple distributional constraints; a worker-firm equilibrium model will be analyzed in Section 5.2.

A considerable amount of new challenges and relevant applications arise in this new framework for $d \geqslant 2$, which will gradually be revealed in this paper. We will explain below several sharp contrasts between the new and the classic frameworks, along with our contributions and results.

First, inspired by the example above, the measures at origin (supplies) do not necessarily have the same mass as the measures at destination (demands to meet). Obviously, there does not exist a possible transport if the demands (in any product type) are larger than the supplies, but there can be transports if the demands are smaller than the supplies. We will say that the transport problem is *balanced* if the vector of total masses at origin is equal to the vector of total masses at destination, and otherwise it is *unbalanced* (see Section 2 for a precise definition). Unbalance is generally not an issue if $d = 1$ since one can glue a point at the destination which incurs no transport cost to reformulate the problem as a balanced problem, but such a trick does not work in the simultaneous transport setting; see Section 2 for an explanation. A connection between the balanced and unbalanced settings is established in Section 4 via a continuity result.

Second, one needs to specify a reference measure with respect to which the transport cost is computed. In classic transport theory, the cost is integrated with respect to the measure at origin (supply). In the example above, it seems that none of the distributions of the product supplies is a natural benchmark for computing the cost; neither are their combinations. A separate benchmark measure needs to be introduced (see Section 2), and it may cause extra technical subtlety depending on whether it is equivalent to a measure dominating the measures at origin.

Third, for two given d-tuples of (probability) measures, a simultaneous transport may not exist, even if there are no atoms in these measures (transports between atomless probabilities always exist in case $d = 1$). As a trivial example, suppose that there are a continuum of factories, each supplying an equal amount of product A and product B, and a continuum of retailers, half demanding a ratio of $2 : 1$ between products A and B and the other half demanding a ratio of $1 : 2$ between A and B. If the total demand vector is equal to the total supply vector, then there is obviously no possible transport plan; indeed, any transport plan would supply the same amount of A and B to any retailer, leading to oversupplying of one product for each supplier. However, if, instead of a $1 : 1$ ratio, half of the factories supply in a $3 : 1$ ratio between A and B, and the other half supply in a $1 : 3$ ratio, then transport plans exist, and we can choose from these plans to minimize the total transport cost. Moreover, it is easy to see from this example that the transport problems are not symmetric in the measures at origin and the measures at destination, in sharp contrast to the classic problem. Even if transport plans exist, the set which it can be chosen from is bound to additional constraints. The existence issue of simultaneous transport will be studied in Section 3 using the notions of joint non-atomicity and heterogeneity order, based on existing results in Torgersen (1991) and Shen et al. (2019). Several other interesting inequalities and observations, which do not appear in the classic setting, are also discussed in Section 3.

Fourth, in the balanced setting, the classic transport problem can be conveniently written in the Kantorovich formulation as each transport corresponds to a joint probability measure with specified marginals but unspecified dependence structure (or a copula, see e.g., Beare (2010) and Joe (2014)). In the framework of simultaneous transport, since there is no "first marginal" or "second marginal" of the problem (instead, two vectors of marginals), the Kantorovich formulation via joint distributions is less clear than in the classic case, and it is studied in Section 4. Assuming joint non-atomicity, we prove that the Monge and Kantorovich (kernel) formulations have the same infimum cost.

Fifth, a duality theorem for balanced simultaneous transport is obtained in Section 5, which has a different form compared with the classic duality formula. Using the duality result, we construct a labour market equilibrium model (see e.g., Galichon (2016) for a classic equilibrium model in case $d = 1$), where workers, each with several types of skills and seeking to optimize their wage, are matched with firms, each seeking to employ these skills of a certain cumulative amount to optimize their profit. The equilibrium wage function and the equilibrium profit function are obtained from the duality formula for given distributions of the skills that workers have and firms seek.

Sixth, the Wasserstein quasi-metric is defined for simultaneous transports, but it does not have symmetry since the transport problem is clearly not symmetric in the two vectors of measures. In case transports from and to the measures at origin both exist, we speak of *two-way* transport problems (in the classic setting via the Kantorovich formulation, all balanced transport problems are two-way). In the setting of measures for which two-way transports exist, the Wasserstein distance can be naturally defined in Section 6. In addition, we provide a decomposition formula of the optimal transport and its optimal cost in the two-way transport setting, which can be solved explicitly based on existing results (e.g., Gangbo and McCann (1996)) if the cost function is convex.

Finally, due to the new structure of the simultaneous transport problem, we are able to discuss *uniqueness* of the transport[1] for some given vectors of measures. Uniqueness is shown in Section 6 for the two-way transport setting under additional conditions on the structure of the measures at origin. Note that uniqueness of the Kantorovich transport does not appear at all in case $d = 1$ except for the trivial case where the measure at origin is degenerate.

In Section 7 we conclude the paper with several promising directions of future research and open challenges. All technical proofs are put in Appendices A-D except for a few very short ones that are kept within the main text. In recent years there has been a growing interest in various generalizations of the classic Monge-Kantorovich optimal transport problem. A few generalizations of optimal transport in higher dimensions are related to our paper, which we collect in Appendix E; the closest to our framework is perhaps Wolansky (2020) who considered a similar setting to our simultaneous transport with a different focus and distinctive mathematical results.

2 Simultaneous optimal transport

We first briefly review the classic Monge-Kantorovich transport problem. For a measurable space X that is also a Polish space equipped with the Borel σ-field $\mathcal{B}(X)$, we denote by $\mathcal{P}(X)$ the set of all Borel probability measures on X. Consider Polish spaces X, Y, and probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Although our results are formulated on general Polish spaces, it does not hurt to think of $X = \mathbb{R}^N$ and $Y = \mathbb{R}^N$ as the primary example. We will always equip $X \times Y$ with the product σ-field. In the following while writing $A \subseteq X,\ B \subseteq Y$, we always assume that A, B are Borel measurable subsets. Given a cost function $c : X \times Y \to [0, +\infty]$, the classic optimal transport problem raised by Monge asks for

$$\inf_{T \in \mathcal{T}(\mu,\nu)} \int_X c(x, T(x))\mu(\mathrm{d}x),$$

where $\mathcal{T}(\mu, \nu)$ consists of transport maps from μ to ν, i.e., measurable functions $T : X \to Y$ such that $\mu \circ T^{-1} = \nu$.

Kantorovich later studied a relaxation of Monge's problem, that is, to solve for

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

where $\Pi(\mu, \nu)$ is the set of transport plans from μ to ν, i.e., the set of probability measures $\pi \in \mathcal{P}(X \times Y)$ such that for any $A \subseteq X$ and $B \subseteq Y$, $\pi(A \times Y) =$

[1] Note this is different from *uniqueness* of the *optimal* transport.

$\mu(A)$ and $\pi(X \times B) = \nu(B)$. These are the celebrated Monge-Kantorovich optimal transport problems.

2.1 Simultaneous transport

Throughout, we denote by $d \in \mathbb{N}$ the dimension of a vector-valued measure, where the more interesting case is when $d \geqslant 2$, and by $[d] = \{1, \ldots, d\}$. We work with d-tuples of finite Borel measures $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ on X and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$ on Y such that for each $j \in [d]$, $\mu_j(X) \geqslant \nu_j(Y) > 0$.

We propose the new framework of *simultaneous optimal transport* by requiring that a certain transport *map* or transport *plan* sends μ_j to cover ν_j simultaneously for all $j \in [d]$. In this setup, the set of all simultaneous transport maps is defined as

$$\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \{T : X \to Y \mid \boldsymbol{\mu} \circ T^{-1} \geqslant \boldsymbol{\nu}\}.$$

Here and throughout, equalities and inequalities are understood componentwise, and for two measures μ and ν on the same space, $\mu \geqslant \nu$ means that $\mu(A) \geqslant \nu(A)$ for all measurable A. If $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$, then we speak of *balanced* simultaneous transports.

The most natural and intuitive way to describe the set of all (simultaneous) transport plans is to use $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, the set of all stochastic kernels κ such that $\kappa_\# \boldsymbol{\mu} \geqslant \boldsymbol{\nu}$, and defined as

$$\kappa_\# \boldsymbol{\mu}(\cdot) := \int_X \kappa(x; \cdot) \boldsymbol{\mu}(\mathrm{d}x) \geqslant \boldsymbol{\nu}(\cdot). \tag{1}$$

Imagine that one would like to distribute goods from an (possibly infinitesimal) point $x \in X$ to different places in Y, then the measure $\kappa(x; \cdot)$ describes such a distribution. In view of this definition, the set of stochastic kernels $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ can be written as an intersection:

$$\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \bigcap_{j=1}^{d} \mathcal{K}(\mu_j, \nu_j).$$

In words, a simultaneous transport plan from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ sends simultaneously μ_j to ν_j for any $j \in [d]$. The non-emptyness of $\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is not guaranteed generally, and will be explained later in Section 3.1.

In the case $d = 1$ and $\mu(X) = \nu(Y)$, our problem reduces to the classic Monge-Kantorovich problem. We first illustrate an example for simultaneous transport problems, which sheds some light on the special structure and technical difference of our problem in contrast to the classic problem.

Example 2.1 (Simultaneous transport of supplies)**.** Suppose that there are m factories; each factory j has a_j units of product A and b_j units of product B. There are m' retailers, each demanding a'_k units of A and b'_k units of B. We assume that the supply is enough to cover the demand, that is, with normalization,

$$1 = \sum_{j=1}^{m} a_j \geqslant \sum_{j=1}^{m'} a'_k \quad \text{and} \quad 1 = \sum_{j=1}^{m} b_j \geqslant \sum_{j=1}^{m'} b'_k.$$

If we assume demand-supply clearance, then, with normalization,

$$\sum_{j=1}^{m} a_j = \sum_{j=1}^{m'} a'_k = \sum_{j=1}^{m} b_j = \sum_{j=1}^{m'} b'_k = 1. \tag{2}$$

Let μ_1 be a probability such that $\mu_1(\{j\}) = a_j$ for each j, and similarly, $\mu_2(\{j\}) = b_j$ for each j, and $\nu_1(\{k\}) = a'_k$ and $\nu_2(\{k\}) = b'_k$ for each k. Write $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\boldsymbol{\nu} = (\nu_1, \nu_2)$.

1. A transport in $\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ or $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, if it exists, is an arrangement to send products from factories to retailers to meet their demand. We cannot transport products within the m' retailers or within the m factories.

2. The transport in $\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is required to be done in single trips: One factory can only supply one retailer. This is illustrated in Figure 1 (a). As a practical example, we may think of the situation where each factory only has one truck that goes to one destination in every production cycle.

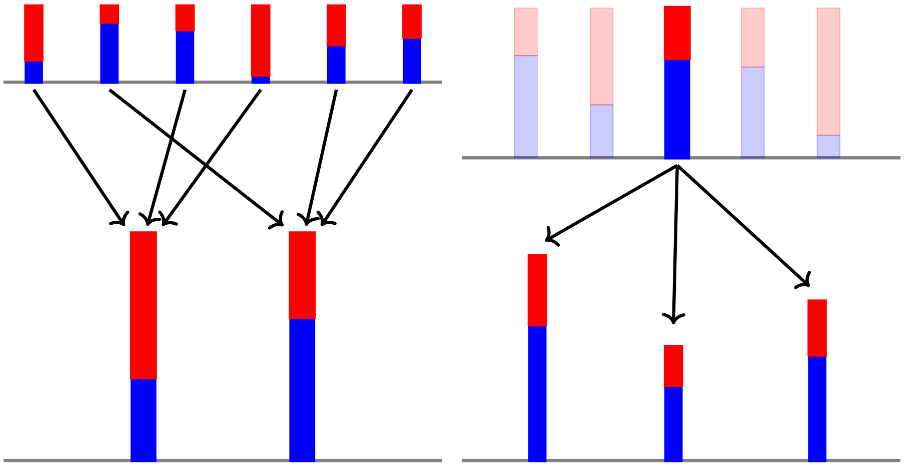
3. We may allow each factory to supply multiple retailers, e.g., a factory with multiple trucks. In this case, we can use the formulation of transport kernels in $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. We note that a non-trivial constraint imposed by the formulation (1) is that the amount of A and that of B are proportional in each truck departing from the same factory (e.g., bundled goods, or worker skills in Section 5.2 which are not divisible). This is illustrated in Figure 1 (b).

4. If demand-supply clearance (2) holds, then the transport is balanced; otherwise it is unbalanced. In case (2) holds, one may consider the backward direction of transporting $\boldsymbol{\nu}$ to $\boldsymbol{\mu}$, and this leads to two-way transports treated in Section 6.

Example 2.1 and its continuous version will serve as a primary example to facilitate understanding of our new framework. To quantify the cost of simultaneous transports, a cost function will be associated to the simultaneous transport problem, as in the classic formulation. Throughout, we define the normalized average measures

$$\bar{\mu} := \frac{\sum_{j=1}^{d} \mu_j}{\sum_{j=1}^{d} \mu_j(X)} \text{ and } \bar{\nu} := \frac{\sum_{j=1}^{d} \nu_j}{\sum_{j=1}^{d} \nu_j(Y)},$$

which are probability measures. In case $\mu_1, \dots, \mu_d$ are themselves probability measures, $\bar{\mu}$ is their arithmetic average. Consider a measurable function c :



(a) Monge

(b) Kernel

Figure 1: A showcase of simultaneous transport of supplies; red and blue represent different types of products.

$X \times Y \to [0, +\infty]$ and a *reference* probability measure η on X such that $\eta \ll \bar{\mu}$. We define the transport costs as follows: for $T \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$, let

$$\mathcal{C}_\eta(T) := \int_X c(x, T(x))\eta(\mathrm{d}x). \tag{3}$$

Such a reference measure η allows us the greatest generality in view of Example 2.1: We allow nonlinear dependencies of η in terms of $\boldsymbol{\mu}$, for example, when computing the petrol cost which is nonlinear in weights of the transported products. We impose the condition $\eta \ll \bar{\mu}$ because it would be unreasonable to assign a cost where there is no transport. (For general $\eta \in \mathcal{M}(X)$, we can always normalize it to a probability without loss of generality.)

In terms of $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, we define the transport cost

$$\mathcal{C}_\eta(\kappa) := \int_{X \times Y} c(x, y)\eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y). \tag{4}$$

The quantities of interest are the minimum (or infimum) costs

$$\inf_{T \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(T) \quad \text{and} \quad \inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa).$$

If $\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ or $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is an empty set, the corresponding minimum cost is set to $+\infty$. In dimension one, this cost coincides with the classic Monge-Kantorovich costs. In case $\eta = \bar{\mu}$, we omit the subscript η in (3) and (4).

Remark 2.2. If the supports of $\bar{\mu}, \bar{\nu}$ are both finite (e.g., Example 2.1), then the optimal transport problem is equivalent to a finite-dimensional linear programming problem, which can be handled conveniently by linear programming solvers. The dimension $d \geqslant 2$ of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ leads to more constraints in this linear program compared to the classic case of $d = 1$. These additional constraints are highly non-trivial. For instance, the additional constraints may rule out the existence of any transport, in contrast to the case $d = 1$; see Section 3.1.

2.2 Balanced simultaneous transport

Although we have set up the problem in greater generality with unbalanced measures, in some parts of this paper we will focus on the balanced case where $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$. We may without loss of generality assume that each μ_j, ν_j are probability measures. In this case, we have

$$\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{T : X \to Y \mid \boldsymbol{\mu} \circ T^{-1} = \boldsymbol{\nu}\}$$

and

$$\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\kappa \mid \kappa_{\#}\boldsymbol{\mu} = \boldsymbol{\nu}\}.$$

The two examples below illustrate some particular applications of this setting, in addition to the supply-demand clearing case (2) of Example 2.1.

Example 2.3 (Financial cost efficiency with multiple distributional constraints)**.** Let $(X, \mathcal{F})$ be a measurable space on which $\mu_1, \dots, \mu_d$ are d probability measures and $\mathcal{L}$ be the set of random variables on $(X, \mathcal{F})$. Let $\nu_1, \dots, \nu_d$ be d distributions on $\mathbb{R}$ and define

$$\mathcal{L}_{\boldsymbol{\nu}}(\boldsymbol{\mu}) := \{L \in \mathcal{L} \mid L \overset{\text{law}}{\sim}_{\mu_i} \nu_i,\ i \in [d]\}$$

where $L \overset{\text{law}}{\sim}_{\mu} \nu$ means that L has distribution ν under μ. The set $\mathcal{L}_{\boldsymbol{\nu}}(\boldsymbol{\mu})$ represents all possible financial positions which has distribution ν_j under a reference probability μ_j. As an example in case $d = 2$, an investor may seek for an investment L which has a target distribution ν_1 under her subjective probability measure μ_1 and is bound by regulation to have a distribution ν_2 under a regulatory measure μ_2; see Shen et al. (2019, Section 5). The investor is interested in the optimization problem

$$\min \{\mathbb{E}^{\eta}[f(L)] \mid L \in \mathcal{L}_{\boldsymbol{\nu}}(\boldsymbol{\mu})\} \tag{5}$$

where $\eta \ll \bar{\mu}$ and f is a nonnegative measurable function. The probability measure η can be seen as a pricing measure on the financial market, and the optimization problem (5) is to find the cheapest financial position $f(L)$ with L satisfying the distributional constraints. In case $d = 1$, i.e., with only one distributional constraint, this problem is the cost-efficient portfolio problem studied by Dybvig (1988), which can be solved by the classic Fréchet-Hoeffding inequality (e.g., Rüschendorf (2013)). For $d \geqslant 2$, the problem becomes much more complicated, and a special case of mutually singular $\mu_1, \dots, \mu_d$ is studied in Wang and Ziegel (2021).

Note that by definition $L_{\boldsymbol{\nu}}(\boldsymbol{\mu}) = \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Hence, $L \in L_{\boldsymbol{\nu}}(\boldsymbol{\mu})$ is a balanced Monge transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$, and

$$\mathbb{E}^{\eta}[f(L)] = \int_X f(L(\omega))\eta(\mathrm{d}\omega),$$

which is simply the transport cost of L as a Monge transport, with cost function $c(x,y)=f(y)$ and reference measure η. We will see from Theorem 4.2 that if f is continuous and $\boldsymbol{\mu}$ is jointly atomless, then the infimum of the cost is the same as the infimum cost among the corresponding transport plans. If $\eta \sim \bar{\mu}$, further duality results from Section 5 are applicable.

Example 2.4 (Time-homogeneous Markov processes with specified marginals)**.** Let $\mu_1,\dots,\mu_T$ be probability measures on $X=\mathbb{R}^N$ and $\xi=(\xi_t)_{t=1,\dots,T}$ be an $\mathbb{R}^N$-valued Markov process with marginal distributions $\mu_1,\dots,\mu_T$. The Markov kernels of ξ, $\kappa_t:\mathbb{R}^N\to\mathcal{P}(\mathbb{R}^N)$ for $t=1,\dots,T-1$, are such that $\kappa_t(\mathbf{x})$ is the distribution of ξ_{t+1} conditional on $\xi_t=\mathbf{x}$,. Here and throughout conditional distributions (probabilities) should be understood as regular conditional distributions (probabilities). The Markov process ξ is time-homogeneous if $\kappa:=\kappa_t$ does not depend on t. In other words, κ needs to satisfy

$$\mu_{t+1}=\int_{\mathbb{R}^N}\kappa(\mathbf{x})\mu_t(\mathrm{d}\mathbf{x})\qquad\text{for } t=1,\dots,T-1.$$

Therefore, the distribution of a time-homogeneous Markov process with marginals $(\mu_1,\dots,\mu_T)$ corresponds to the Markov kernel $\kappa\in\mathcal{K}(\boldsymbol{\mu},\boldsymbol{\nu})$ where $\boldsymbol{\mu}=(\mu_1,\dots,\mu_{T-1})$, and $\boldsymbol{\nu}=(\mu_2,\dots,\mu_T)$, which is a simultaneous transport kernel. With the tool of simultaneous optimal transport, we can study *optimal* (in some sense) time-homogeneous Markov processes. A special case of this example will be given in Proposition 3.7.

In the classic optimal transport framework with $d=1$, an unbalanced transport problem can be converted to a balanced transport problem by adjoining a point y_0 to the space Y with mass $\mu(X)-\nu(Y)$ and such that $c(x,y_0)=0$ for all x. However, for $d\geqslant 2$ the two problems are not equivalent. The reason that the conversion works for $d=1$ is that the set of unbalanced transports

$$\mathcal{K}(\boldsymbol{\mu},\boldsymbol{\nu})=\{\kappa\mid\kappa_{\#}\boldsymbol{\mu}\geqslant\boldsymbol{\nu}\}=\{\kappa\mid\kappa_{\#}\boldsymbol{\mu}=\tilde{\boldsymbol{\nu}}\text{ for some }\tilde{\boldsymbol{\nu}}\geqslant\boldsymbol{\nu}\}$$

is identical to the set of transports

$$\mathcal{K}'(\boldsymbol{\mu},\boldsymbol{\nu}):=\{\kappa\mid\kappa_{\#}\tilde{\boldsymbol{\mu}}=\boldsymbol{\nu}\text{ for some }\tilde{\boldsymbol{\mu}}\leqslant\boldsymbol{\mu}\}.$$

This is not necessarily true in case $d\geqslant 2$. For example, take $\mu_1=\mu_2$ be two times the Dirac measure at 0, ν_1 be uniform on $[-1,0]$ and ν_2 uniform on $[0,1]$. Then the kernel κ sending 0 uniformly to $[-1,1]$ belongs to $\mathcal{K}(\boldsymbol{\mu},\boldsymbol{\nu})$ while $\mathcal{K}'(\boldsymbol{\mu},\boldsymbol{\nu})$ is clearly empty. This subtle issue also hints on the additional technical challenges when dealing with simultaneous transports.

2.3 Assumptions and standing notation

We will focus on different levels of generality in the subsequent sections, with the following hierarchical structure on the imposed assumptions.

i. In Sections 3 and 4 we will prove general results in the unbalanced setting;

ii. in Sections 5 and 6 we work within the balanced setting;

iii. in Section 6 we further require that both $\mathcal{K}(\boldsymbol{\mu},\boldsymbol{\nu})$ and $\mathcal{K}(\boldsymbol{\nu},\boldsymbol{\mu})$ are non-empty.

In terms of the reference measure we have the following hierarchy of considerations.

i. In Section 3 we make no further assumption on the reference measure η except that $\eta \ll \bar{\mu}$;

ii. in Sections 4 and 5 we assume that $\eta \sim \bar{\mu}$;[2]

iii. in Section 6 and throughout our examples we assume for simplicity that $\eta = \bar{\mu}$.

Throughout, we consider the general setting where X and Y are Polish spaces, unless otherwise stated (in Theorem 5.2 and some examples). We let $\mathbb{1}_A$ stand for the indicator of a set A, and $\mathbb{R}_+ := [0,+\infty)$. The set $\mathcal{M}(X)$ is the collection of all finite and non-zero Borel measures on X.

3 Existence, inequalities, and examples

3.1 Existence of simultaneous transports

We first state a condition to guarantee that $\mathcal{K}(\boldsymbol{\mu},\boldsymbol{\nu})$ and $\mathcal{T}(\boldsymbol{\mu},\boldsymbol{\nu})$ are non-empty. We write $\mu \gg \boldsymbol{\mu}$ if μ dominates each component of $\boldsymbol{\mu}$. The following definition is adapted from Shen et al. (2019) where $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$ is assumed.

Definition 3.1. Let $\boldsymbol{\mu} \in \mathcal{M}(X)^d$ and $\boldsymbol{\nu} \in \mathcal{M}(Y)^d$.

(i) We denote by $\boldsymbol{\mu} \succeq_{\mathrm{h}} \boldsymbol{\nu}$ ("h" stands for *heterogeneity*) if there exist $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ satisfying $\mu \gg \boldsymbol{\mu}$, $\nu \gg \boldsymbol{\nu}$ and $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\mu \geqslant_{\mathrm{icx}} \mathrm{d}\boldsymbol{\nu}/\mathrm{d}\nu$, where $\geqslant_{\mathrm{icx}}$ is the multivariate increasing convex order.[3]

(ii) We say that $\boldsymbol{\mu}$ is *jointly atomless* if there exist $\mu \gg \boldsymbol{\mu}$ and a random variable ξ such that under μ, ξ is continuously distributed and independent of $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\mu$.

We note that $\boldsymbol{\mu} \succeq_{\mathrm{h}} \boldsymbol{\nu}$ implies $\boldsymbol{\mu}(X) \geqslant \boldsymbol{\nu}(Y)$ by taking a linear function $f(x_1,\dots,x_d) = x_j$ for $j \in [d]$ in the definition of the increasing convex order. Hence, it makes sense to discuss the set $\mathcal{K}(\boldsymbol{\mu},\boldsymbol{\nu})$ under this condition.

Remark 3.2. In Definition 3.1, if $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$, then $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\mu \geqslant_{\mathrm{icx}} \mathrm{d}\boldsymbol{\nu}/\mathrm{d}\nu$ is equivalent to $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\mu \geqslant_{\mathrm{cx}} \mathrm{d}\boldsymbol{\nu}/\mathrm{d}\nu$, where $\geqslant_{\mathrm{cx}}$ is the multivariate convex order. It is shown in Shen et al. (2019) that in this case, μ and ν can be safely chosen as $\bar{\mu}$ and $\bar{\nu}$.

[2] As we will see, this is necessary for the Kantorovich reformulation to make sense.

[3] This means $\int f(\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\mu)\mathrm{d}\mu \geqslant \int f(\mathrm{d}\boldsymbol{\nu}/\mathrm{d}\nu)\mathrm{d}\nu$ for all increasing convex $f : \mathbb{R}^n \to \mathbb{R}$ such that the integrals are well-defined.

Remark 3.3. Shen et al. (2019) called the notion of joint non-atomicity in Definition 3.1 as "conditional non-atomicity". We choose the term "joint non-atomicity" because this notion is indeed a collective property of $(\mu_1, \dots, \mu_d)$, and it is stronger than non-atomicity of each μ_j. There are many parallel results between non-atomicity for $d = 1$ and joint non-atomicity for $d \geqslant 2$; see Remark 4.4.

Proposition 3.4. *Let $\boldsymbol{\mu} \in \mathcal{M}(X)^d$ and $\boldsymbol{\nu} \in \mathcal{M}(Y)^d$.*

(i) The set $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is non-empty if and only if $\boldsymbol{\mu} \succeq_{\mathrm{h}} \boldsymbol{\nu}$.

(ii) Assume that $\boldsymbol{\mu}$ is jointly atomless. The set $\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is non-empty if and only if $\boldsymbol{\mu} \succeq_{\mathrm{h}} \boldsymbol{\nu}$.

Proof. The first statement is implied by Proposition 9.7.1 of Torgersen (1991) and the remarks that follow it. The second statement can be shown by the same arguments as in Theorem 3.17 of Shen et al. (2019) where $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$ is assumed. □

By Proposition 3.4, $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\mathcal{K}(\boldsymbol{\nu}, \boldsymbol{\mu})$ are both non-empty if and only if $\boldsymbol{\mu}' \overset{\mathrm{law}}{=} \boldsymbol{\nu}'$, where the laws are under $\bar{\mu}$ and $\bar{\nu}$, respectively.

Remark 3.5. To understand $\boldsymbol{\mu} \succeq_{\mathrm{h}} \boldsymbol{\nu}$ intuitively, which means that $\boldsymbol{\mu}$ is more heterogeneous than $\boldsymbol{\nu}$, one could look at some special cases (treated in Proposition 3.7 of Shen et al. (2019)), by assuming $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$. Suppose that $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is non-empty. Then $\boldsymbol{\mu}$ has identical components (smallest in $\preceq_{\mathrm{h}}$) $\Rightarrow$ so does $\boldsymbol{\nu}$; $\boldsymbol{\mu}$ has equivalent components $\Rightarrow$ so does $\boldsymbol{\nu}$; $\boldsymbol{\nu}$ has mutually singular components (largest in $\preceq_{\mathrm{h}}$) $\Rightarrow$ so does $\boldsymbol{\mu}$. Moreover, $\boldsymbol{\mu}$ has mutually singular components or $\boldsymbol{\nu}$ has identical components $\Rightarrow$ $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is non-empty.

In case $\boldsymbol{\mu}(X) \geqslant \boldsymbol{\nu}(Y)$ in which equality does not hold, simple sufficient conditions exist. For example, suppose that

$$\min_{j\in[d]}(\mu_j(X)) \geqslant \left(\max_{j\in[d]} \nu_j\right)(Y),$$

then $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is non-empty.[4] To see this, we may assume each μ_j is a probability measure. For $\nu := (\max_j \nu_j)/((\max_j \nu_j)(Y))$, we have that $\mathcal{K}(\boldsymbol{\mu}, (\nu, \dots, \nu))$ is non-empty since the constant kernel $x \mapsto \nu$ is in $\mathcal{K}(\boldsymbol{\mu}, (\nu, \dots, \nu))$. Then for $\kappa \in \mathcal{K}(\boldsymbol{\mu}, (\nu, \dots, \nu))$,

$$\kappa_{\#}\mu_j = \nu = \frac{\max_{j\in[d]} \nu_j}{(\max_{j\in[d]} \nu_j)(Y)} \geqslant \max_{j\in[d]} \nu_j \geqslant \nu_j.$$

This shows $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$.

We record an immediate corollary of Proposition 3.4 for future use, which can also be shown by directly using definition.

[4] For a collection of (signed) measures μ_j, $j \in J$ on X, their maximum (or supremum) is defined as $\sup_{j\in J} \mu_j(A) = \sup\{\sum_{j\in J} \mu_j(A_j) \mid \bigcup_{j\in J} A_j = A \text{ and } A_j \text{ are disjoint}\}$ for $A \subseteq X$. Moreover, the positive part of μ, denoted by μ_+, is $\max\{\mu, 0\}$ where 0 is the zero measure.

Corollary 3.6. *Suppose that $\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\eta}$ are $\mathbb{R}^d$-valued probability measures on Polish spaces such that $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\mathcal{K}(\boldsymbol{\nu}, \boldsymbol{\eta})$ are non-empty. Then $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\eta})$ is non-empty.*

Proposition 3.4 can also be applied to give a necessary condition for the existence of a time-homogeneous Markov process (see Example 2.4) for centered Gaussian marginals on $\mathbb{R}$.

Proposition 3.7. *Suppose that $\mu_t = \mathrm{N}(0, \sigma_t^2)$, $\sigma_t > 0$, $t = 1, \dots, T$. For the existence of a transport from $(\mu_1, \dots, \mu_{T-1})$ to $(\mu_2, \dots, \mu_T)$, it is necessary that the mapping $t \mapsto \sigma_t$ on $\{1, \dots, T\}$ is increasing log-concave or decreasing log-convex. If $T = 3$, this condition is also sufficient.*

The necessary condition in Proposition 3.7 is not sufficient for $T > 3$. See Appendix A for a counterexample. In the case $T = 3$, the Markov process in Proposition 3.7 can be realized by an AR(1) process with Gaussian noise.

3.2 Some simple lower bounds for on the minimum cost

We collect some lower bounds for the infimum cost based only on classic ($d = 1$) transports. Since every $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ transports each μ_j to cover ν_j, it must transport each $\boldsymbol{\lambda} \cdot \boldsymbol{\mu}$ to cover $\boldsymbol{\lambda} \cdot \boldsymbol{\nu}$ for $\boldsymbol{\lambda} \in \mathbb{R}_+^d$. Denoting by Δ_d the standard simplex in $\mathbb{R}^d$, we have

$$\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu}) \subseteq \bigcap_{\boldsymbol{\lambda} \in \Delta_d} \mathcal{K}(\boldsymbol{\lambda} \cdot \boldsymbol{\mu}, \boldsymbol{\lambda} \cdot \boldsymbol{\nu}).$$

Therefore, we obtain

$$\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa) \geqslant \sup_{\boldsymbol{\lambda} \in \Delta_d} \inf_{\kappa \in \mathcal{K}(\boldsymbol{\lambda} \cdot \boldsymbol{\mu}, \boldsymbol{\lambda} \cdot \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa). \tag{6}$$

In particular, if $\kappa \in \mathcal{K}(\bar{\mu}, \bar{\nu})$ is an optimal transport from $\bar{\mu}$ to $\bar{\nu}$ and $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, then κ is also an optimal transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$. However, as we will see in Example 3.10, the inequality (6) is not sharp in general.

We record yet another lower bound for the minimum cost as an application of the kernel formulation. For simplicity we consider the balanced setting. The following proposition follows intuitively by observing that, for example in case $d = 2$, the parts where $\nu_1 \geqslant \nu_2$ must be transported from the parts where $\mu_1 \geqslant \mu_2$ (see Figure 2).

Proposition 3.8. *Suppose that $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$, and for each $x \in X$, $c(x, y) = 0$ for some $y \in Y$. Then*

$$\begin{aligned} &\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa) \\ &\geqslant \max_{i,j \in [d]} \left(\inf_{\kappa \in \mathcal{K}((\mu_i - \mu_j)_+, (\nu_i - \nu_j)_+)} \mathcal{C}_\eta(\kappa) + \inf_{\kappa \in \mathcal{K}((\mu_j - \mu_i)_+, (\nu_j - \nu_i)_+)} \mathcal{C}_\eta(\kappa) \right). \end{aligned} \tag{7}$$

In particular, if $(\mu_i - \mu_j)_+(X) < (\nu_i - \nu_j)_+(Y)$ for some $i, j \in [d]$, then both sides of (7) are equal to $+\infty$.

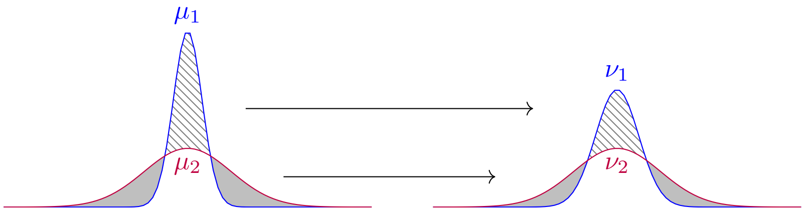


Figure 2: Part of the shaded region $(\mu_1 - \mu_2)_+$ on the left is transported to cover all of the shaded region $(\nu_1 - \nu_2)_+$ on the right; similarly, part of the gray region $(\mu_2 - \mu_1)_+$ is transported to cover all of the gray region $(\nu_2 - \nu_1)_+$.

Note that the quantities on the right-hand side of (7) arise from two separate one-dimensional transport problems.

If $\mu_1, \dots, \mu_d$ have mutually disjoint supports (in particular, if $d = 1$), then the simultaneous transport problem is reduced to d classic transport problems and the optimal cost is the sum of corresponding optimal costs. In this case, both (6) and (7) are sharp.

3.3 Peculiarities of the simultaneous transport

We consider a few simple but instructional examples showing that simultaneous transport is very different from classic transport ($d = 1$). We will focus on the balanced case (i.e., $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$) for simplicity.

We first provide some immediate observations which help to explain some novel features of simultaneous transport and its connection to the classic optimal transport problem. Denote $\mu_j'(x), \nu_j'(y)$ the corresponding Radon-Nikodym derivatives of μ_j, ν_j with respect to $\bar{\mu}, \bar{\nu}$, respectively.

First, suppose that $\eta \sim \bar{\mu}$. If there exist measurable functions ϕ on X and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)$ on Y such that

$$c(x, y) = \phi(x) + \boldsymbol{\psi}(y)^\top \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x),$$

then all transports (should any exist) from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ have the same cost

$$\int_{X \times Y} c \,\mathrm{d}(\eta \otimes \kappa) = \int_X \phi \,\mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu}, \tag{8}$$

because $\kappa_\# \boldsymbol{\mu} = \boldsymbol{\nu}$. This extends the fact that in the case $d = 1$, the cost functions of the form $c(x, y) = \phi(x) + \psi(y)$ are trivial and can be "decomposed into marginal costs". We now have a larger class of such cost functions. If $\eta = \bar{\mu}$, then a term $\psi(y)$ for $\psi : Y \to \mathbb{R}$ can also be included in $c(x, y)$, by noting that

$$\psi(y) = \frac{1}{d} \psi(y) \mathbf{1} \cdot \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\bar{\mu}}(x).$$

Moreover, (8) also hints on how a duality result would look like in this setting, which will be discussed in Section 5.

Example 3.9. Consider $X = \mathbb{R}$ on which Borel probability measures $\boldsymbol{\mu}$ are supported and $\eta = \bar{\mu}$. Assume that μ_1' is linear in $x \in \mathbb{R}$ on the support of $\bar{\mu}$, say, equal to $ax + b$, $a \neq 0$. Let $\boldsymbol{\nu}$ be probability measures on $\mathbb{R}$ such that $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is non-empty. Consider a quadratic cost function $c(x, y) = (x - y)^2$. Then we may write

$$c(x,y) = x^2 + (ax+b)\left(-\frac{2y}{a}\right) + \left(y^2 + \frac{2by}{a}\right).$$

Therefore, for any $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$,

$$\mathcal{C}(\kappa) = \int_X x^2 \bar{\mu}(\mathrm{d}x) + \int_Y \left(y^2 + \frac{2by}{a}\right) \bar{\nu}(\mathrm{d}y) - \frac{2}{a}\int_Y y\, \nu_1(\mathrm{d}y).$$

Example 3.10. As a concrete but slightly more general example, we consider $X = Y = [0, 1]$ on which Borel probability measures $\mu_j, \nu_j,\ j = 1, 2$ are supported. Assume μ_1 has density $2x$ and μ_2 has density $2 - 2x$ with respect to Lebesgue measure on $[0, 1]$, and $\nu_1 = \nu_2$ be any identical probability measures on $[0, 1]$ such that $\nu_1((1/4, 3/4)) = 1/2$ (see Figure 3). Thus the Radon-Nikodym derivatives are $\mu_1'(x) = 2x$ and $\nu_1'(y) = 1$. Denote the set $A = (1/4, 3/4) \times ([0, 1/4) \cup (3/4, 1])$ and consider the cost function

$$c(x, y) = (x - y)^2 + \alpha \mathbb{1}_A,\ \ \alpha > 0.$$

For any $s \in [0, 1/2]$ and any S such that $\nu_1(S) = 2s$, the transport kernel

$$\kappa(x; B) := \frac{\nu_1(B \cap S)}{\nu_1(S)} \mathbb{1}_{\{x\in(s,1-s)\}} + \frac{\nu_1(B \setminus S)}{\nu_1([0,1] \setminus S)} \mathbb{1}_{\{x\in[0,s]\cup[1-s,1]\}}$$

belongs to $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. In case $S = (1/4, 3/4)$ and $s = 1/4$, we denote such a transport by κ_0.

We show that κ_0 is indeed an optimal transport. Similarly as in Example 3.9, for a kernel $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ we compute its transport cost

$$\mathcal{C}(\kappa) = \frac{1}{3} + \int_0^1 (y^2 - y)\bar{\nu}(\mathrm{d}y) + \alpha(\bar{\mu} \otimes \kappa)(A) \geqslant \frac{1}{3} + \int_0^1 (y^2 - y)\bar{\nu}(\mathrm{d}y),$$

where inequality holds if and only if $(\bar{\mu} \otimes \kappa)(A) = 0$. Since by definition $\kappa_0(x; (1/4, 3/4)) = 1$ for $x \in (1/4, 3/4)$, we have $(\bar{\mu} \otimes \kappa_0)(A) = 0$. Therefore, κ_0 is an optimal transport.

Trivial as it looks, Example 3.10 provides us some interesting aspects of the simultaneous optimal transport in contrast to the classic optimal transport.

i. It is well-known that for the classic Kantorovich transport problem in $\mathbb{R}$, if the cost function is a convex function in $y - x$, then the comonotone map is always optimal (see e.g., Theorem 2.9 of Santambrogio (2015)). However, this effect no longer exists in simultaneous transport, since there may not exist an admissible comonotone map.

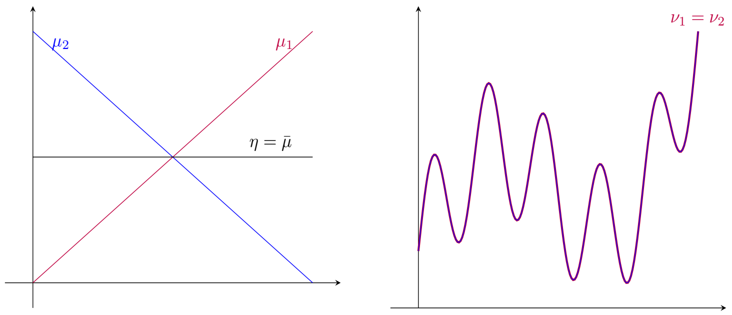


Figure 3: Densities of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ in Example 3.10.

ii. It is also easy to see that the equality in (6) may not hold, for example when ν_1 is uniform on $[0,1]$. In addition, the inequality (7) becomes trivial since it gives a lower bound 0.

After developing our theory, we discuss a few more interesting examples in Section 6.5.

4 The Kantorovich formulation and approximation results

Recall we consider d-tuples of finite measures $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$ on X and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d)$ on Y, and a reference measure $\eta \sim \bar{\mu}$. Also recall that $\boldsymbol{\mu}', \boldsymbol{\nu}'$ denote the Radon-Nikodym derivatives of $\boldsymbol{\mu}, \boldsymbol{\nu}$ with respect to $\bar{\mu}, \bar{\nu}$.

4.1 The Kantorovich formulation

Sometimes it is mathematically more convenient to adopt the Kantorovich formulation, which describes the set of all transport plans as probability measures in $\mathcal{P}(X \times Y)$. More precisely, for a probability measure $\eta \sim \bar{\mu}$, we define

$$\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}) := \{\eta \otimes \kappa \mid \kappa_\# \boldsymbol{\mu} \geqslant \boldsymbol{\nu}\}.$$

The subscript η incorporates the way we calculate costs: see (3) and (4). It is immediate that

$$\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\pi) := \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c(x,y) \pi(\mathrm{d}x, \mathrm{d}y) = \inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa). \tag{9}$$

Equivalently, we have the following reformulation for $\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$. We call this the Kantorovich reformulation, whose reasons are explained below.

Proposition 4.1. *For each $\eta \sim \bar{\mu}$, we have*

$$\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu}) = \Bigg\{ \pi \in \mathcal{P}(X \times Y) \mid \pi(A \times Y) = \eta(A) \textit{ for all } A \subseteq X \textit{ and } \int_{X\times B} \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x)\pi(\mathrm{d}x,\mathrm{d}y) \geqslant \boldsymbol{\nu}(B) \textit{ for all } B \subseteq Y \Bigg\}. \tag{10}$$

In a way similar to Proposition 4.1, in the balanced case, i.e., $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$, we have

$$\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu}) = \Bigg\{ \pi \in \mathcal{P}(X \times Y) \mid \pi(A \times Y) = \eta(A) \text{ for all } A \subseteq X \text{ and } \int_{X\times B} \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x)\pi(\mathrm{d}x,\mathrm{d}y) = \boldsymbol{\nu}(B) \text{ for all } B \subseteq Y \Bigg\}. \tag{11}$$

In particular, if $\eta = \bar{\mu}$, we denote by $\Pi(\boldsymbol{\mu},\boldsymbol{\nu}) = \Pi_{\bar{\mu}}(\boldsymbol{\mu},\boldsymbol{\nu})$, and (11) reads as

$$\Pi(\boldsymbol{\mu},\boldsymbol{\nu}) = \Bigg\{ \pi \in \mathcal{P}(X \times Y) \mid \pi(A \times Y) = \bar{\mu}(A) \text{ for all } A \subseteq X \text{ and } \int_{X\times B} \boldsymbol{\mu}'(x)\pi(\mathrm{d}x,\mathrm{d}y) = \boldsymbol{\nu}(B) \text{ for all } B \subseteq Y \Bigg\}. \tag{12}$$

It seems worthwhile to explain the similarities and differences of (12) compared to the classic definition $\Pi(\mu,\nu)$ in the case $d = 1$ (see (14) below). First, by summing over and normalizing the second constraint in (12), we see that π is a transport from $\bar{\mu}$ to $\bar{\nu}$. Thus, one may think of $\pi(A \times B)$ as the amount of $\bar{\mu}$-mass moving from A to B. With $j \in [d]$ fixed, the second constraint in (12) means that the mass sent from the contribution of μ_j covers exactly the corresponding portion of ν_j in Y.

We can reformulate (12) as

$$\Pi(\boldsymbol{\mu},\boldsymbol{\nu}) = \Bigg\{ \pi \in \mathcal{P}(X \times Y) \mid \int_{X\times Y} f(x)\pi(\mathrm{d}x,\mathrm{d}y) = \int_X f \,\mathrm{d}\bar{\mu} \text{ and } \int_{X\times Y} \boldsymbol{\mu}'(x)g(y)\pi(\mathrm{d}x,\mathrm{d}y) = \int_Y g \,\mathrm{d}\boldsymbol{\nu} \text{ for all measurable } f,g \Bigg\}. \tag{13}$$

In the case $d = 1$, our formulation coincides with the classic Kantorovich formulation, where the admissible transports are defined as

$$\tilde{\Pi}(\mu,\nu) := \{\pi \in \mathcal{P}(X \times Y) \mid \text{for all } A \subseteq X,\ B \subseteq Y,\ \pi(A \times Y) = \mu(A) \text{ and } \pi(X \times B) = \nu(B)\}. \tag{14}$$

In some sense, one can also recover transports in $\tilde{\Pi}(\mu_j,\nu_j)$ from $\Pi(\boldsymbol{\mu},\boldsymbol{\nu})$. For example, taking $f(x) = \mathbb{1}_{\{x\in A\}}\mu_j'(x)$ and $g(y) = \mathbb{1}_{\{y\in B\}}$ in (13), we have for any $j \in [d]$, the measure $\mu_j'(x)\pi(\mathrm{d}x,\mathrm{d}y)$ belongs to $\tilde{\Pi}(\mu_j,\nu_j)$.

Unlike the classic Kantorovich optimal transport problem in the case $d = 1$, the simultaneous transport problem is not symmetric with respect to the measures $\boldsymbol{\mu}, \boldsymbol{\nu}$, as expected from Proposition 3.4. It seems unlikely that $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ can be defined in a similar way as (14) using only projections of measures. The fact that the classic Kantorovich formulation uses projections and is symmetric, is nothing more than a nice consequence of the kernel formulation and does not reflect the general structure.

4.2 Equivalence between Monge and Kantorovich costs

Below, we prove that under suitable conditions, the set of transport maps and plans have the same infimum cost. This serves as an extension of Theorem 2.1 of Ambrosio (2003) in the case $d = 1$ and we also assume for simplicity that $\bar{\mu}, \bar{\nu}$ have compact supports. As expected from Proposition 3.4, joint non-atomicity plays an important role since it guarantees the existence of Monge transports.

We first prove the following more general result using the kernel formulation. Observe that for a Monge transport $T \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$, we can associate a kernel $\kappa_T \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ defined by $\kappa_T(x; B) := \mathbb{1}_{\{T_n(x) \in B\}}$. In view of (3) and (4), they have the same transport cost.

Theorem 4.2. *Let $\eta \sim \bar{\mu}$. Suppose that X, Y are compact spaces on which $\boldsymbol{\mu}, \boldsymbol{\nu}$ are supported, $\boldsymbol{\mu}$ is jointly atomless, and c is continuous. Then the transport plans and transport maps admit the same infimum cost. That is,*

$$\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa) = \inf_{T \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(T).$$

Combining with (9) yields the following.

Corollary 4.3. *Consider a reference measure $\eta \sim \bar{\mu}$. Suppose that X is a compact space on which $\boldsymbol{\mu}, \boldsymbol{\nu}$ are supported, $\boldsymbol{\mu}$ is jointly atomless, and c is continuous, then Monge and Kantorovich transport costs have the same infimum value. That is,*

$$\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\pi) = \inf_{T \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(T).$$

The proof of Theorem 4.2 follows a similar path as the classic result in the case $d = 1$, except that we need a few new lemmas on joint non-atomicity. Recall from the classic proof that non-atomicity allows us to approximate a transport plan using a transport map on each small piece of X. In our setting, we need joint non-atomicity to achieve this; see Proposition 3.4.

Remark 4.4. Heuristically, there is a parallel between non-atomicity in the classic setting and joint non-atomicity in our setting. For example,

i. Under joint non-atomicity, a Monge transport exists if and only if a Kantorovich transport exists (Proposition 3.4)[5]. In $d = 1$ with non-atomicity,

[5] The converse does not hold. There are examples where $\boldsymbol{\mu}$ is not jointly atomless but there exists a unique Kantorovich transport that is also Monge.

this equivalence also holds, although a Kantorovich transport between μ and ν exists as soon as μ and ν have the same mass.

ii. Marginal non-atomicity is equivalent to the existence of a uniform random variable and joint non-atomicity is equivalent to the existence of a uniform random variable independent of a σ-field (Lemma B.2).

iii. The joint non-atomicity condition enables us to conclude Monge and Kantorovich problems have the same infimum (Corollary 4.3), which is true in the case $d = 1$ given marginal non-atomicity.

4.3 Connecting the balanced and unbalanced settings

So far we have discussed simultaneous optimal transport in the unbalanced setting. In real applications such as the setting of Example 2.1, it likely holds that $\boldsymbol{\mu}(X) \geqslant \boldsymbol{\nu}(Y)$ with strict inequality in some components. For instance, in an economy, the total demand for each product may be approximately 95% of the total supply, leading to $\boldsymbol{\nu}(Y) \approx 0.95 \times \boldsymbol{\mu}(X)$.

As we will see in the subsequent sections, results on duality, equilibria, uniqueness and the Wasserstein distance will be obtained in the setting of balanced transport, since the balanced setting has much richer mathematical structure than the unbalanced setting.

Nevertheless, we show below that the balanced setting of simultaneous transport can be used as an approximation of the unbalanced setting. A special situation is when $\boldsymbol{\nu}(Y) \approx (1-\varepsilon) \times \boldsymbol{\mu}(X)$ for a small $\varepsilon > 0$, which is more realistic in applications.

Suppose that $\boldsymbol{\nu}^n \leqslant \boldsymbol{\nu}$ for $n \in \mathbb{N}$ and $\boldsymbol{\nu}^n \to \boldsymbol{\nu}$ weakly as $n \to +\infty$. By definition, $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu}) \subseteq \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu}^n)$, which means that each transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ is also a transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}^n$. Moreover, under a continuity assumption, the minimum transport cost from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ is the limit of that from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}^n$. Therefore, an optimal transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ can be seen as an approximation of an optimal transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}^n$. Note that $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$ is not needed for this continuity result.

Proposition 4.5. *Suppose that X, Y are compact Polish spaces, $\boldsymbol{\mu} \in \mathcal{M}(X)^d$, and $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\eta$ and c are continuous. Suppose that $(\boldsymbol{\nu}^n)_{n\in\mathbb{N}} \subseteq \mathcal{M}(Y)^d$ is a sequence of measures converging weakly to $\boldsymbol{\nu} \in \mathcal{M}(Y)^d$ such that $\boldsymbol{\nu}^n \leqslant \boldsymbol{\nu}$ for each $n \in \mathbb{N}$. Then*

$$\lim_{n\to\infty} \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}^n)} \mathcal{C}(\pi) = \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\pi).$$

Proposition 4.5 provides a link between two settings, allowing us to use results in the balanced setting to approximate the unbalanced setting. Starting from the next section, we concentrate on the balanced setting.

5 Duality and an equilibrium model

Consider $\mathbb{R}^d$-valued measures $\boldsymbol{\mu}, \boldsymbol{\nu}$ on Polish spaces X, Y satisfying $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$ (e.g., when they are probability measures), a reference probability $\eta \sim \bar{\mu}$, and $\boldsymbol{\mu}, \boldsymbol{\nu}$ are absolutely continuous with respect to $\bar{\mu}, \bar{\nu}$ with densities $\boldsymbol{\mu}'$ on X and $\boldsymbol{\nu}'$ on Y respectively. Also recall that (11) is the set of all transport plans from the vector-valued measure $\boldsymbol{\mu}$ to the vector-valued measure $\boldsymbol{\nu}$.

5.1 Duality for compact spaces and for $\mathbb{R}^N$

We first give a duality theorem for simultaneous optimal transport on compact Polish spaces. For the readers' convenience we have provided a detailed proof in Appendix C.1. Our proof is adapted from Santambrogio (2015), Section 1.6.3, where the convex analysis trick originates from Bouchitte and Buttazzo (2001).

Theorem 5.1. *Suppose that X, Y are compact, $\eta \sim \bar{\mu}$, and $c : X \times Y \to [0, +\infty]$ is lower semi-continuous.*[6] *Duality holds as*

$$\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c \, \mathrm{d}\pi = \sup_{(\phi, \boldsymbol{\psi}) \in \Phi_c} \int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu}, \tag{15}$$

where

$$\Phi_c = \left\{ (\phi, \boldsymbol{\psi}) \in C(X) \times C(Y)^d \mid \phi(x) + \boldsymbol{\psi}(y) \cdot \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x) \leqslant c(x, y) \right\}.$$

If moreover $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\eta$ is continuous, the infimum in (15) is attained.

In the case $d = 1$ and $\eta = \mu$, this recovers Theorem 1.3 in Villani (2003) under the assumption of compactness.

If $\mathrm{d}\eta/\mathrm{d}\bar{\mu}$ is bounded (e.g., when $\eta = \bar{\mu}$), even if X, Y are not compact, we still have $\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$ is tight and hence weakly relatively compact. This follows from the definition of tightness and

$$\bar{\nu}(B) = \int_{X \times B} \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\eta}(x) \pi(\mathrm{d}x, \mathrm{d}y) \geqslant \left(\sup_{x \in X} \frac{\mathrm{d}\eta}{\mathrm{d}\bar{\mu}}(x) \right)^{-1} \pi(X \times B).$$

Furthermore, if $\eta = \bar{\mu}$, $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is weakly compact if $\boldsymbol{\mu}'$ is assumed to be continuous, as can be seen by taking limits in (13). In particular, we will see in the proof in Section C.1 that in this case the attainability of the infimum in (15) does not require the spaces X, Y to be compact.

In the case where $\eta \ll \bar{\mu}$ but $\bar{\mu} \not\ll \eta$, we still have the lower bound

$$\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c(x, y) \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y) \geqslant \sup_{(\phi, \boldsymbol{\psi}) \in \Phi_c} \int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu},$$

where

$$\Phi_c := \left\{ (\phi, \boldsymbol{\psi}) \in C(X) \times C^d(Y) \mid \phi(x) \mathrm{d}\eta(x) + \boldsymbol{\psi}(y)^\top \mathrm{d}\boldsymbol{\mu}(x) \leqslant c(x, y) \mathrm{d}\eta(x) \right\}.$$

[6] Recall that a function f is lower semi-continuous if and only if for any $y \in \mathbb{R}$, $\{\mathbf{x} \mid f(\mathbf{x}) > y\}$ is open.

This is because for $(\phi, \boldsymbol{\psi}) \in \Phi_c$ and $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$,

$$\begin{aligned}\int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu} &= \int_{X \times Y} \phi(x) \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y) + \int_{X \times Y} \boldsymbol{\psi}(y)^\top \boldsymbol{\mu} \otimes \kappa(\mathrm{d}x, \mathrm{d}y) \\ &\leqslant \int_{X \times Y} c(x,y) \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y).\end{aligned}$$

In more generality, we also have the same duality result for Euclidean spaces, where no compactness is assumed.

Theorem 5.2. *Suppose that $X = Y = \mathbb{R}^N$ and $\mathrm{d}\eta/\mathrm{d}\bar{\mu}$ is bounded. Then duality formula (15) holds.*

Our proof uses specific properties of the Euclidean space and does not generalize to generic Polish spaces. The idea is to shrink the supports of $\bar{\mu}, \bar{\nu}$ to bounded sets using a homeomorphism of $\mathbb{R}^N$. This is illustrated by the following Figure 4.

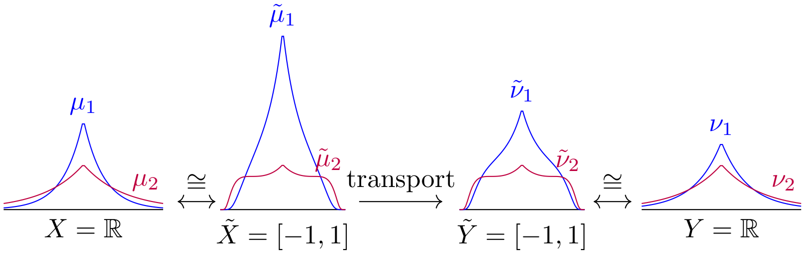


Figure 4: To analyze the transport from $(X, \boldsymbol{\mu})$ to $(Y, \boldsymbol{\nu})$, we transform homeomorphically via the arctan function to shrink the measures to $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}}$ supported on the compact set $\tilde{X} = \tilde{Y} = [-1, 1]$.

In general, some technical steps are needed to ensure that after the homeomorphism, the new cost function is still lower semi-continuous. Having reduced to compact supports, we apply Theorem 5.1 to conclude the proof.

5.2 A labour market equilibrium model

We discuss a matching equilibrium model in a labour market via simultaneous transport, similar to that in the classic transport setting. First, we state the relevant version of the duality formula in Theorem 5.1. Suppose that $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$ is a vector of probabilities on X, $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d)$ is a vector of probabilities on Y, and $\eta \sim \bar{\mu}$. Assume that X and Y are compact, and $g : X \times Y \to [-\infty, +\infty)$ is upper semi-continuous. The duality formula, with a maximization in place of a minimization in (15), is

$$\sup_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} g \, \mathrm{d}\pi = \inf_{(\phi, \boldsymbol{\psi}) \in \Phi_g} \int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \, \mathrm{d}\boldsymbol{\nu}, \tag{16}$$

where

$$\Phi_g := \left\{ (\phi, \boldsymbol{\psi}) \in C(X) \times C^d(Y) : \ \phi(x) + \boldsymbol{\psi}(y) \cdot \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x) \geqslant g(x,y) \right\}.$$

Let $x \in X$ represent worker labels and $y \in Y$ represent firms. The interpretation of η, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is given below.

1. η is the distribution of the workers, i.e., how much proportion of the workers are labelled with $x \in X$. In a discrete setting of n workers in total, it would not hurt to imagine that $\eta(x) = 1/n$; i.e., each worker has a different label.

2. There are d types of skills in this production problem. Workers with the same label have the same skills. The distribution μ_i describes the supply of type-i skill provided by the workers. In a discrete setting, $\mu_i(x)$ is the type-i skill provided by each worker label x. We denote by $\boldsymbol{\mu}' = \mathrm{d}\boldsymbol{\mu}/\mathrm{d}\eta$, that is, the (per-worker) skill vector.

3. The distribution ν_i describes the demand of type-i skill from the firms. In a discrete setting, $\nu_i(y)$ is the type-i skill demanded by each firm y.

Assume that the total demand and the total supply of skills are equal, and hence both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are normalized to have total mass of $(1, \dots, 1)$. A matching between the workers and the firms is an element π of $\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$. Let $g(x,y)$ represent the production of firm y hiring worker x per unit of worker. For a given matching π, the total production in the economy is $\int g \, \mathrm{d}\pi$.

Take two arbitrary functions $w : X \to \mathbb{R}$ and $\mathbf{p} : Y \to \mathbb{R}^d$. As usual, $w(x)$ represents the wage of worker x. The function $\mathbf{p}$ represents the profit-per-skill vector of firm y in the following sense: if firm y employs a skill vector $\mathbf{q} \in \mathbb{R}^d_+$, then the total profit of the firm is $\mathbf{p}(y) \cdot \mathbf{q}$. Taking $\mathbf{q} = \boldsymbol{\mu}'(x)$, the profit generated from hiring each worker x is $\mathbf{p}(y) \cdot \boldsymbol{\mu}'(x)$. The total profit of all firms is

$$\int_{X \times Y} \mathbf{p}(y) \cdot \boldsymbol{\mu}'(x) \pi(\mathrm{d}x, \mathrm{d}y) = \int_Y \mathbf{p}^\top \mathrm{d}\boldsymbol{\nu},$$

which follows from the definition of π.

For worker x, their objective is to choose a firm to maximize their wage, that is

$$\max_{y \in Y} \left\{ g(x,y) - \mathbf{p}(y) \cdot \boldsymbol{\mu}'(x) \right\}.$$

For firm y, its objective is to hire workers to maximize its profit, that is

$$\max_{x \in X} \left\{ g(x,y) - w(x) \right\}.$$

For a social assignment $(w, \mathbf{p})$ and a matching $\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$, an *equilibrium* is attained if

(a) the social assignment is optimal, that is

$$w(x) = \max_{y \in Y} \{g(x,y) - \mathbf{p}(y) \cdot \boldsymbol{\mu}'(x)\}$$

and

$$\mathbf{p}(y) \cdot \boldsymbol{\mu}'(x_y) = g(x_y, y) - w(x_y) = \max_{x \in X} \{g(x,y) - w(x)\}\,.$$

(b) the total production in the economy is at least as large as the total wage plus the total profit, that is,

$$\int_{X \times Y} g \, \mathrm{d}\pi \geqslant \int_X w \, \mathrm{d}\eta + \int_Y \mathbf{p}^\top \mathrm{d}\boldsymbol{\nu}. \tag{17}$$

Since (a) implies

$$w(x) + \mathbf{p}(y) \cdot \boldsymbol{\mu}'(x) \geqslant g(x,y) \tag{18}$$

for all $x \in X$ and $y \in Y$, integrating (18) with respect to π gives

$$\int_X w \, \mathrm{d}\eta + \int_Y \mathbf{p}^\top \mathrm{d}\boldsymbol{\nu} \geqslant \int_{X \times Y} g \, \mathrm{d}\pi,$$

and hence, (17) has to hold as an equality, and this implies the duality (16). Again, an equilibrium exists if and only if duality holds with both the infimum and the supremum attained. In the finite-state setting, the above attainability is automatic.

6 Two-way transport and the Wasserstein distance

The aim of this section is to propose a notion of the Wasserstein distance between $\mathbb{R}^d$-valued probability measures on a Polish space X equipped with a metric ρ, using the optimal cost in simultaneous transport. Throughout this section, we consider the reference measure $\eta = \bar{\mu}$ and a number $p \geqslant 1$.

6.1 Defining the Wasserstein quasi-metric

Let us first recall the classic definition of the Wasserstein distance. Consider a Polish space (X, ρ) and define

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) \mid \int_X \rho(x_0, x)^p \mu(\mathrm{d}x) < +\infty \text{ for some } x_0 \in X \right\}.$$

The Wasserstein distance between probability measures $\mu, \nu \in \mathcal{P}_p(X)$ is the metric given by

$$\mathcal{W}_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} \rho(x,y)^p \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}.$$

The space $(\mathcal{P}_p(X), \mathcal{W}_p)$ is again a Polish space.

For $\mathbb{R}^d$-valued measures, we may similarly define

$$\mathcal{P}(X)^d_{p,\rho} := \left\{ \boldsymbol{\mu} \in \mathcal{P}(X)^d \mid \text{there exists } x_0 \in X, \int_X \rho(x, x_0)^p \bar{\mu}(\mathrm{d}x) < +\infty \right\}.$$

A straightforward way to define a Wasserstein quasi-metric is to let for $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(X)^d_{p,\rho}$,

$$\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) := \left(\inf_{\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times X} \rho(x, y)^p \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}.$$

That $\mathcal{W}_p$ satisfies the triangle inequality is straightforward to check. However, due to the asymmetry of $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ mentioned in Section 4.1,[7] $\mathcal{W}_p$ is not a metric, and it may even take the value $+\infty$. To overcome this, a natural attempt is to consider a subset of $\mathbb{R}^d$-valued probability measures $\mathcal{E} \subseteq \mathcal{P}(X)^d_{p,\rho}$ such that for any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}$, $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{W}_p(\boldsymbol{\nu}, \boldsymbol{\mu}) < +\infty$. A first observation is that, we must need that both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty. It turns out that in this case, the problem is completely solved by reducing it to a collection of classic optimal transport problems with $d = 1$; see Theorem 6.4. To achieve this goal, we first prove some results on the uniqueness of the transport, which is of its own interest.

6.2 Uniqueness of the simultaneous transport

We are interested in the uniqueness of the transport in $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, and using the Kantorovich reformulation it suffices to consider the case $\eta = \bar{\mu}$. The reason for considering this is twofold:

(i) If the transport is unique, then it is automatically the optimal transport for any cost function c.

(ii) As we will see in Section 6.3, the results in this section eventually provide a complete solution of the optimal transport problem when both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty (equivalently, $\boldsymbol{\mu}' \overset{\text{law}}{=} \boldsymbol{\nu}'$ as random variables on measure spaces $(X, \bar{\mu})$ and $(Y, \bar{\nu})$), in the sense that it is decomposed into a collection of individual classic optimal transport problems. This in turn gives a notion of Wasserstein distances between $\mathbb{R}^d$-valued probability measures.

We consider the situation where $\boldsymbol{\mu}, \boldsymbol{\nu}$ are $\mathbb{R}^d$-valued probability measures on Polish spaces X, Y. Our main result is the following.

Theorem 6.1. *Suppose that both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty and $\boldsymbol{\mu}'$ is injective on the support of $\bar{\mu}$.*

(i) There exist a unique $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and a unique $\tilde{\pi} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$.

[7] For example, $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ may be non-empty while $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ is empty, so that the transport cost from $\boldsymbol{\nu}$ to $\boldsymbol{\mu}$ is $+\infty$.

(ii) We have $\pi(A \times B) = \tilde{\pi}(B \times A)$ *for all* $(A, B) \in \mathcal{B}(X) \times \mathcal{B}(Y)$.

(iii) It holds that

$$\pi\left(\{(x, y) \mid \boldsymbol{\mu}'(x) \neq \boldsymbol{\nu}'(y)\}\right) = 0. \tag{19}$$

(iv) The measure $\tilde{\pi}$ *is induced by the Monge transport* $(\boldsymbol{\mu}')^{-1} \circ \boldsymbol{\nu}'$.

Remark 6.2. By Proposition 3.4, a necessary and sufficient condition that both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty is that $\boldsymbol{\mu}' \overset{\text{law}}{=} \boldsymbol{\nu}'$ as random variables on measure spaces $(X, \bar{\mu})$ and $(Y, \bar{\nu})$. Moreover, a sufficient condition of $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ being non-empty is that there exist $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and a measurable function $h(y)$ such that $\pi(\{(h(y), y) \mid y \in Y\}) = d$. Indeed, $h \in \mathcal{T}(\boldsymbol{\nu}, \boldsymbol{\mu})$ since for $A \subseteq X$,

$$\boldsymbol{\nu}(h^{-1}(A)) = \int_{X \times h^{-1}(A)} \boldsymbol{\mu}'(x) \pi(\mathrm{d}x, \mathrm{d}y) = \int_{A \times Y} \boldsymbol{\mu}'(x) \pi(\mathrm{d}x, \mathrm{d}y) = \boldsymbol{\mu}(A),$$

where the last step follows from (13).

In the language of σ-fields, an equivalent condition of $\boldsymbol{\mu}'$ being injective on the support X of $\boldsymbol{\mu}$ is that $\mathcal{B}(X) = \sigma(\mathrm{d}\mu_1/\mathrm{d}\bar{\mu}, \dots, \mathrm{d}\mu_d/\mathrm{d}\bar{\mu})$. Intuitively, $\boldsymbol{\mu}$ is "jointly degenerate", which in particular implies $\boldsymbol{\mu}$ cannot be jointly atomless.

The condition that $\boldsymbol{\mu}'(x)$ is injective is necessary, because if not, one may "exchange" the parts where $\boldsymbol{\mu}'$ are equal. For example, if each $\mu_j = \nu_j$ is supported on $[-1, 1]$ and symmetric around 0, then both transports $x \mapsto x$ and $x \mapsto -x$ belong to $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$. In other words, the symmetry of $\boldsymbol{\mu}$ prevents the transport from being unique. This observation will be studied in detail in Section 6.3.

The analogue of Theorem 6.1 is trivial in the case $d = 1$, since $\mathrm{d}\mu/\mathrm{d}\mu = 1$ is never injective unless X consists of a single point, in which case the transport is of course unique.

Example 6.3. Let $X = Y = [0, 1]$, $\mu_2 = \nu_2$ be Lebesgue, μ_1 have density $2x$, and ν_1 have density $|2 - 4x|$ with respect to Lebesgue measure on $[0, 1]$. A transport is given by the stochastic kernel $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ where $\kappa : [0, 1] \to \mathcal{P}([0, 1])$ is given by

$$\kappa(x) = \frac{1}{2}\left(\delta_{(1+x)/2} + \delta_{(1-x)/2}\right).$$

See Figure 5 for an illustration. This kernel also gives, by Remark 6.2, that $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ is non-empty. Since μ_1' is strictly increasing, it follows from Theorem 6.1 that κ is the unique transport. In particular, we recover Example 3.8 in Shen et al. (2019) that there is no Monge transport map from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$.

Although we will prove Theorem 6.1 first, it can be derived from Theorem 6.4 below. We discuss some intuition of the proof after presenting Theorem 6.4.

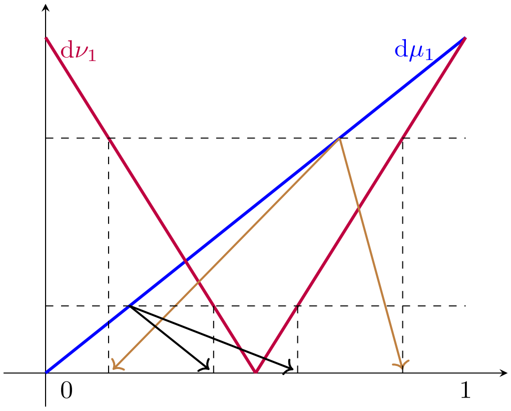


Figure 5: An illustration of Example 6.3

6.3 Explicit solutions for the two-way transport

Define an equivalence relation $\simeq$ among $\mathbb{R}^d$-valued measures as follows: $\boldsymbol{\mu} \simeq \boldsymbol{\nu}$ if $\bar{\mu} \circ (\boldsymbol{\mu}')^{-1} = \bar{\nu} \circ (\boldsymbol{\nu}')^{-1}$ (or equivalently, both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are nonempty). We denote by $\mathcal{E}_P$ the equivalence class under $\simeq$ where $P = \bar{\mu} \circ (\boldsymbol{\mu}')^{-1}$ for each $\boldsymbol{\mu} \in \mathcal{E}_P$ being a probability measure on $\mathbb{R}^d_+$. The transitivity of $\simeq$ follows from Corollary 3.6.

Consider $\boldsymbol{\mu} \in \mathcal{P}(X)^d$ and $\boldsymbol{\nu} \in \mathcal{P}(Y)^d$ such that $\boldsymbol{\mu} \simeq \boldsymbol{\nu}$. By the disintegration theorem, there exist measures $\{\mu_{\mathbf{z}}\}_{\mathbf{z} \in \mathbb{R}^d_+}$ such that

$$\mu_{\mathbf{z}}(X \setminus A_{\mathbf{z}}) := \mu_{\mathbf{z}}\left(X \setminus (\boldsymbol{\mu}')^{-1}(\mathbf{z})\right) = 0$$

and for any Borel measurable function $f : X \to [0, \infty)$,

$$\int_X f(x)\bar{\mu}(\mathrm{d}x) = \int_{\mathbb{R}^d_+} \int_{A_{\mathbf{z}}} f(x)\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}).$$

Moreover, the family of measures $\{\mu_{\mathbf{z}}\}_{\mathbf{z} \in \mathbb{R}^d_+}$ is uniquely determined for P-a.s. $\mathbf{z} \in \mathbb{R}^d_+$. Similarly for $\mathbf{z} \in \mathbb{R}^d_+$ we define $B_{\mathbf{z}} \subseteq Y$ and a probability measure $\nu_{\mathbf{z}}$ on Y. We also write the minimum cost as

$$\mathcal{I}_c(\boldsymbol{\mu}, \boldsymbol{\nu}) := \inf_{\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c(x, y)\pi(\mathrm{d}x, \mathrm{d}y) = \inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\kappa).$$

Theorem 6.4. *Suppose that c is continuous. For $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}_P$ and $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, the following are equivalent:*

(i) κ is an optimal transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$;

(ii) κ is an optimal transport from $\mu_{\mathbf{z}}$ to $\nu_{\mathbf{z}}$ for P-a.s. $\mathbf{z}$;

(iii) we have

$$\mathcal{C}_{\bar{\mu}}(\kappa) = \int_{\mathbb{R}_+^d} \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) P(\mathrm{d}\mathbf{z}). \tag{20}$$

In particular,

$$\mathcal{I}_c(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int_{\mathbb{R}_+^d} \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) P(\mathrm{d}\mathbf{z}).$$

Remark 6.5. That the right-hand side of (20) is indeed well-defined will be discussed in the proof using a measure selection argument.

A few comments are in place. Roughly speaking, two-way transports exist if and only if each transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ (provided it exists) can be inverted to produce a transport from $\boldsymbol{\nu}$ to $\boldsymbol{\mu}$. This inversion is not in general possible, because multiple points with different Radon-Nikodym derivatives may be transported to the same point in the destination, while any inversion transports back with the same Radon-Nikodym derivative as the destination point; see Figure 1 (a).

If both X, Y are discrete, the two-way transports exist if and only if for each $\mathbf{z} \in \mathbb{R}_+^d$,

$$\sum_{x \in X:\ \boldsymbol{\mu}'(x) = \mathbf{z}} \bar{\mu}(\{x\}) = \sum_{y \in Y:\ \boldsymbol{\nu}'(y) = \mathbf{z}} \bar{\nu}(\{y\}).$$

Theorem 6.4 provides us with an explicit expression of the minimum cost $\mathcal{I}_c(\boldsymbol{\mu}, \boldsymbol{\nu})$. Intuitively, it amounts to optimizing a (possibly infinite) collection of individual classic transport problems with the same cost function. For the important case of a convex cost, i.e., $c(x, y) = h(y - x)$ with h strictly convex, existing techniques can be applied to solve these individual problems; see Gangbo and McCann (1996). In the special case where $X = Y = \mathbb{R}$ and c is continuous and strictly submodular[8] on $\mathbb{R}^2$, this transport problem is uniquely optimized by taking comonotone transport plans from $\mu_{\mathbf{z}}$ to $\nu_{\mathbf{z}}$ for each $\mathbf{z} \in \mathbb{R}_+^d$ by the Fréchet-Hoeffding theorem. We summarize this in the following corollary.

Corollary 6.6. *Suppose that $X = Y = \mathbb{R}$, $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}_P$ and c is continuous and submodular. Then*

$$\mathcal{I}_c(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int_{\mathbb{R}_+^d} \int_0^1 c\left(F_{\mathbf{z}}^{-1}(t), G_{\mathbf{z}}^{-1}(t)\right) \mathrm{d}t \, P(\mathrm{d}\mathbf{z}),$$

where $F_{\mathbf{z}}^{-1}, G_{\mathbf{z}}^{-1}$ are the distribution functions of $\mu_{\mathbf{z}}, \nu_{\mathbf{z}}$ respectively.[9]

Below we explain intuitive ideas behind the proof of Theorem 6.4. The general case is boosted from the case $d = 2$ and $X = \mathbb{R}$, so we will illustrate only this latter case. We have noted in Proposition 3.8 that if $(\mu_1 - \mu_2)_+(X) <$

[8] A function c on $X \times Y$ is submodular if $c(x, y) + c(x', y') \leqslant c(x, y') + c(x', y)$ whenever $x \leqslant x'$ and $y \leqslant y'$. It is strictly submodular if the above inequality is strict as soon as $(x, y) \neq (x', y')$. An example of a (strictly) submodular function on $\mathbb{R}^2$ is $(x, y) \mapsto h(y - x)$ for a (strictly) convex h.

[9] See above Lemma D.3 for the definitions of $\mu_{\mathbf{z}}, \nu_{\mathbf{z}}$.

$(\nu_1 - \nu_2)_+(Y)$, then $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is empty. Therefore, if both $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\mathcal{K}(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty, we must have $(\mu_1 - \mu_2)_+(X) = (\nu_1 - \nu_2)_+(Y)$. Moreover, let S_1 denote the set where $\mu_1 \geqslant \mu_2$ and S_2 the set where $\mu_1 < \mu_2$, and similarly $\nu_1 \geqslant \nu_2$ on T_1 and $\nu_1 < \nu_2$ on T_2, then any transport $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ must send $\boldsymbol{\mu}|_{S_1}$ to $\boldsymbol{\nu}|_{T_1}$ and $\boldsymbol{\mu}|_{S_2}$ to $\boldsymbol{\nu}|_{T_2}$. This shrinks the space where the transport can take place.

(i) The places with $\mu_1' \geqslant 1$ must be sent to the places with $\nu_1' \geqslant 1$.

(ii) The places with $\mu_1' < 1$ must be sent to the places with $\nu_1' < 1$.

See Figure 6 below.

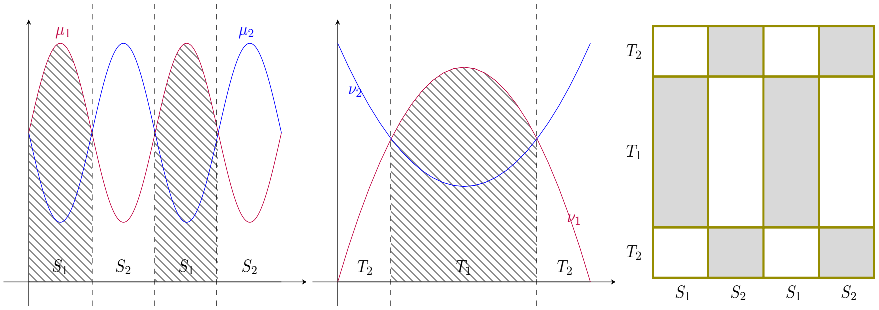


Figure 6: The transports are divided into shaded (S_1, T_1) where $\mu_1' \geqslant 1$ and $\nu_1' \geqslant 1$, and unshaded parts (S_2, T_2) where $\mu_1' < 1$ and $\nu_1' < 1$; any transport must be supported in the gray area in the right panel.

This procedure also reduces the original transport problem into two separate transport problems, and thus can be iterated. After renormalizing the restricted measures of $\boldsymbol{\mu}, \boldsymbol{\nu}$, the next step yields two points, $a \in (0,1)$ and $b \in (1,2)$, such that for I being any of the four intervals $[0,a], (a,1], (1,b], (b,2]$, the transport must send places where $\mu_1' \in I$ to places where $\nu_1' \in I$. In some sense, this procedure can be continued to produce finer and finer intervals restricting the fluctuation of Radon-Nikodym derivatives before and after the transport. In the end, the transport cannot change the Radon-Nikodym derivative at all.

In addition, some technicalities using the disintegration theorem show that these collection of transport problems, each corresponding to a value of the Radon-Nikodym derivative, can be treated independently. Moreover, we are able to not only decompose any transport $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ like this, but also "glue up" a collection of *optimal* transports[10], each corresponding a value of the Radon-Nikodym derivative, to form the *optimal* transport in $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$.

We may recover Theorem 6.1 from Theorem 6.4 since $\boldsymbol{\mu}'$ being injective means that for each $\mathbf{z}$, $\mu_{\mathbf{z}}$ is concentrated on a single point, and the transport from a Dirac mass to any measure is unique.

[10] It is not true in general that we can glue up arbitrary transports, because the resulting transport kernel may not be measurable.

6.4 The Wasserstein distance

Recall that our goal is to construct a subset of $\mathbb{R}^d$-valued probability measures $\mathcal{E} \subseteq \mathcal{P}(X)^d_{p,\rho}$ such that for any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}$, $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{W}_p(\boldsymbol{\nu}, \boldsymbol{\mu}) < +\infty$. The following consequence of Theorem 6.4 provides such a collection $\mathcal{E}$.

Theorem 6.7. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(X)^d$ and suppose that both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty and $c(x, y)$ is continuous and symmetric in x, y. Then*

$$\mathcal{I}_c(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{I}_{\tilde{c}}(\boldsymbol{\nu}, \boldsymbol{\mu})$$

where $\tilde{c}(y, x) = c(y, x)$.

Proof. By Theorem 6.4, we have

$$\mathcal{I}_c(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int_{\mathbb{R}^d_+} \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) P(\mathrm{d}\mathbf{z}) = \int_{\mathbb{R}^d_+} \mathcal{I}_{\tilde{c}}(\nu_{\mathbf{z}}, \mu_{\mathbf{z}}) P(\mathrm{d}\mathbf{z}) = \mathcal{I}_{\tilde{c}}(\boldsymbol{\nu}, \boldsymbol{\mu}),$$

where the second step follows since the classic optimal transport problem is symmetric. □

Remark 6.8. As mentioned in Remark 6.2, both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty if and only if $\boldsymbol{\mu}' \overset{\text{law}}{=} \boldsymbol{\nu}'$ considered as random variables on measure spaces (X, μ) and (Y, ν).

Corollary 6.9. *For $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}_P$, we have*

$$\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu})^p = \int_{\mathbb{R}^d_+} \mathcal{W}_p(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})^p P(\mathrm{d}\mathbf{z}).$$

Corollary 6.10. *For each $1 \leqslant p < +\infty$, the metric space $(\mathcal{E}_P, \mathcal{W}_p)$ is complete and separable, hence a Polish space.*

Proof. This follows from standard results on the analysis on the space of random variables taking values in a Polish space; see Crauel (2002). □

The upshot of Theorem 6.7 is that, for $\boldsymbol{\mu}, \boldsymbol{\nu}$ belonging to the same equivalence class we can define the Wasserstein distance

$$\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) := \left(\inf_{\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times X} \rho(x, y)^p \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}.$$

Since for each $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$,

$$\int_{X \times X} c(x, y) \bar{\mu} \otimes \kappa(\mathrm{d}x, \mathrm{d}y) = \frac{1}{d} \sum_{j=1}^{d} \int_{X \times X} c(x, y) \mu_j \otimes \kappa(\mathrm{d}x, \mathrm{d}y),$$

we have by taking infimum that

$$\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu})^p \geqslant \frac{1}{d} \sum_{j=1}^{d} \mathcal{W}_p(\mu_j, \nu_j)^p. \tag{21}$$

Example 6.11. As a sanity check, let us consider the case where $\mu_1 = \cdots = \mu_d$ and $\nu_1 = \cdots = \nu_d$. Then according to discussions in Section 3.3, the optimal transport from μ_1 to ν_1 is also optimal from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$. This means

$$\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu})^p = \frac{1}{d}\sum_{j=1}^{d} \mathcal{W}_p(\mu_j, \nu_j)^p = \mathcal{W}_p(\mu_1, \nu_1)^p.$$

In other words, in the trivial case where all measures are equal, our Wasserstein distance is the same as the classic Wasserstein distance between such measures.

As another sanity check, consider $d = 1$, then for any μ, ν, both $\Pi(\mu, \nu)$ and $\Pi(\nu, \mu)$ are non-empty, so that $\mathcal{W}_p$ is a metric on $\mathcal{P}(X)_{p,\rho}$ and it coincides with the classic Wasserstein distance.

Example 6.12. Suppose that $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R})^d$, define $T(x) = ax + b$ for some $a > 0, b \in \mathbb{R}$ and $\boldsymbol{\nu} := \boldsymbol{\mu} \circ T^{-1}$. Consider the convex cost $c(x, y) = |x - y|^p$, $p \geqslant 1$. Then since the linear transformation is comonotone, $\kappa_T \in \mathcal{K}(\bar{\mu}, \bar{\nu})$ is an optimal transport from $\bar{\mu}$ to $\bar{\nu}$. By arguments in Section 3.3, κ_T is also optimal from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$. In particular, (21) is an equality. Moreover, by the arguments in Section 3.3, in case $\mu_1, \ldots, \mu_d$ have disjoint supports, (21) is also an equality.

6.5 Two further examples

We discuss a few interesting examples illustrating the peculiarities of simultaneous transport (complementing Section 3.3) even if we restrict to the same equivalence class $\mathcal{E}_P$.

From classic optimal transport theory ($d = 1$), we recall the following result (see Theorem 1.17 in Santambrogio (2015)).

Proposition 6.13. *Suppose that probability measures μ, ν are supported on a compact domain $\Omega \subseteq \mathbb{R}^N$ where $\partial\Omega$ is μ-negligible, μ is absolutely continuous, and $c(x, y) = h(y - x)$ with h strictly convex, then there exists a unique transport that is optimal among all Kantorovich transports and such a transport is Monge.*

Example 6.14 below shows that, in the setting of simultaneous transport ($d = 2$), there may not exist an optimal Monge transport even if we assume moreover that both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are absolutely continuous with respect to the Lebesgue measure on $[0, 1]^2$ and are jointly atomless.

We first recall from Exercise 2.14 in Villani (2003) that if we remove the absolute continuity condition of μ while still assuming μ is atomless, Proposition 6.13 may fail to hold. A counterexample is given by μ being uniformly distributed on $[0, 1] \times \{a, b\}$ where $a \neq b$ and $ab \neq 0$, and ν uniform on $[0, 1] \times \{0\}$, with $N = 2$ and $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$.

Example 6.14. Consider $N = 2$ and $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$. Define μ_1, ν_1 being uniformly distributed on $[0, 1] \times [0, 1]$ and $[0, 1] \times [2, 3]$ respectively. Define μ_2 supported on $[0, 1] \times [0, 1]$ such that $\mathrm{d}\mu_2/\mathrm{d}\mu_1(x, y) = 2y$ and ν_2 supported on $[0, 1] \times [2, 3]$ such that $\mathrm{d}\nu_2/\mathrm{d}\nu_1(x, y) = 2 - 2|y - 5/2|$.

Observe that $\bar{\mu}, \bar{\nu}$ are compactly supported and $\boldsymbol{\mu}, \boldsymbol{\nu}$ are jointly atomless (e.g., the uniform distribution on $[0,1] \times \{0\}$ and μ_1' are independent). For each $z \in \mathbb{R}_+$, using notations similarly as in Section 6, we have $A_z := (\mu_1')^{-1}(z) = [0,1] \times \{(1-z)/2z\}$ and $B_z := (\nu_1')^{-1}(z) = [0,1] \times \{(5/2) \pm ((3z-1)/2z)\}$. Moreover, μ_z, ν_z are uniformly distributed on A_z, B_z respectively. Thus, from the counterexample mentioned above, the unique optimal transport from μ_z to ν_z is not Monge unless $z = 1/3$. This proves that the unique optimal transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ is not Monge.

We next discuss an example of simultaneous transport between Gaussian measures. For simplicity we focus on the case $d = 2$ with L^2 cost. First, we record a general result stating that for $\boldsymbol{\mu}, \boldsymbol{\nu}$ in the same equivalence class $\mathcal{E}_P$, there must exist a linear transport between them. That is, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ differ by a nonsingular linear transformation. This can be seen as a complement to Example 6.12 above.

Proposition 6.15. *If $\boldsymbol{\mu}, \boldsymbol{\nu}$ are $\mathbb{R}^2$-valued Gaussian measures on $\mathbb{R}^N$ with positive densities everywhere, then $\boldsymbol{\mu}, \boldsymbol{\nu}$ belong to the same equivalence class $\mathcal{E}_P$ if and only if $\boldsymbol{\mu} \circ T^{-1} = \boldsymbol{\nu}$ where $T(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ with A invertible.*

In the following we discuss an example where the optimal transport may not be the linear transport given in Proposition 6.15.

Consider $\delta > 0$ and Gaussian measures $\mu_1, \nu_1 \sim N(0, I_2)$, $\mu_2 \sim N(0, \Sigma)$, and $\nu_2 \sim N(0, \Omega)$ where

$$\Sigma = \begin{pmatrix} 1+\delta & 0 \\ 0 & 1 \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} 1 & 0 \\ 0 & 1+\delta \end{pmatrix}.$$

It is straightforward to compute all the linear transports in $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. These are given by reflections along $y = \pm x$ axes and rotations of $\pm\pi/2$ degrees at zero. Our goal is to show that these transports are not optimal, in contrast to the case $d = 1$ where optimal transports are linear. Observe that $\boldsymbol{\mu}, \boldsymbol{\nu}$ belong to the same equivalence class $\mathcal{E}_P$ (since two-way transports exist), so that we may apply Theorem 6.4. Consider $z \in (0,2)$, then computing the density yields that $\mathrm{d}\mu_1/\mathrm{d}\bar{\mu}((x,y)) = z$ if and only if

$$x = \pm\sqrt{\frac{2(1+\delta)^{3/2}}{\delta} \log\left(\frac{2}{z} - 1\right)} =: \pm h_\delta(z).$$

Similarly, $\mathrm{d}\nu_1/\mathrm{d}\bar{\nu}((x,y)) = z$ if and only if $y = \pm h_\delta(z)$. The optimal transport problem from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ is then reduced to transporting from $\{(x,y) \mid x = \pm c\}$ to $\{(x,y) \mid y = \pm c\}$ for each $c = h_\delta(z) \geqslant 0$ on which some copies of Gaussian measures are equipped. Direct computation shows that a transport from μ_z to ν_z is given by

$$T((x,y)) = (\mathrm{sgn}(x)|y|, \mathrm{sgn}(y)|x|).$$

This is illustrated by the following Figure 7.

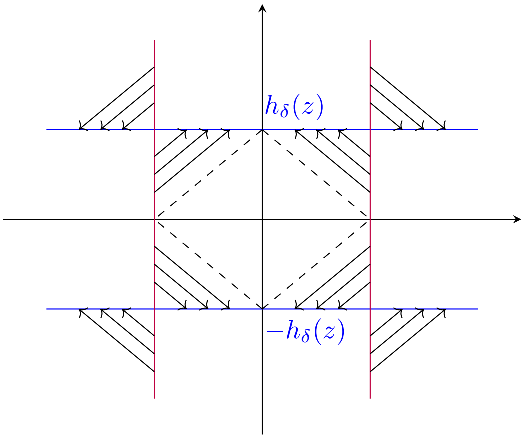


Figure 7: Optimal transport from μ_z to ν_z: the red and blue lines indicate the supports of μ_z and ν_z respectively. Black arrows indicate the transports.

Recall from Theorem 3.2.9 of Rachev and Rüschendorf (1998) that T is optimal if and only if

$$(\operatorname{sgn}(x)|y|, \operatorname{sgn}(y)|x|) \in \partial f(x, y)$$

for some lower semi-continuous convex function f on $\mathbb{R}^2$, where the subdifferential ∂f is given by

$$\partial f(\mathbf{x}) := \{\mathbf{x}^* \in X^* \mid f(\mathbf{x}) - f(\mathbf{y}) \leqslant \langle \mathbf{x} - \mathbf{y}, \mathbf{x}^* \rangle \text{ for all } \mathbf{y} \in X\}, \ \mathbf{x} \in X = \mathbb{R}^2.$$

It is straightforward to check that $f(x, y) = |y|$ meets these criteria, so that T is indeed an optimal transport from μ_z to ν_z. Since this holds for all $z \in (0, 2)$, by Theorem 6.4, the transport T on $\mathbb{R}^2$ is an optimal transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$. Evidently, this is not given by a linear map. Intuitively, even if there always exists an optimal linear transport map when considering transports of a single measure, in our case the measures are weaved in such a way that under both constraints, none of the linear maps become optimal.

It is also straightforward to see that (21) is not an equality in this example.

7 Concluding remarks

The simultaneous optimal transport is introduced and analyzed in this paper. The framework is shown to be much more complicated than the classic setting which corresponds to $d = 1$ and many new mathematical results are obtained. Due to the additional technical richness, there are many future directions to explore within the framework of simultaneous optimal transport. We discuss a few directions below, each deserving a series of follow-up studies.

1. We have focused on the case where d is an integer. The problem can be naturally formulated for infinite dimension, by looking at

$$\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \bigcap_{j \in J} \mathcal{K}(\mu_j, \nu_j)$$

where J is an infinite set which is possibly a continuum. The optimal transport problem in this setting can be seen as a limit in some sense of our setting as $d \to \infty$. A significant technical challenge arises because $\{\mu_j \mid j \in J\}$ may not admit a dominating measure. For studies involving collections of probabilities without a dominating measure, see e.g., Soner et al. (2011) in the context of stochastic analysis with applications to mathematical finance.

2. The setting of this paper involves two tuples of measures to transport between. A natural question is how to generalize the framework to accommodate multiple marginals $\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^n \in \mathcal{P}(X)^d$. For simplicity, assume all marginals are probabilities and defined on the same space X. In case $d = 1$, such a generalization can be conveniently described via the Kantorovich formulation such that the optimal transport problem is

$$\inf_{\pi \in \Pi(\mu^1, \dots, \mu^n)} \int_{X^n} c \, \mathrm{d}\pi,$$

where $c : X^n \to \mathbb{R}$ is the cost function and $\Pi(\mu^1, \dots, \mu^n)$ is the collection of measures with marginals $\mu^1, \dots, \mu^n \in \mathcal{P}(X)$; see e.g., Rüschendorf (2013) and Pass (2015) for results in multi-marginal transports for $d = 1$. In contrast to the case $d = 1$ or $n = 2$, such a generalization cannot be easily described via the Kantorovich formulation for $d \geqslant 2$ and $n \geqslant 3$. A possible formulation via kernels is given by defining, for each $j \in [d]$,

$$\mathcal{K}(\mu_j^1, \dots, \mu_j^n) = \{\kappa : X \to \mathcal{P}(X^{n-1}) \mid \kappa_{\#}\mu_j^1 \in \Pi(\mu_j^2, \dots, \mu_j^n)\}$$

and letting

$$\mathcal{K}(\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^n) = \bigcap_{j=1}^{d} \mathcal{K}(\mu_j^1, \dots, \mu_j^n).$$

Each $\kappa \in \mathcal{K}(\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^n)$ corresponds to a multi-marginal simultaneous transport, with $n = 2$ corresponding to the setting in our paper and $d = 1$ corresponding to the classic multi-marginal transport setting.

3. Recall that in the Monge formulation, the objective is to minimize

$$\mathcal{C}_\eta(T) = \int_X c(x, T(x))\eta(\mathrm{d}x). \tag{22}$$

One may consider a nonlinear reference, i.e., η in (22) is a Choquet capacity[11] instead of a measure. The motivation of this formulation can be

[11] A Choquet capacity η on a σ-field $\mathcal{B}$ of X is a function $\eta : \mathcal{B} \to [0, \infty]$ such that $\eta(\emptyset) = 0$ and $\eta(A) \leqslant \eta(B)$ for $A \subseteq B \subseteq X$, and the integration of $L : X \to \mathbb{R}$ with respect to η is defined as $\int L \mathrm{d}\eta = \int_0^\infty \eta(L > t)\mathrm{d}t + \int_{-\infty}^0 (\eta(L > t) - \eta(X))\mathrm{d}t$.

easily explained in the context of Example 2.3, where the objective is

$$\text{to minimize } \int f(L)\mathrm{d}\eta, \qquad \text{subject to } L \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu}). \tag{23}$$

By taking η as a capacity, (23) includes many popular objectives in risk management and decision analysis. For instance, if η is given by $\eta : A \mapsto \mathbb{1}_{\{\mathbb{P}(A)>1-\alpha\}}$ where $\mathbb{P} \in \mathcal{P}(X)$, then $\int f(L)\mathrm{d}\eta$ is the (left) α-quantile of $f(L)$, and the problem (23) is a quantile optimization problem; see e.g., Rostek (2010) for an axiomatization of quantile optimization in decision theory. This formulation also includes optimization of risk measures (Föllmer and Schied (2016)) or rank-dependent utilities (Quiggin (1993)) of the financial position $f(L)$. More generally, one may optimize $\mathcal{R}(L)$ subject to $L \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ for a general mapping $\mathcal{R} : \mathcal{L} \to \mathbb{R}$, such as many other quantities developed in decision theory (e.g., Hansen and Sargent (2001); Maccheroni et al. (2006); Strzalecki (2011)). Alternatively, instead of choosing η as a capacity, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ may also be chosen as tuples of capacities instead of measures.

4. Our optimal transport is allowed to be chosen from the entire set of transports $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ (kernel) or $\mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ (Monge). There is an active stream of research on optimal transport with constraints such as martingale optimal transport (e.g., Beiglböck et al. (2013, 2017), Galichon et al. (2014), and Beiglböck and Juillet (2016)) and directional simultaneous transport (e.g., Nutz and Wang (2022)). Adding these constraints to the simultaneous transport gives rise to many new challenges and requires further studies.

5. There are a few technical open questions related to results in this paper.

    (a) There are several places in the paper where compactness of X and Y is assumed. In particular, duality (Theorems 5.1 and 5.2) is shown in the compact case or in Euclidean spaces. We expect that duality holds for general Polish spaces. Compactness is also used in Theorem 4.2 and Proposition 4.5. We expect that this condition can be removed. In particular, we note that Theorem 4.2 for $d = 1$ holds without the compactness assumption as shown by Pratelli (2007).

    (b) A natural question to ask (especially motivated from the equilibrium model in Section 5.2) is whether the dual problem admits a solution, i.e., whether the supremum is attained on the right-hand side of (15), given some mild assumptions on c and the reference measure η. There is a trivial counterexample when c is continuous bounded and $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is empty. In this case, the supremum is $+\infty$, which cannot be attained. However, we do not know whether this is the only case.

    (c) Duality, equilibrium, uniqueness, and decomposition results in Sections 5 and 6 are obtained only in the balanced setting. Whether analogous results hold in the unbalanced case requires different techniques and further study.

References

Ambrosio, L. (2003). Lecture notes on optimal transport problems. In *Mathematical Aspects of Evolving Interfaces* (pp. 1–52). Springer, Berlin, Heidelberg.

Attouch, H., Buttazzo, G., and Michaille, G. (2014). *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization.* Society for Industrial and Applied Mathematics.

Bacon, X. (2019). Optimal transportation of vector-valued measures. *arXiv preprint arXiv:* 1901.04765.

Beare, B. K. (2010). Copulas and temporal dependence. *Econometrica*, **78**(1), 395–410.

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

Beiglböck, M. and Juillet, N. (2016). On a problem of optimal transport under marginal martingale constraints. *Annals of Probability*, **44**(1), 42–106.

Beiglböck, M., Nutz, M. and Touzi, N. (2017). Complete duality for martingale optimal transport on the line. *Annals of Probability*, **45**(5), 3038–3074.

Blanchet, A. and Carlier, G. (2016). Optimal transport and Cournot-Nash equilibria. *Mathematics of Operations Research*, **41**(1), 125–145.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, **44**(2), 565–600.

Boerma, J., Tsyvinski, A. and Zimin, A. P. (2021). Sorting with team formation. *arXiv preprint arXiv:* 2109.02730.

Bouchitté, G. and Buttazzo, G. (2001). Characterization of optimal shapes and masses through Monge-Kantorovich equation. *Journal of the European Mathematical Society*, **3**(2), 139–168.

Carlier, G., Chernozhukov, V. and Galichon, A. (2016). Vector quantile regression: An optimal transport approach. *Annals of Statistics*, **44**(3), 1165–1192.

Chen, Y., Conforti, G. and Georgiou, T. T. (2018a). Measure-valued spline curves: An optimal transport viewpoint. *SIAM Journal on Mathematical Analysis*, **50**(6), 5947–5968.

Chen, Y., Georgiou, T. T., and Tannenbaum, A. (2018b). Vector-valued optimal mass transport. *SIAM Journal on Applied Mathematics*, **78**(3), 1682–1696.

Ciosmak, K. J. (2021). Optimal transport of vector measures. *Calculus of Variations and Partial Differential Equations*, **60**(6), 1–22.

Crauel, H. (2002). *Random Probability Measures on Polish Spaces* (Vol. **11**). CRC press.

Daskalakis, C., Deckelbaum, A. and Tzamos, C. (2017). Strong duality for a multiple-good monopolist. *Econometrica*, **85**(3), 735–767.

Delbaen, F. (2021). Commonotonicity and time-consistency for Lebesgue-continuous monetary utility functions. *Finance and Stochastics*, **25**, 597–614.

Dybvig, P. (1988), Distributional analysis of portfolio choice. *Journal of Business*, **61**(3), 369–393.

Embrechts, P., Puccetti, G. and Rüschendorf, L. (2013). Model uncertainty and VaR aggregation. *Journal of Banking and Finance*, **37**(8), 2750–2764.

Ekeland, I. (2010). Notes on optimal transportation. *Economic Theory*, **42**(2), 437–459.

Föllmer, H. and Schied, A. (2016). *Stochastic Finance. An Introduction in Discrete Time.* Fourth Edition. Walter de Gruyter, Berlin.

Friedland, S. (1983). Simultaneous similarity of matrices. *Advances in Mathematics*, **50**(3), 189–265.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A., Henry-Labordère, P. and Touzi, N. (2014). A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Annals of Applied Probability*, **24**(1), 312–336.

Gangbo, W. and McCann, R. J. (1996). The geometry of optimal transportation. *Acta Mathematica*, **177**(2), 113–161.

Gladkov, N. A., Kolesnikov, A. V., and Zimin, A. P. (2019). On multistochastic Monge–Kantorovich problem, bitwise operations, and fractals. *Calculus of Variations and Partial Differential Equations*, **58**(5), 1–33.

Gladkov, N. A., Kolesnikov, A. V., and Zimin, A. P. (2021). The multistochastic Monge–Kantorovich problem. *Journal of Mathematical Analysis and Applications*, **506**, 125666.

Good, I. J. and Welch, L. R. (1963). On the independence of quadratic expressions. *Journal of the Royal Statistical Society: Series B (Methodological)*, **25**(2), 377–382.

Hansen, L. P. and Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review.* **91**(2), 60–66.

Hirsch, F., Profeta, C., Roynette, B., and Yor, M. (2011). *Peacocks and Associated Martingales, with Explicit Constructions*. Springer Science & Business Media.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic Forms in Random Variables: Theory and Applications.* Dekker.

Joe, H. (2014). *Dependence Modeling with Copulas.* London: Chapman & Hall.

Maccheroni, F., Marinacci, M. and Rustichini, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, **74**(6), 1447–1498.

Nöldeke, G. and Samuelson, L. (2018). The implementation duality. *Econometrica*, **86**(4), 1283–1324.

Nutz, M. and Wang, R. (2022). The directional optimal transport. *Annals of Applied Probability*, forthcoming.

Pass, B. (2015). Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, **49**(6), 1771–1790.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, **11**(5-6), 355–607.

Pratelli, A. (2007). On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, **43**(1), 1–13.

Quiggin, J. (1993). *Generalized Expected Utility Theory: The Rank-dependent Model.* Kluwer, the Netherlands.

Rachev, S. T., and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory* (Vol. **1**). Springer Science & Business Media.

Rostek, M. (2010). Quantile maximization in decision theory. *Review of Economic Studies*, **77**, 339–371.

Rüschendorf, L. (2013). *Mathematical Risk Analysis. Dependence, Risk Bounds, Optimal Allocations and Portfolios.* Springer, Heidelberg.

Ryu, E. K., Chen, Y., Li, W., and Osher, S. (2018). Vector and matrix optimal mass transport: theory, algorithm, and applications. *SIAM Journal on Scientific Computing*, **40**(5), A3675–A3698.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians.* Springer, New York.

Sergeichuk, V. V. (1998). Unitary and Euclidean representations of a quiver. *Linear Algebra and its Applications*, **278**(1-3), 37–62.

Shen, J., Shen, Y., Wang, B. and Wang, R. (2019). Distributional compatibility for change of measures. *Finance and Stochastics*, **23**(3), 761–794.

Soner, M., Touzi, N. and Zhang, J. (2011). Quasi-sure stochastic analysis through aggregation. *Electronic Journal of Probability*, **16**, 1844–1879.

Strzalecki, T. (2011). Axiomatic foundations of multiplier preferences. *Econometrica*, **79**(1), 47–73.

Torgersen, E. N. (1991). *Comparison of Statistical Experiments.* Cambridge University Press, Cambridge, England.

Villani, C. (2003). *Topics in Optimal Transportation* (No. 58). American Mathematical Soc.

Villani, C. (2009). *Optimal Transport: Old and New.* Springer-Verlag, Berlin.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics*, **49**(3), 1736–1754.

Wang, B. and Wang, R. (2016). Joint mixability. *Mathematics of Operations Research*, **41**(3), 808–826.

Wang, R., and Ziegel, J. F. (2021). Scenario-based risk evaluation. *Finance and Stochastics*, **25**, 725–756.

Wolansky, G. (2020). Semi-discrete optimal transport. *arXiv preprint arXiv:* 1911.04348v4.

Appendices

A Proofs of results in Section 3

Proof of Proposition 3.7. Note that it suffices to prove the case $T = 3$, as the necessity statement for $T > 3$ follows from that for $T = 3$. Write $\alpha = \sigma_2^2/\sigma_1^2 > 0$ and $\beta = \sigma_3^2/\sigma_2^2 > 0$. Increasing log-concavity of $t \mapsto \sigma_t$ means $\alpha \geqslant \beta \geqslant 1$ (case i) and decreasing log-convexity of $t \mapsto \sigma_t$ means $\alpha \leqslant \beta \leqslant 1$ (case ii).

Using Lemma 3.5 of Shen et al. (2019), the following are equivalent:

(a) $(\mu_2, \mu_3) \preceq_{\mathrm{h}} (\mu_1, \mu_2)$;

(b) $\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_3}\big|_{\mu_3} \preceq_{\mathrm{cx}} \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\big|_{\mu_2}$;

(c) $\frac{\mathrm{d}\mu_3}{\mathrm{d}\mu_2}\big|_{\mu_2} \preceq_{\mathrm{cx}} \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}\big|_{\mu_1}$,

where $\preceq_{\mathrm{cx}}$ is the one-dimensional convex order on $\mathcal{P}$. We shall use the equivalent condition (b) for the case $\alpha, \beta \geqslant 1$ and the condition (c) for the case $\alpha, \beta \leqslant 1$. Writing ξ as a standard Gaussian random variable, and $\overset{\mathrm{law}}{=}$ as equality in distribution, by direct calculation,

$$\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\bigg|_{\mu_2} \overset{\mathrm{law}}{=} \frac{\sigma_2}{\sigma_1} e^{-\frac{Z^2}{2\sigma_1^2}+\frac{Z^2}{2\sigma_2^2}}\bigg|_{Z\sim\mu_2} \overset{\mathrm{law}}{=} \sqrt{\alpha}\, e^{\xi^2(\frac{1}{2}-\frac{\alpha}{2})},$$

$$\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_3}\bigg|_{\mu_3} \overset{\mathrm{law}}{=} \frac{\sigma_3}{\sigma_2} e^{-\frac{Z^2}{2\sigma_2^2}+\frac{Z^2}{2\sigma_3^2}}\bigg|_{Z\sim\mu_3} \overset{\mathrm{law}}{=} \sqrt{\beta}\, e^{\xi^2(\frac{1}{2}-\frac{\beta}{2})},$$

$$\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}\bigg|_{\mu_1} \overset{\mathrm{law}}{=} \frac{\sigma_1}{\sigma_2} e^{-\frac{Z^2}{2\sigma_2^2}+\frac{Z^2}{2\sigma_1^2}}\bigg|_{Z\sim\mu_1} \overset{\mathrm{law}}{=} \sqrt{\frac{1}{\alpha}}\, e^{\xi^2(\frac{1}{2}-\frac{1}{2\alpha})},$$

and

$$\frac{\mathrm{d}\mu_3}{\mathrm{d}\mu_2}\bigg|_{\mu_2} \overset{\mathrm{law}}{=} \frac{\sigma_2}{\sigma_3} e^{-\frac{Z^2}{2\sigma_3^2}+\frac{Z^2}{2\sigma_2^2}}\bigg|_{Z\sim\mu_2} \overset{\mathrm{law}}{=} \sqrt{\frac{1}{\beta}}\, e^{\xi^2(\frac{1}{2}-\frac{1}{2\beta})}.$$

Therefore, $(\mu_2, \mu_3) \preceq_{\mathrm{h}} (\mu_1, \mu_2)$ is equivalent to

$$\beta^{1/2} e^{\xi^2(\frac{1}{2}-\frac{\beta}{2})} \preceq_{\mathrm{cx}} \alpha^{1/2} e^{\xi^2(\frac{1}{2}-\frac{\alpha}{2})} \iff \beta^{-1/2} e^{\xi^2(\frac{1}{2}-\frac{1}{2\beta})} \preceq_{\mathrm{cx}} \alpha^{-1/2} e^{\xi^2(\frac{1}{2}-\frac{1}{2\alpha})}. \tag{24}$$

A convenient result we use here is Corollary 1.2 of Hirsch et al. (2011), which says that the stochastic process $((1+2t)^{1/2} e^{-\xi^2 t})_{t\geqslant 0}$ is a peacock; moreover, it is obvious that this process is non-stationary. This implies that, for $x, y \geqslant 1$, $\sqrt{y} e^{\xi^2(\frac{1}{2}-\frac{y}{2})} \preceq_{\mathrm{cx}} \sqrt{x} e^{\xi^2(\frac{1}{2}-\frac{x}{2})}$ if and only if $y \leqslant x$. Hence, if $\alpha, \beta \geqslant 1$, then (24) is equivalent to $\beta \leqslant \alpha$, thus case (i), and if $\alpha, \beta \leqslant 1$, then (24) is equivalent to $\beta \geqslant \alpha$, thus case (ii).

To show that (i) and (ii) are the only cases where a transport from (μ_1, μ_2) to (μ_2, μ_3) exists, it suffices to exclude the case $\alpha < 1 < \beta$ or $\beta < 1 < \alpha$. Note that in this case $\sqrt{\beta}e^{\xi^2(\frac{1}{2}-\frac{\beta}{2})}$ and $\sqrt{\alpha}e^{\xi^2(\frac{1}{2}-\frac{\alpha}{2})}$ have mismatch supports (one bounded away from $-\infty$ and one bounded away from ∞), and the hence either order in (24) is not possible. □

The condition in Proposition 3.7 is not sufficient when $T > 3$. For example, consider $(\sigma_t) = (8, 4, 2, \sqrt{2}, 1)$. If $\kappa \in \mathcal{K}((\mu_1, \dots, \mu_{d-1}), (\mu_2, \dots, \mu_d))$, then by Lemma D.4 in Appendix D, $\kappa(x; \{\pm x/2\}) = \kappa(x; \{\pm x/\sqrt{2}\}) = 1$, a contradiction.

Proof of Proposition 3.8. By symmetry, it suffices to consider the case $d = 2$ and we assume that $i = 1, j = 2$.

Consider the decomposition $Y = Y_1 \cup Y_2$ where $Y_1 = \{y \in Y \mid \nu_1'(y) \geqslant 1\} = Y_2^c$. Also fix an arbitrary $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Then for $B \subseteq Y_1$, we have

$$\begin{aligned}
\kappa_\#(\mu_1 - \mu_2)_+(B) &= \int_X \kappa(x; B)(\mu_1 - \mu_2)_+(\mathrm{d}x) \\
&\geqslant \int_X \kappa(x; B)(\mu_1 - \mu_2)(\mathrm{d}x) = (\nu_1 - \nu_2)(B) = (\nu_1 - \nu_2)_+(B).
\end{aligned}$$

In fact, this holds with κ replaced by $\kappa|_{X_1}$, where we define $X_1 = \{x \in X \mid \mu_1'(x) \geqslant 1\} = X_2^c$. Similarly, for $B \subseteq Y_2$, we have

$$(\kappa|_{X_2})_\#(\mu_2 - \mu_1)_+(B) \geqslant (\nu_2 - \nu_1)_+(B).$$

Therefore, denoting by η_1 the restriction of η on the set $\{x \in X \mid \mu_1'(x) \geqslant 1\}$ and $\eta_2 = \eta - \eta_1$, we obtain

$$\begin{aligned}
\mathcal{C}_\eta(\kappa) &= \int_{X \times Y} c(x, y)\kappa(x; \mathrm{d}y)\eta(\mathrm{d}x) \\
&= \int_{X_1 \times Y} c(x, y)\kappa|_{X_1}(x; \mathrm{d}y)\eta_1(\mathrm{d}x) + \int_{X_2 \times Y} c(x, y)\kappa|_{X_2}(x; \mathrm{d}y)\eta_2(\mathrm{d}x) \\
&\geqslant \inf_{\kappa \in \mathcal{K}((\mu_1 - \mu_2)_+, (\nu_1 - \nu_2)_+} \mathcal{C}_{\eta_1}(\kappa) + \inf_{\kappa \in \mathcal{K}((\mu_2 - \mu_1)_+, (\nu_2 - \nu_1)_+} \mathcal{C}_{\eta_2}(\kappa) \\
&= \inf_{\kappa \in \mathcal{K}((\mu_1 - \mu_2)_+, (\nu_1 - \nu_2)_+)} \mathcal{C}_\eta(\kappa) + \inf_{\kappa \in \mathcal{K}((\mu_2 - \mu_1)_+, (\nu_2 - \nu_1)_+)} \mathcal{C}_\eta(\kappa),
\end{aligned}$$

where in the last step we used the condition that for any $x \in X$, there exists $y \in Y$ such that $c(x, y) = 0$. Taking infimum over $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ proves (7). □

B Proof of results in Section 4

B.1 Proof of Proposition 4.1

Proof of Proposition 4.1. For each stochastic kernel $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, we can define a measure $\pi \in \mathcal{P}(X \times Y)$ such that

$$\pi(A \times B) = \int_A \kappa(x; B)\eta(\mathrm{d}x) \text{ for all } A \subseteq X,\ B \subseteq Y. \tag{25}$$

Such a measure π exists and is unique by Carathéodory's extension theorem. It follows that for a nonnegative measurable function $f : X \to \mathbb{R}$,

$$\int_{X\times B} f(x)\pi(\mathrm{d}x,\mathrm{d}y) = \int_X \kappa(x;B)f(x)\eta(\mathrm{d}x),$$

which can be proved by considering indicator functions first and then using monotone convergence. Plugging in $f := \mathrm{d}\mu_j/\mathrm{d}\eta$ we obtain for any $B \subseteq Y$,

$$\nu_j(B) \leqslant \int_X \kappa(x;B)\mu_j(\mathrm{d}x) = \int_X \kappa(x;B)\frac{\mathrm{d}\mu_j}{\mathrm{d}\eta}(x)\eta(\mathrm{d}x) = \int_{X\times B} \frac{\mathrm{d}\mu_j}{\mathrm{d}\eta}(x)\pi(\mathrm{d}x,\mathrm{d}y).$$

In addition, for any $A \subseteq X$, $\pi(A\times Y) = \int_A \kappa(x;Y)\eta(\mathrm{d}x) = \eta(A)$, so by definition, $\pi \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})$.

On the other hand, given $\pi \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})$, we have by definition $\pi\circ\pi_1^{-1} = \eta$ where π_1 is projection onto X. By the disintegration theorem for product spaces, there exists a stochastic kernel $\kappa : \mathbb{R} \to \mathcal{P}(Y)$ such that for $A \subseteq X,\ B \subseteq Y$,

$$\pi(A\times B) = \int_A \kappa(x;B)\pi\circ\pi_1^{-1}(\mathrm{d}x) = \int_A \kappa(x;B)\eta(\mathrm{d}x),$$

which is exactly (25). Similarly as above, we have

$$\nu_j(B) \leqslant \int_{X\times B} \frac{\mathrm{d}\mu_j}{\mathrm{d}\eta}(x)\pi(\mathrm{d}x,\mathrm{d}y) = \int_X \kappa(x;B)\frac{\mathrm{d}\mu_j}{\mathrm{d}\eta}(x)\eta(\mathrm{d}x) = \int_X \kappa(x;B)\mu_j(\mathrm{d}x),$$

thus $\kappa \in \mathcal{K}(\boldsymbol{\mu},\boldsymbol{\nu})$. □

B.2 On jointly atomless measures

We first show a useful lemma, Lemma B.3, which will be used to show that joint non-atomicity is sufficient for the equality between the optimal values of Monge and Kantorovich formulations of simultaneous transport in Section 4.

In what follows, $\mathcal{B}$ is always the Borel σ-field on $\mathbb{R}$. We first define another notion of joint non-atomicity introduced by Delbaen (2021). This notion is similar to our Definition 3.1, which was proposed by Shen et al. (2019), but this time defined for σ-fields. Both Shen et al. (2019) and Delbaen (2021) called their properties as being "conditionally atomless" (and they are indeed equivalent in some sense as discussed by Delbaen (2021); see Lemma B.2). Recall that we renamed the notion from Shen et al. (2019) as joint non-atomicity. All inequalities below involving conditional expectations are in the almost sure sense.

Definition B.1. Let $(\Omega,\mathcal{G},\mu)$ be a measure space. We say that $(\mathcal{G},\mu)$ is atomless conditionally to the sub-σ-field $\mathcal{F}\subseteq\mathcal{G}$, if for all $A\in\mathcal{G}$ with $\mu(A)>0$, there exists $A'\subseteq A$, $A'\in\mathcal{G}$, such that

$$\mathbb{E}^\mu[\mathbb{1}_A|\mathcal{F}] > 0 \implies 0 < \mathbb{E}^\mu[\mathbb{1}_{A'}|\mathcal{F}] < \mathbb{E}^\mu[\mathbb{1}_A|\mathcal{F}].$$

Intuitively, the requirement in Definition B.1 means that any set A can be divided into smaller (measured by μ) sets, conditionally on $\mathcal{F}$. Delbaen (2021) showed that the two notions of conditional non-atomicity are equivalent in the sense of Lemma B.2. This equivalence is anticipated because, in the unconditional setting, any set being divisible (corresponding to Definition B.1) is equivalent to the existence of a continuously distributed random variable (corresponding to Definition 3.1); see e.g., Lemma D.1 of Vovk and Wang (2021).

Lemma B.2. *Let μ be any strictly positive convex combination of $\boldsymbol{\mu} \in \mathcal{P}(X)^d$. Then $\boldsymbol{\mu}$ is jointly atomless if and only if $(\mathcal{B}(X), \mu)$ is atomless conditionally to $\sigma(\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\mu)$.*

Proof. This statement follows from Theorem 2.3 of Delbaen (2021). The connection between the two notions of conditional non-atomicity is discussed in Remark 2.11 of Delbaen (2021). □

Next, we are ready to give a useful lemma for non-atomicity on a subset of the sample space.

Lemma B.3. *Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d) \in \mathcal{P}(X)^d$ be jointly atomless. Consider an arbitrary Borel set $B \subseteq \mathbb{R}$ and, without loss of generality, assume $\mu_j(B) > 0$ for $1 \leqslant j \leqslant m$ where $m \leqslant d$. The normalized tuple $\boldsymbol{\mu}_B$ of probability measures on B, given by*

$$\boldsymbol{\mu}_B = \left(\frac{\mu_1|_B}{\mu_1(B)}, \dots, \frac{\mu_m|_B}{\mu_m(B)} \right),$$

is again jointly atomless.

Proof. Let $\mu = (\mu_1 + \dots + \mu_m)/m$ and $\mathcal{F} = \sigma(\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\mu) = \sigma(\mathrm{d}\mu_1/\mathrm{d}\mu, \dots, \mathrm{d}\mu_m/\mathrm{d}\mu)$. Define $\mathcal{F}_B = \{A \cap B \mid A \in \mathcal{F}\}$ and similarly for $\mathcal{B}_B$, and $\mu_B(A) = \mu(A \cap B)/\mu(B)$ for $A \in \mathcal{B}$.

Take $A \in \mathcal{B}_B$ with $\mu(A) = \mu(B)\mu_B(A) > 0$. Note that $(\mu_1, \dots, \mu_m)$ is jointly atomless. Using Lemma B.2, $(\mathcal{B}, \mu)$ is atomless conditionally to $\mathcal{F}$. By definition, there exists $A' \subseteq A$, $A' \in \mathcal{B}$ such that

$$\mathbb{E}^{\mu}[\mathbb{1}_A|\mathcal{F}] > 0 \implies 0 < \mathbb{E}^{\mu}[\mathbb{1}_{A'}|\mathcal{F}] < \mathbb{E}^{\mu}[\mathbb{1}_A|\mathcal{F}]. \tag{26}$$

Since $A' \subseteq A \subseteq B$, we have

$$\mathbb{E}^{\mu_B}[\mathbb{1}_A|\mathcal{F}_B] = \mathbb{E}^{\mu_B}[\mathbb{1}_A|\mathcal{F}] = \mathbb{E}^{\mu}[\mathbb{1}_A|\mathcal{F}],$$

and the same holds for A' in place of A. As a consequence, (26) leads to

$$\mathbb{E}^{\mu_B}[\mathbb{1}_A|\mathcal{F}_B] > 0 \implies 0 < \mathbb{E}^{\mu_B}[\mathbb{1}_{A'}|\mathcal{F}_B] < \mathbb{E}^{\mu_B}[\mathbb{1}_{A'}|\mathcal{F}_B] \tag{27}$$

Note also that $A' \in \mathcal{B}_B$ by definition. Therefore, by treating μ_B as a probability measure on $\mathcal{B}_B$, (27) implies that $(\mathcal{B}_B, \mu_B)$ is atomless conditionally to $\mathcal{F}_B$. Noting that μ_B is a strictly positive convex combination of components of $\boldsymbol{\mu}_B$, and using Lemma B.2 again, we conclude that $\boldsymbol{\mu}_B$ is jointly atomless. □

B.3 Proof of Theorem 4.2

Proof of Theorem 4.2. We can without loss of generality assume that $X = Y$ by considering $\boldsymbol{\mu}, \boldsymbol{\nu}$ as measures on the compact space $X \times Y$, and that each μ_j is a probability measure. We have shown above that Monge transports are special cases as Kantorovich transports, thus the infimum cost among Monge transports is bounded below by that among Kantorovich transports.

To prove the other direction, we first assume that there is $\delta > 0$ such that $\frac{\mathrm{d}\eta}{\mathrm{d}\bar{\mu}}(x) \geqslant \delta$ for all $x \in X$. For each $n \in \mathbb{N}$ we partition X into countably many Borel sets $\{K_{i,n}\}_{i\in\mathbb{N}}$ of diameter smaller than $1/n$ and such that for each i,

$$\frac{\sup_{x\in K_{i,n}} \frac{\mathrm{d}\eta}{\mathrm{d}\bar{\mu}}(x)}{\inf_{x\in K_{i,n}} \frac{\mathrm{d}\eta}{\mathrm{d}\bar{\mu}}(x)} \leqslant 1 + \frac{1}{n}.$$

Consider a transport plan $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Define

$$\boldsymbol{\mu}^{i,n} := \boldsymbol{\mu}|_{K_{i,n}} \text{ and for } B \subseteq X,\ \boldsymbol{\nu}^{i,n}(B) := \int_{K_{i,n}} \kappa(x; B)\boldsymbol{\mu}(\mathrm{d}x). \tag{28}$$

It is then obvious that $\kappa_{i,n} := \kappa|_{K_{i,n}} \in \mathcal{K}(\boldsymbol{\mu}^{i,n}, \boldsymbol{\nu}^{i,n})$. Consider the normalized probability measures

$$\mathrm{d}\tilde{\mu}_j^{i,n} = \frac{\mathrm{d}\mu_j^{i,n}}{\mu_j^{i,n}(K_{i,n})};\ \mathrm{d}\tilde{\nu}_j^{i,n} = \frac{\mathrm{d}\nu_j^{i,n}}{\nu_j^{i,n}(X)}.$$

It is also easy to check that $\kappa_{i,n} \in \mathcal{K}(\tilde{\boldsymbol{\mu}}^{i,n}, \tilde{\boldsymbol{\nu}}^{i,n})$. By Propositions 3.4, $\tilde{\boldsymbol{\mu}}^{i,n} \succeq_{\mathrm{h}} \tilde{\boldsymbol{\nu}}^{i,n}$. By Lemma B.3, $\tilde{\boldsymbol{\mu}}^{i,n}$ is jointly atomless, so that applying Proposition 3.4 again, we conclude that $\mathcal{T}(\tilde{\boldsymbol{\mu}}^{i,n}, \tilde{\boldsymbol{\nu}}^{i,n})$ is non-empty.[12] That is, there exist Monge transports $T_{i,n} : K_{i,n} \to X$ such that $\boldsymbol{\mu}^{i,n} \circ T_{i,n}^{-1} = \boldsymbol{\nu}^{i,n}$. By gluing these, we obtain a Monge transport $T_n : X \to X$. Note that $T_n \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ since

$$\sum_{i\in\mathbb{N}} \boldsymbol{\nu}^{i,n}(B) = \int_X \kappa(x; B)\boldsymbol{\mu}(\mathrm{d}x) \geqslant \bar{\nu}(B).$$

Define $\kappa_n(x; B) := \mathbb{1}_{\{T_n(x)\in B\}}$, then $\kappa_n \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Our goal now is to show that

$$\mathcal{C}_\eta(T_n) = \int_{X\times X} c(x,y)\eta \otimes \kappa_n(\mathrm{d}x, \mathrm{d}y) \to \int_{X\times X} c(x,y)\eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y). \tag{29}$$

Let us define cost functions

$$\bar{c}_n(x,y) := \sup_{x_0\in K_{i,n}, y_0\in K_{\ell,n}} c(x_0, y_0) \text{ if } x \in K_{i,n} \text{ and } y \in K_{\ell,n}.$$

[12] We can forget about the components j where $\tilde{\mu}_j^{i,n}(K_{i,n}) = 0$ because the transport condition is trivially satisfied there.

Then since c is uniform continuous on $X \times X$, we have

$$\int_{X\times X} |\bar{c}_n(x,y) - c(x,y)| \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y) \to 0. \tag{30}$$

On the other hand,

$$\begin{aligned}
\int_{K_{i,n}\times K_{\ell,n}} \eta \otimes \kappa_n(\mathrm{d}x, \mathrm{d}y) &= \int_{K_{i,n}} \mathbb{1}_{\{T_n(x)\in K_{\ell,n}\}} \eta(\mathrm{d}x) \\
&= \int_{K_{i,n}} \mathbb{1}_{\{T_n(x)\in K_{\ell,n}\}} \frac{\mathrm{d}\eta|_{K_{i,n}}}{\mathrm{d}\bar{\mu}^{i,n}}(x) \bar{\mu}^{i,n}(\mathrm{d}x) \\
&\leqslant \sup_{x\in K_{i,n}} \frac{\mathrm{d}\eta}{\mathrm{d}\bar{\mu}}(x) \bar{\nu}^{i,n}(K_{\ell,n}) \\
&\leqslant \left(1 + \frac{1}{n}\right) \inf_{x\in K_{i,n}} \frac{\mathrm{d}\eta}{\mathrm{d}\bar{\mu}}(x) \bar{\nu}^{i,n}(K_{\ell,n}) \\
&\leqslant \left(1 + \frac{1}{n}\right) \int_{K_{i,n}\times K_{\ell,n}} \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y).
\end{aligned}$$

Applying this in the second inequality below yields that

$$\begin{aligned}
\mathcal{C}_\eta(T_n) &\leqslant \int_{X\times X} \bar{c}_n(x,y) \eta \otimes \kappa_n(\mathrm{d}x, \mathrm{d}y) \\
&= \sum_{i\in\mathbb{N}} \sum_{\ell\in\mathbb{N}} \sup_{x\in K_{i,n}, y\in K_{\ell,n}} c(x,y) \int_{K_{i,n}\times K_{\ell,n}} \eta \otimes \kappa_n(\mathrm{d}x, \mathrm{d}y) \\
&\leqslant \left(1 + \frac{1}{n}\right) \sum_{i\in\mathbb{N}} \sum_{\ell\in\mathbb{N}} \sup_{x\in K_{i,n}, y\in K_{\ell,n}} c(x,y) \int_{K_{i,n}\times K_{\ell,n}} \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y) \\
&= \left(1 + \frac{1}{n}\right) \int_{X\times X} \bar{c}_n(x,y) \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y). \qquad (31)
\end{aligned}$$

Combining (30) and (31), and since $c \geqslant 0$, we obtain

$$\limsup_{n\to\infty} \mathcal{C}_\eta(T_n) \leqslant \int_{X\times X} c(x,y) \eta \otimes \kappa(\mathrm{d}x, \mathrm{d}y).$$

The liminf part is similar. We have thus proved (29).

In the general case where $\mathrm{d}\eta/\mathrm{d}\bar{\mu}$ is not bounded below by $\delta > 0$, we consider $\eta_\delta := \eta + \delta\bar{\mu}$. Since c is bounded, we have uniformly for $T \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$,

$$\mathcal{C}_{\eta_\delta}(T) \to \mathcal{C}_\eta(T) \text{ and } \mathcal{C}_{\eta_\delta}(\kappa) \to \mathcal{C}_\eta(\kappa) \text{ as } \delta \to 0.$$

This completes the proof. □

B.4 Proof of Proposition 4.5

Proof of Proposition 4.5. Denote by

$$J_n := \inf_{\pi\in\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}^n)} \mathcal{C}(\pi).$$

It suffices to show for each subsequence $\{n_k\}$ there exists a further subsequence $\{n_{k_\ell}\}$ such that $J_{n_{k_\ell}} \to \inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} \mathcal{C}(\pi)$.

Consider for each n a measure $\pi_n \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu}^n)$. Then since X, Y are compact, the sequence (π_{n_k}) is tight, so a subsequence $(\pi_{n_{k_\ell}})$ converges weakly to some $\pi \in \mathcal{P}(X\times Y)$. Since $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\eta$ is continuous, the operations defining $\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})$ is continuous with respect to weak topology in (10), thus we have $\pi \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})$. Since $c(x,y)$ is continuous, this gives that

$$\lim_{k\to\infty} \mathcal{C}(\pi_{n_k}) = \mathcal{C}(\pi).$$

Taking infimum yields that

$$\liminf_{\ell\to\infty} J_{n_{k_\ell}} \geqslant \inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} \mathcal{C}(\pi).$$

Since $\boldsymbol{\nu}^n \leqslant \boldsymbol{\nu}$, we also have

$$J_{n_{k_\ell}} \leqslant \inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} \mathcal{C}(\pi),$$

thus $J_{n_{k_\ell}} \to \inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} \mathcal{C}(\pi)$, completing the proof. □

C Proof of results in Section 5

C.1 Proof of Theorem 5.1

Proof of Theorem 5.1. We first assume c is continuous. Observe that by definition,

$$\begin{aligned}
&\sup_{\substack{\phi\in C(X)\\ \boldsymbol{\psi}\in C^d(Y)}} \Bigg\{ \int_X \phi\,\mathrm{d}\eta - \int_{X\times Y} \phi(x)\pi(\mathrm{d}x,\mathrm{d}y) \\
&\qquad + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu} - \int_{X\times Y} \boldsymbol{\psi}(y)^\top \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x)\pi(\mathrm{d}x,\mathrm{d}y) \Bigg\} \\
&= \begin{cases} 0 & \text{if } \pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu}); \\ +\infty & \text{elsewhere.} \end{cases}
\end{aligned} \tag{32}$$

For every $p \in C(X\times Y)$, define

$$\begin{aligned}
H(p) := -\sup\Bigg\{ &\int_X \phi\,\mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu} \mid \phi\in C(X),\ \boldsymbol{\psi}\in C^d(Y), \\
&\text{and } \phi(x) + \boldsymbol{\psi}(y)^\top \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x) \leqslant c(x,y) - p(x,y) \Bigg\}.
\end{aligned}$$

We next check that H is convex and lower semi-continuous for the uniform convergence on $X \times Y$. First, consider $\varepsilon > 0$ and $p_1, p_2 \in C(X\times Y)$ with functions $\phi^k,\ \boldsymbol{\psi}^k,\ k=1,2$, such that

$$\int_X \phi^k(x)\eta(\mathrm{d}x)+\int_Y \boldsymbol{\psi}^k(y)^\top \boldsymbol{\nu}(\mathrm{d}y)\geqslant -H(p_k)-\varepsilon,\ k=1,2.$$

For $t\in[0,1]$, let $p_t=tp_1+(1-t)p_2$ and $\phi^t=t\phi^1+(1-t)\phi^2$, $\boldsymbol{\psi}^t=t\boldsymbol{\psi}^1+(1-t)\boldsymbol{\psi}^2$. Then $(\phi^t,\boldsymbol{\psi}^t)$ forms an admissible pair, so that

$$-H(p_t)\geqslant\int_X \phi^t(x)\eta(\mathrm{d}x)+\int_Y \boldsymbol{\psi}^t(y)^\top \boldsymbol{\nu}(\mathrm{d}y)\geqslant -(tH(p_1)+(1-t)H(p_2))-\varepsilon.$$

Letting $\varepsilon\to 0$ proves convexity.

For lower semi-continuity of H, consider a sequence p_n converging to p uniformly in $C(X\times Y)$. Given any $\varepsilon>0$, we choose N large such that for any $n>N$, $\|p_n(x,y)-p(x,y)\|_\infty<\varepsilon/2$. For any $n>N$, choose $(\phi^n,\boldsymbol{\psi}^n)$ such that

$$\int_X \phi^n(x)\eta(\mathrm{d}x)+\int_Y \boldsymbol{\psi}^n(y)^\top \boldsymbol{\nu}(\mathrm{d}y)\geqslant -H(p_n)-\varepsilon/2.$$

Consider $(\phi^{(n)},\boldsymbol{\psi}^{(n)})$ where we define $\boldsymbol{\psi}^{(n)}=\boldsymbol{\psi}^n$ and $\phi^{(n)}=\phi^n-\|p_n-p\|_\infty$. Thus $(\phi^{(n)},\boldsymbol{\psi}^{(n)})$ is admissible with p_n replaced with p. By definition,

$$\begin{aligned}-H(p)&\geqslant\int_X \phi^{(n)}(x)\eta(\mathrm{d}x)+\int_Y \boldsymbol{\psi}^{(n)}(y)^\top \boldsymbol{\nu}(\mathrm{d}y)\\&>\int_X \phi^n(x)\eta(\mathrm{d}x)+\int_Y \boldsymbol{\psi}^n(y)^\top \boldsymbol{\nu}(\mathrm{d}y)-\varepsilon/2\geqslant -H(p_n)-\varepsilon.\end{aligned}$$

That is, for any $\varepsilon>0$ we have found an $N>0$ such that for any $n>N$, $-H(p)>-H(p_n)-\varepsilon$. Therefore, $-H(p)\geqslant\limsup_{n\to\infty}-H(p_n)$, thus proving lower semi-continuity.

By Proposition 9.3.2 in Attouch et al. (2014), $H^{**}=H$ where H^* is the Fenchel-Legendre transform of H. For $\pi\in\mathcal{P}(X\times Y)$, we have by definition

$$\begin{aligned}&H^*(\pi)\\&=\sup\Bigg\{\int_{X\times Y}p(x,y)\pi(\mathrm{d}x,\mathrm{d}y)+\int_X\phi\,\mathrm{d}\eta+\int_Y\boldsymbol{\psi}^\top\mathrm{d}\boldsymbol{\nu}\ \Big|\\&\qquad\qquad \phi\in C(X),\ \boldsymbol{\psi}\in C^d(Y),\ \phi(x)+\boldsymbol{\psi}(y)^\top\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x)\leqslant c(x,y)-p(x,y)\Bigg\}\\&=\sup\Bigg\{\int_{X\times Y}c(x,y)\pi(\mathrm{d}x,\mathrm{d}y)+\int_X\phi\,\mathrm{d}\eta-\int_{X\times Y}\phi(x)\pi(\mathrm{d}x,\mathrm{d}y)\\&\qquad+\int_Y\boldsymbol{\psi}^\top\mathrm{d}\boldsymbol{\nu}-\int_{X\times Y}\boldsymbol{\psi}(y)^\top\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\eta}(x)\pi(\mathrm{d}x,\mathrm{d}y)\ \Big|\ \phi\in C(X),\ \boldsymbol{\psi}\in C^d(Y)\Bigg\}\\&=\begin{cases}\int_{X\times Y}c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) & \text{if } \pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu});\\ +\infty & \text{elsewhere,}\end{cases}\end{aligned}$$

where the last step follows from (32). Since $H^{**} = H$,

$$H(0) = H^{**}(0) = -\inf_{\pi \in \mathcal{P}(X\times Y)} H^*(\pi) = -\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} \int_{X\times Y} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y),$$

proving the duality formula (15) in the case where c is continuous.

Consider the general case where c is lower semi-continuous, possibly taking values in $\{+\infty\}$. As in Villani (2003), we can write $c = \sup c_n$ where each c_n is continuous bounded and c_n is nondecreasing in n. For $(\phi, \boldsymbol{\psi}) \in \Phi_c$, we denote $\varphi^d(\phi,\boldsymbol{\psi}) := \int_X \phi \,\mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu}$. Also write $I_n(\pi) = \int_{X\times Y} c_n \mathrm{d}\pi$. We aim to show that

$$\begin{aligned}\inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} I(\pi) &\leqslant \sup_n \inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} I_n(\pi)\\ &\leqslant \sup_n \sup_{(\phi,\boldsymbol{\psi})\in\Phi_{c_n}} \varphi^d(\phi,\boldsymbol{\psi}) \leqslant \sup_{(\phi,\boldsymbol{\psi})\in\Phi_c} \varphi^d(\phi,\boldsymbol{\psi}). \qquad (33)\end{aligned}$$

The second inequality follows from the first part of the proof, and the third inequality follows from that $\{c_n\}$ is nondecreasing in n, so it suffices to prove the first equality.

Since X, Y are assumed to be compact, $\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})$ is tight. Consider a minimizing sequence $\{\pi_{n,k}\}$ for $\inf I_n(\pi)$. By Prokhorov's theorem, we can extract a subsequence, say $\pi_{n,k} \to \pi_n$ weakly as $k \to \infty$. Note that $\pi_n \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})$ since $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\eta$ is continuous. Thus the infimum is attained at π_n. Again by Prokhorov's theorem, $\pi_n \to \pi_*$ up to extracting a subsequence. By monotone convergence, $I_n(\pi_*) \to I(\pi_*)$. Thus for any $\varepsilon > 0$, we can find N, M such that

$$\inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} I(\pi) \leqslant I(\pi_*) < I_N(\pi_*) + \varepsilon < I_N(\pi_M) + 2\varepsilon.$$

Letting $\varepsilon \to 0$ proves the first inequality of (33). Combining with the trivial bound

$$\inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} I(\pi) \leqslant \sup_{(\phi,\boldsymbol{\psi})\in\Phi_c} \varphi^d(\phi,\boldsymbol{\psi})$$

completes the proof of (15).

To show that the infimum of (15) is attained we still apply Prokhorov's theorem. For a minimizing sequence $\{\pi_k\}$ it has a subsequence converging to $\pi_* \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})$ (since $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\eta$ is continuous) and

$$I(\pi_*) = \lim_{n\to\infty} I_n(\pi_*) \leqslant \lim_{n\to\infty} \limsup_{k\to\infty} I_n(\pi_k) \leqslant \limsup_{k\to\infty} I(\pi_k) = \inf_{\pi\in\Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} I(\pi).$$

This shows the desired attainability. □

C.2 Isomorphic primal and dual problems

Our ultimate goal is to prove Theorem 5.2, which states that Kantorovich duality holds for simultaneous transport in a finite-dimensional Euclidean space $\mathbb{R}^N$ with nonnegative lower semi-continuous cost functions given that $\mathrm{d}\eta/\mathrm{d}\bar{\mu}$

is bounded. Our approach introduces a notion of *isomorphic transports* and in particular, gives an alternative proof of the classic Kantorovich duality for $\mathbb{R}^N$. In this section, X, Y can be general measurable spaces that may not be Polish. Recall that a simultaneous transport problem can be described as a tuple $((X, \boldsymbol{\mu}), (Y, \boldsymbol{\nu}), c, \eta)$.

Definition C.1. We say that two simultaneous Kantorovich transport problems $((X, \boldsymbol{\mu}), (Y, \boldsymbol{\nu}), c, \eta)$ and $((\tilde{X}, \tilde{\boldsymbol{\mu}}), (\tilde{Y}, \tilde{\boldsymbol{\nu}}), \tilde{c}, \tilde{\eta})$ are *isomorphic* if there exist measurable bijections $\Phi : X \to \tilde{X}$ and $\Psi : Y \to \tilde{Y}$ such that $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} \circ \Phi^{-1}$, $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} \circ \Psi^{-1}$, $\tilde{\eta} = \eta \circ \Phi^{-1}$, and for any $\tilde{x} \in \tilde{X}, \tilde{y} \in \tilde{Y}$, $\tilde{c}(\tilde{x}, \tilde{y}) = c(\Phi^{-1}(\tilde{x}), \Psi^{-1}(\tilde{y}))$.

In addition, we say two dual simultaneous Kantorovich transport problems are isomorphic if they are dual problems of two isomorphic simultaneous Kantorovich transport problems.

This is an equivalence relation, and is illustrated by the following diagram (note that c is not a function from X to Y, but represents the transport cost from X to Y):

$$\begin{tikzcd}
(X, \boldsymbol{\mu}, \eta) \arrow[r, "c"] \arrow[d, "\Phi"', "\cong"] & (Y, \boldsymbol{\nu}) \arrow[d, "\Psi", "\cong"'] \\
(\tilde{X}, \tilde{\boldsymbol{\mu}}, \tilde{\eta}) \arrow[r, "\tilde{c}"] & (\tilde{Y}, \tilde{\boldsymbol{\nu}})
\end{tikzcd}$$

It is not difficult to check using a change of variable that given $\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$, there exists $\tilde{\pi} \in \Pi_{\tilde{\eta}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}})$ defined by

$$\tilde{\pi}(\tilde{A} \times \tilde{B}) := \pi(\Phi^{-1}(\tilde{A}) \times \Psi^{-1}(\tilde{B})).$$

Moreover, they have the same transport cost, i.e.,

$$\int_{X \times Y} c(x, y)\pi(\mathrm{d}x, \mathrm{d}y) = \int_{\tilde{X} \times \tilde{Y}} \tilde{c}(\tilde{x}, \tilde{y})\tilde{\pi}(\mathrm{d}\tilde{x}, \mathrm{d}\tilde{y}),$$

which can be seen by approximating $c, \tilde{c}$ by indicator functions and then using monotone convergence. This proves the following proposition.

Proposition C.2. *Two isomorphic Kantorovich transport problems have the same infimum cost.*

Similarly, consider the dual problems of the isomorphic Kantorovich transport problems $((X, \boldsymbol{\mu}), (Y, \boldsymbol{\nu}), c, \eta)$ and $((\tilde{X}, \tilde{\boldsymbol{\mu}}), (\tilde{Y}, \tilde{\boldsymbol{\nu}}), \tilde{c}, \tilde{\eta})$. Given $(\phi, \boldsymbol{\psi}) \in \Phi^d_\eta(c)$, let $\tilde{\phi} := \phi \circ \Phi^{-1}$ and $\tilde{\psi}_j := \psi_j \circ \Psi^{-1}$. Then it is easy to check that $(\tilde{\phi}, \tilde{\boldsymbol{\psi}}) \in \tilde{\Phi}_{\tilde{c}}$ and that

$$\int_X \phi \,\mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu} = \int_{\tilde{X}} \tilde{\phi} \,\mathrm{d}\tilde{\eta} + \int_{\tilde{Y}} \tilde{\boldsymbol{\psi}}^\top \mathrm{d}\tilde{\boldsymbol{\nu}}.$$

This proves the following result.

Proposition C.3. *Two isomorphic dual Kantorovich transport problems have the same supremum value.*

C.3 Proof of Theorem 5.2

We now apply the results from Section C.2 to prove a duality formula in $\mathbb{R}^N$.

Definition C.4. We say a cost function $c : \mathbb{R}^N \times \mathbb{R}^N \to [0, +\infty]$ *has infinite limits* if $c(\mathbf{x}, \mathbf{y}) \geqslant h(\max(\|\mathbf{x}\|, \|\mathbf{y}\|))$ for some continuous function h such that $h(t) \to +\infty$ as $t \to +\infty$.

Lemma C.5. *Suppose $X = Y = \mathbb{R}^N$, the duality formula (15) holds if c has infinite limits in addition to the existing conditions in Theorem 5.1.*

Proof. Consider $\tilde{X} = \tilde{Y} = (-1, 1)^N$ and $\Phi(\mathbf{x}) = \Psi(\mathbf{x}) = 2 \arctan(\mathbf{x})/\pi$ where arctan acts coordinatewise. Define $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} \circ \Phi^{-1}$, $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} \circ \Psi^{-1}$, $\tilde{\eta} = \eta \circ \Phi^{-1}$, and $\tilde{c}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = c(\Phi^{-1}(\tilde{\mathbf{x}}), \Psi^{-1}(\tilde{\mathbf{y}}))$, so that $((X, \boldsymbol{\mu}), (Y, \boldsymbol{\nu}), c, \eta)$ and $((\tilde{X}, \tilde{\boldsymbol{\mu}}), (\tilde{Y}, \tilde{\boldsymbol{\nu}}), \tilde{c}, \tilde{\eta})$ are isomorphic transport problems.

Denote $D = \tilde{X} \times \tilde{Y} = (-1, 1)^{2N}$ with boundary ∂D. We may extend $\tilde{X}, \tilde{Y}$ to $[-1, 1]^N$ and define $\tilde{c} = +\infty$ on ∂D, with $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}}$ kept the same. This does not change the values of primal and dual problems, because:

i. Any $\tilde{\pi} \in \Pi_{\tilde{\eta}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}})$ satisfies $\tilde{\pi}(\partial D) = 0$, as is easily checked from (12).

ii. An admissible tuple $(\tilde{\phi}, \tilde{\boldsymbol{\psi}})$ can be extended to D using continuity and still satisfies $\tilde{\phi}(\mathbf{x}) + \tilde{\boldsymbol{\psi}}(\mathbf{y})^\top \frac{\mathrm{d}\tilde{\boldsymbol{\mu}}}{\mathrm{d}\tilde{\eta}}(\mathbf{x}) \leqslant \tilde{c}(\mathbf{x}, \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in [-1, 1]^N$ because $\tilde{c}(\mathbf{x}, \mathbf{y}) = +\infty$ if $(\mathbf{x}, \mathbf{y}) \in \partial D$.

Observe that since c has infinite limits and Φ^{-1}, Ψ^{-1} are continuous, the extended cost function $\tilde{c}$ on D is lower semi-continuous. Therefore, combining Theorem 5.1, Propositions C.2 and C.3 completes the proof. $\square$

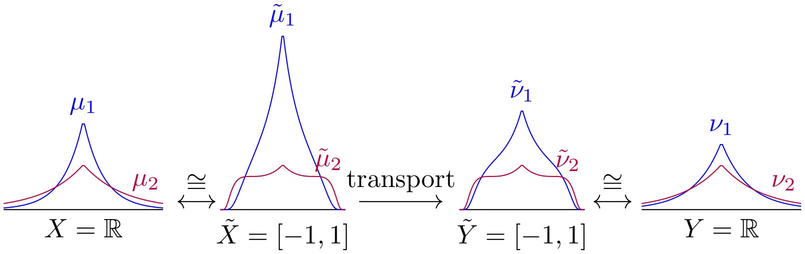


Figure 8: Scheme of the proof of Lemma C.5: to analyze the transport from $(X, \boldsymbol{\mu})$ to $(Y, \boldsymbol{\nu})$, we transform isomorphically via the arctan function to shrink the measures to $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}}$ supported on the compact set $\tilde{X} = \tilde{Y} = [-1, 1]$.

We need an elementary lemma from real analysis, whose proof is left as an exercise.

Lemma C.6. *For any finite measure μ on $\mathbb{R}^N$, there exists a symmetric*[13] *continuous function $f : \mathbb{R}^N \to [0, \infty)$ depending only on $\|\mathbf{x}\|$ that is increasing on $\mathbb{R}^N_{\geqslant 0}$ in $\|\mathbf{x}\|$, satisfies $\lim_{\|\mathbf{x}\| \to +\infty} f(\mathbf{x}) = +\infty$, and is such that $\int f \mathrm{d}\mu < +\infty$.*

Proof of Theorem 5.2. By virtue of Lemma C.6, we choose symmetric continuous functions $f, g : \mathbb{R}^N \to [0, \infty)$ that are increasing on $\mathbb{R}^N_{\geqslant 0}$ in $\|\mathbf{x}\|$, tend to infinity at both infinities, and are such that $\int f \mathrm{d}\eta + \int g \mathrm{d}\bar{\nu} < +\infty$. Define $h(t) = \min(f(\mathbf{x}), g(\mathbf{x}))$ for $\|\mathbf{x}\| = t$, $C := \sup_{\mathbf{x} \in X} \frac{\mathrm{d}\eta}{\mathrm{d}\bar{\mu}}(\mathbf{x})$, and

$$c'(\mathbf{x}, \mathbf{y}) := c(\mathbf{x}, \mathbf{y}) + f(\mathbf{x}) + C g(\mathbf{y}) \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\eta}(\mathbf{x}),$$

so that c' is nonnegative, lower semi-continuous, and satisfies

$$\begin{aligned} c'(\mathbf{x}, \mathbf{y}) &\geqslant f(\mathbf{x}) + C g(\mathbf{y}) \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\eta}(\mathbf{x}) \\ &\geqslant \begin{cases} f(\mathbf{x}) \geqslant h(\|\mathbf{x}\|) = h(\max(\|\mathbf{x}\|, \|\mathbf{y}\|)) & \text{if } \|\mathbf{x}\| \geqslant \|\mathbf{y}\|; \\ C g(\mathbf{y}) \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\eta}(\mathbf{x}) \geqslant h(\|\mathbf{y}\|) = h(\max(\|\mathbf{x}\|, \|\mathbf{y}\|)) & \text{if } \|\mathbf{x}\| \leqslant \|\mathbf{y}\|. \end{cases} \end{aligned}$$

That is, c' has infinite limits. By Lemma C.5, we obtain

$$\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c' \mathrm{d}\pi = \sup_{(\phi, \boldsymbol{\psi}) \in \Phi_{c'}} \int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu}.$$

Note that for any $\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$, by (11),

$$\begin{aligned} \int_{X \times Y} (c' - c) \mathrm{d}\pi &= \int_{X \times Y} f(\mathbf{x}) \pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) + \int_{X \times Y} C g(\mathbf{y}) \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\eta}(\mathbf{x}) \pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \\ &= \int_X f \mathrm{d}\eta + C \int_Y g \mathrm{d}\bar{\nu} =: A(f, g), \end{aligned}$$

which is independent of π.

Moreover, for each $(\phi, \boldsymbol{\psi}) \in \Phi_{c'}$, we let $\tilde{\phi} := \phi - f$ and for each $j \in [d]$, $\tilde{\psi}_j := \psi_j - Cg$. Then $(\tilde{\phi}, \tilde{\boldsymbol{\psi}}) \in \Phi_c$, which implies

$$\begin{aligned} \sup_{(\tilde{\phi}, \tilde{\boldsymbol{\psi}}) \in \Phi_c} \int_X \tilde{\phi} \, \mathrm{d}\eta + \int_Y \tilde{\boldsymbol{\psi}}^\top \mathrm{d}\boldsymbol{\nu} &\geqslant \int_X \tilde{\phi} \, \mathrm{d}\eta + \int_Y \tilde{\boldsymbol{\psi}}^\top \mathrm{d}\boldsymbol{\nu} \\ &= \int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu} - A(f, g). \end{aligned}$$

Taking supremum among $(\phi, \boldsymbol{\psi}) \in \Phi_{c'}$, it follows that

$$\begin{aligned} \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c \, \mathrm{d}\pi &= \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c' \mathrm{d}\pi - A(f, g) \\ &= \sup_{(\phi, \boldsymbol{\psi}) \in \Phi_{c'}} \int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu} - A(f, g) \\ &\leqslant \sup_{(\tilde{\phi}, \tilde{\boldsymbol{\psi}}) \in \Phi_c} \int_X \tilde{\phi} \, \mathrm{d}\eta + \int_Y \tilde{\boldsymbol{\psi}}^\top \mathrm{d}\boldsymbol{\nu}. \end{aligned}$$

[13] By symmetric we mean $f(x_1, \ldots, x_N) = f(|x_1|, \ldots, |x_N|)$ for all $x_1, \ldots, x_N \in \mathbb{R}$.

Since the other inequality

$$\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu},\boldsymbol{\nu})} \int_{X\times Y} c \, \mathrm{d}\pi \geqslant \sup_{(\phi,\boldsymbol{\psi})\in\Phi_c} \int_X \phi \, \mathrm{d}\eta + \int_Y \boldsymbol{\psi}^\top \mathrm{d}\boldsymbol{\nu}$$

is obvious as can be similarly proved as in Proposition 1.5 in Villani (2003), the proof of (15) is finished. □

D Proof of results in Section 6

D.1 Proof of Theorem 6.1

To prove Theorem 6.1 we bootstrap from simple cases.

Lemma D.1. *Consider $d = 2$ and a compact set $X \subseteq \mathbb{R}$. Suppose that both $\Pi(\boldsymbol{\mu},\boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu},\boldsymbol{\mu})$ are non-empty, and μ_1' is strictly increasing on the support of $\bar{\mu}$.*

(i) There exist a unique $\pi \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})$ and a unique $\tilde{\pi} \in \Pi(\boldsymbol{\nu},\boldsymbol{\mu})$.

(ii) We have $\pi(A \times B) = \tilde{\pi}(B \times A)$ for all $(A, B) \in \mathcal{B}(X) \times \mathcal{B}(Y)$.

(iii) It holds

$$\pi\left(\{(x,y) \mid \mu_1'(x) \neq \nu_1'(y)\}\right) = 0. \tag{34}$$

Proof. We describe a process that iteratively divide the transport into disjoint parts. The basic observation is as follows. Consider $I_1^1 = \{x \mid \mu_1'(x) \geqslant 1\}$ and $I_2^1 = \{x \mid \mu_1'(x) < 1\}$. Since μ_1' is increasing, I_1^1, I_2^1 are intervals that form a partition of X. Define similarly $J_1^1 = \{y \mid \nu_1'(y) \geqslant 1\}$ and $J_1^2 = \{y \mid \nu_1'(y) < 1\}$. Since both $\Pi(\boldsymbol{\mu},\boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu},\boldsymbol{\mu})$ are non-empty,

$$(\mu_1 - \mu_2)_+(I_1^1) = (\mu_1 - \mu_2)_+(X) = (\nu_1 - \nu_2)_+(Y) = (\nu_1 - \nu_2)_+(J_1^1).$$

Thus for any $\pi \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})$, we must have $\pi((I_1^1 \times J_2^1) \cup (I_2^1 \times J_1^1)) = 0$. The transport π is thus divided into two disjoint parts, $\pi_1 \in \Pi(\boldsymbol{\mu}|_{I_1^1}, \boldsymbol{\nu}|_{J_1^1})$ and $\pi_2 \in \Pi(\boldsymbol{\mu}|_{I_2^1}, \boldsymbol{\nu}|_{J_2^1})$. We illustrate this in the following figures.

To iterate this procedure we consider normalized measures of $\boldsymbol{\mu}|_{I_1^1}, \boldsymbol{\nu}|_{J_1^1}$ on I_1^1, J_1^1. Call the normalized probability measures $\boldsymbol{\mu}_1^1, \boldsymbol{\nu}_1^1$. Then π up to a constant transports $\boldsymbol{\mu}_1^1$ to $\boldsymbol{\nu}_1^1$. Similarly as above, write $I_1^2 = \{x \in I_1^1 \mid (\mu_1^1)_1'(x) \geqslant 1\}$ and $I_2^2 = \{x \in I_1^1 \mid (\mu_1^1)_1'(x) < 1\}$. Then I_1^2, I_2^2 form a partition of I_1^1. Define in a similar way a partition I_3^2, I_4^2 of I_2^1, so that $\{I_k^2\}_{1\leqslant k\leqslant 4}$ forms a finer partition of X. Also define similarly partitions $\{J_k^2\}_{1\leqslant k\leqslant 4}$ of Y, so that by previous analysis,

$$\pi(X \times Y) = \sum_{k=1}^{4} \pi(I_k^2 \times J_k^2).$$

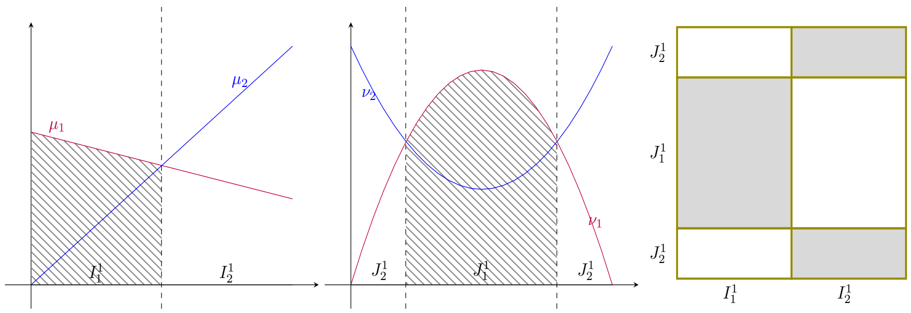


Figure 9: The transports are divided into shaded (I_1^1, J_1^1) and unshaded parts (I_2^1, J_2^1); π must be supported in the gray area.

We continue this process and find partitions $\mathbb{T}_n = \{I_k^n\}_{1 \leqslant k \leqslant 2^n}$ of X and $\{J_k^n\}_{1 \leqslant k \leqslant 2^n}$ of Y such that

$$\pi(X \times Y) = \sum_{k=1}^{2^n} \pi(I_k^n \times J_k^n).$$

This implies for any k and any $B \subseteq J_k^n$, since μ_1' is injective,

$$\pi(I_k^n \times B) = \pi(X \times B). \tag{35}$$

We next show that μ_1' being strictly increasing implies that the mesh of the partition $\mathbb{T}_n$ goes to 0 with $n \to \infty$. In particular, since each $\mathbb{T}_n$ consists of intervals, the ring of subsets $\{\emptyset\} \cup \mathbb{T}_n$ generates the Borel σ-field $\mathcal{B}(X)$.

Since $\mathbb{T}_n$ is a refining sequence of partitions into intervals, we may suppose for contradiction that I is an interval with positive length in the support of $\bar{\mu}$ (so that $\bar{\mu}(I) > 0$) such that there exists a nested collection of intervals $I_1 \supseteq I_2 \supseteq \dots$ with $I_n \in \mathbb{T}_n$ and such that $I = \bigcap_n I_n$. Define for an interval J that $A_J := \mu_1(J)/\bar{\mu}(J)$. Since for each ℓ, $I \subseteq I_{\ell+1} \subseteq I_\ell$, we have by our construction that either $\sup_I \mu_1' \leqslant A_{I_\ell}$ or $\inf_I \mu_1' \geqslant A_{I_\ell}$. Taking the limit as $\ell \to \infty$ and using $\bar{\mu}(I) > 0$ give that either $\sup_I \mu_1' \leqslant A_I$ or $\inf_I \mu_1' \geqslant A_I$. Since μ_1' is increasing, this shows that μ_1' is constant on I, contradicting μ_1' being strictly increasing.

To show that $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is unique, note that for any $A = I_k^n \in \mathbb{T}_n$ and any $B \in \mathcal{B}(Y)$,

$$\pi(A \times B) = \pi(A \times (B \cap J_k^n)) = \pi(X \times (B \cap J_k^n)) = \bar{\nu}(B \cap J_k^n),$$

where the second equality follows from (35) and the third from (12). By the Carathéodory extension theorem applied to the ring $\bigcup_n \mathbb{T}_n \cup \{\emptyset\}$ (on which π is a pre-measure by the σ-additivity of ν), such π is uniquely determined as a measure on $(X \times Y, \mathcal{B}(X) \otimes \mathcal{B}(Y))$.

To show that $\tilde{\pi} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ is unique, observe that a similar argument as above gives that for any k and any $B \subseteq J_k^n$, $\tilde{\pi}(B \times I_k^n) = \tilde{\pi}(B \times X)$. Thus for any $A = I_k^n \in \mathbb{T}_n$ and any $B \in \mathcal{B}(Y)$,

$$\tilde{\pi}(B \times A) = \tilde{\pi}((B \cap J_k^n) \times A) = \tilde{\pi}((B \cap J_k^n) \times X) = \bar{\nu}(B \cap J_k^n) = \pi(A \times B).$$

This proves the uniqueness of $\tilde{\pi} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ and moreover, $\pi(A \times B) = \tilde{\pi}(B \times A)$ for all $A \subseteq X$ and $B \subseteq Y$.

It suffices to prove (34). By (13) and since $\tilde{\pi} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ with $\pi(A \times B) = \tilde{\pi}(B \times A)$, we have for any $A \subseteq X$ and $B \subseteq Y$,

$$\int_{X \times B} \mu_1'(x)\pi(\mathrm{d}x, \mathrm{d}y) = \bar{\nu}_1(B) = \int_{X \times B} \nu_1'(y)\pi(\mathrm{d}x, \mathrm{d}y)$$

and

$$\int_{A \times Y} \nu_1'(y)\pi(\mathrm{d}x, \mathrm{d}y) = \bar{\mu}_1(A) = \int_{A \times Y} \mu_1'(x)\pi(\mathrm{d}x, \mathrm{d}y).$$

It follows from Dynkin's π-λ theorem that for any $A \subseteq X$ and $B \subseteq Y$,

$$\int_{A \times B} (\mu_1'(x) - \nu_1'(y))\pi(\mathrm{d}x, \mathrm{d}y) = 0.$$

Therefore, (34) holds. □

We need another lemma to prove Theorem 6.1. Consider $\mathbb{R}^2$-valued measures $\boldsymbol{\mu}, \boldsymbol{\nu}$ with supports X, Y[14] such that both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty. By the disintegration theorem, there exist measures $\{\mu_z\}_{z \in \mathbb{R}_+}$ such that

$$\mu_z(X \setminus A_z) := \mu_z\left(X \setminus (\mu_1')^{-1}(z)\right) = 0$$

and for any Borel measurable function $f : X \to [0, \infty)$,

$$\int_X f(x)\bar{\mu}(\mathrm{d}x) = \int_{\mathbb{R}_+} \int_{A_z} f(x)\mu_z(\mathrm{d}x)\bar{\mu} \circ (\mu_1')^{-1}(\mathrm{d}z). \tag{36}$$

Lemma D.2. *In the above setting, let $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Then for $\bar{\mu}$-a.e. x, $\kappa(x; B_{\mu_1'(x)}) = 1$, where $B_z = (\nu_1')^{-1}(z)$. In particular, for any $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$,*

$$\pi(\{(x, y) \mid \mu_1'(x) \neq \nu_1'(y)\}) = 0.$$

Proof. The idea is to shrink the parts where μ_1' is not injective into a single point and then use Lemma D.1. Consider $Z = [0, 2]$ and define the measure $\tilde{\mu}$ on Z by

$$\tilde{\mu}(D) := \bar{\mu}\left((\mu_1')^{-1}(D)\right), \ D \subseteq Z,$$

and $\tilde{\boldsymbol{\mu}}$ such that $\tilde{\mu}_1'(z) = z$ for $z \in Z$. It is easy to see by a change of variable that

$$\frac{\mathrm{d}\mu_1 \circ (\mu_1')^{-1}}{\mathrm{d}\bar{\mu} \circ (\mu_1')^{-1}}(z) = z, \tag{37}$$

[14] We no longer assume $X \subseteq \mathbb{R}$.

which implies for $j = 1, 2$, $\tilde{\mu}_j(D) = \mu_j\left((\mu_1')^{-1}(D)\right)$.

We next show that both $\mathcal{K}(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}})$ and $\mathcal{K}(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu})$ are non-empty. Indeed, define $\bar{\kappa}$ by $\bar{\kappa}(x; \{\mu_1'(x)\}) = 1$, then

$$\int_X \bar{\kappa}(x; D)\mu_j(\mathrm{d}x) = \int_X \mathbb{1}_{\{\mu_1'(x)\in D\}}\mu_j(\mathrm{d}x) = \mu_j((\mu_1')^{-1}(D)) = \tilde{\mu}_j(D).$$

Thus $\bar{\kappa} \in \mathcal{K}(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}})$. Define $\tilde{\kappa}$ by $\tilde{\kappa}(z; A) = \mu_z(A \cap A_z)$, then since $\mathrm{d}\mu_1/\mathrm{d}\bar{\mu}(x) = z$ for $x \in A_z$, we have by (36),

$$\begin{aligned}\mu_1(A) &= \int_{\mathbb{R}_+}\int_{A_z\cap A} \frac{\mathrm{d}\mu_1}{\mathrm{d}\bar{\mu}}(x)\mu_z(\mathrm{d}x)\bar{\mu}\circ(\mu_1')^{-1}(\mathrm{d}z)\\ &= \int_{\mathbb{R}_+} z\mu_z(A_z \cap A)\bar{\mu}\circ(\mu_1')^{-1}(\mathrm{d}z) = \int_Z \tilde{\kappa}(z; A)\tilde{\mu}_1(\mathrm{d}z),\end{aligned}$$

where the last step follows from (37). A similar relation holds with μ_1 replaced by μ_2. Thus $\tilde{\kappa} \in \mathcal{K}(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu})$.

By Corollary 3.6 we see that both $\mathcal{K}(\boldsymbol{\nu}, \tilde{\boldsymbol{\mu}})$ and $\mathcal{K}(\tilde{\boldsymbol{\mu}}, \boldsymbol{\nu})$ are non-empty as well. Since $\tilde{\mu}_1'$ is strictly increasing, by Lemma D.1 we may denote the unique transports $\kappa_1 \in \mathcal{K}(\boldsymbol{\nu}, \tilde{\boldsymbol{\mu}})$ and $\kappa_2 \in \mathcal{K}(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}})$. Then for any $D \subseteq \mathbb{R}_+$,

$$\kappa_2(x; D) = \int_Y \kappa_1(y; D)\kappa(x; \mathrm{d}y).$$

By (34) in Lemma D.1, we have

$$1 = \int_{A_z} \kappa_2(x; \{z\})\mu_z(\mathrm{d}x) = \int_{A_z}\int_Y \kappa_1(y; \{z\})\kappa(x; \mathrm{d}y)\mu_z(\mathrm{d}x).$$

Then for μ_z-a.e. $x \in A_z$,

$$1 = \int_Y \kappa_1(y; \{z\})\kappa(x; \mathrm{d}y) = \int_{B_z} \kappa_1(y; \{z\})\kappa(x; \mathrm{d}y) \leqslant \int_{B_z} \kappa(x; \mathrm{d}y) = \kappa(x; B_z),$$

where the second equality follows from (34) in Lemma D.1. This implies that

$$\int_X \kappa(x; B_{\mu_1'(x)})\bar{\mu}(\mathrm{d}x) = \int_{\mathbb{R}_+}\int_{A_z} \kappa(x; B_z)\mu_z(\mathrm{d}x)\bar{\mu}\circ(\mu_1')^{-1}(\mathrm{d}z) = 1.$$

Proof of Theorem 6.1. By definition, if $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$, then $\pi \in \Pi((\mu_1, \mu_2 + \cdots \boxplus \mu_d), (\nu_1, \nu_2 + \cdots + \nu_d))$. By Lemma D.2,

$$\pi\left(\{(x, y) \mid \mu_1'(x) \neq \nu_1'(y)\}\right) = 0.$$

Repeating this argument yields (19). In addition, since $\boldsymbol{\mu}'$ is injective, (19) gives that

$$\pi(\{((\boldsymbol{\mu}')^{-1}\circ\boldsymbol{\nu}'(y), y) \mid y \in Y\}) = d = \pi(X \times Y), \tag{38}$$

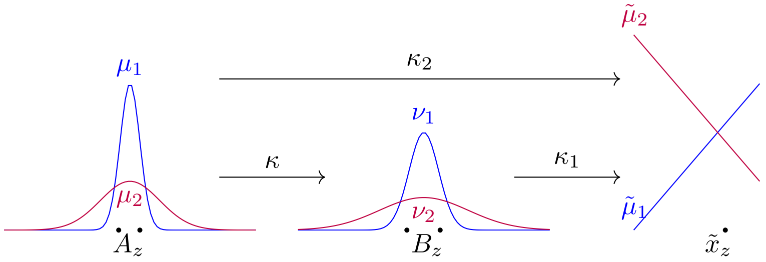


Figure 10: The intuition behind the proof of Lemma D.2: Any $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ must send A_z to B_z, because after transporting B_z using κ_1 to $\tilde{x}_z$, we must obtain the transport κ_2 sending A_z to $\tilde{x}_z$. Note that in the last plot, the density is not with respect to Lebesgue.

so that for $A \subseteq X$ and $B \subseteq Y$,

$$\begin{aligned}\pi(A \times B) &= \pi(A \times (B \cap ((\boldsymbol{\mu}')^{-1} \circ \boldsymbol{\nu}')^{-1}(A))) \\ &= \pi(X \times (B \cap h^{-1}(A))) = \bar{\nu}(B \cap ((\boldsymbol{\mu}')^{-1} \circ \boldsymbol{\nu}')^{-1}(A)).\end{aligned}$$

Thus such $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is unique by Carathéodory's extension theorem. Similarly, again by (19), if $\tilde{\pi} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$,

$$\begin{aligned}\tilde{\pi}(B \times A) &= \tilde{\pi}((B \cap ((\boldsymbol{\mu}')^{-1} \circ \boldsymbol{\nu}')^{-1}(A)) \times A) \\ &= \tilde{\pi}((B \cap ((\boldsymbol{\mu}')^{-1} \circ \boldsymbol{\nu}')^{-1}(A)) \times X) \\ &= \bar{\nu}(B \cap ((\boldsymbol{\mu}')^{-1} \circ \boldsymbol{\nu}')^{-1}(A)) = \pi(A \times B).\end{aligned}$$

This proves (i) and (ii). (iv) follows from (38) and similar arguments as in Remark 6.2. □

D.2 Proof of Theorem 6.4

In the remaining of this appendix we prove Theorem 6.4. First, Lemma D.2 has the following general version for $d \geqslant 2$, which can be proved similarly by using Theorem 6.1 in place of Lemma D.1.

Lemma D.3. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}_P$ and $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Then for $\bar{\mu}$-a.e. x, $\kappa(x; B_{\boldsymbol{\mu}'(x)}) = 1$, where $B_{\mathbf{z}} = \{y \in Y \mid \boldsymbol{\nu}'(y) = \mathbf{z}\}$. In particular, for any $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$,*

$$\pi(\{(x, y) \mid \boldsymbol{\mu}'(x) \neq \boldsymbol{\nu}'(y)\}) = 0.$$

Lemma D.4. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}_P$ and $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. We have $\kappa \in \mathcal{K}(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$ for P-a.s. $\mathbf{z} \in \mathbb{R}^d_+$. More precisely, for any $B \subseteq Y$,*

$$\int_{A_{\mathbf{z}}} \kappa(x; B) \mu_{\mathbf{z}}(\mathrm{d}x) = \nu_{\mathbf{z}}(B).$$

Proof. The measurability of κ as a kernel is obvious. Define measures $\{\tilde{\nu}_{\mathbf{z}}\}$ by

$$\tilde{\nu}_{\mathbf{z}}(B) := \int_{A_{\mathbf{z}}} \kappa(x; B)\mu_{\mathbf{z}}(\mathrm{d}x).$$

Using Lemma D.3 in the third equality yields that

$$\begin{aligned}
\bar{\nu}(B) &= \int_X \kappa(x; B)\bar{\mu}(\mathrm{d}x) \\
&= \int_{\mathbb{R}^d_+} \int_{A_{\mathbf{z}}} \kappa(x; B)\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbb{R}^d_+} \int_{A_{\mathbf{z}}} \kappa(x; B \cap B_{\mathbf{z}})\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbb{R}^d_+} \int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}}} \mathbb{1}_B(y)\kappa(x; \mathrm{d}y)\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbb{R}^d_+} \int_{B_{\mathbf{z}}} \mathbb{1}_B(y) \int_{A_{\mathbf{z}}} \kappa(x; \mathrm{d}y)\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbb{R}^d_+} \int_{B_{\mathbf{z}}} \mathbb{1}_B(y)\tilde{\nu}_{\mathbf{z}}(\mathrm{d}y)P(\mathrm{d}\mathbf{z}).
\end{aligned}$$

This completes the proof by the uniqueness part of the disintegration theorem. □

Lemma D.5. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}_P$ and suppose κ is a stochastic kernel from $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\kappa \in \mathcal{K}(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$ for P-a.s. $\mathbf{z} \in \mathbb{R}^d_+$. Then $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$.*

Proof. Note that for $x \in A_{\mathbf{z}}$ and $B \subseteq Y$, $\kappa(x; B) = \kappa(x; B \cap B_{\mathbf{z}})$. This gives

$$\begin{aligned}
\nu_j(B) &= \int_{\mathbb{R}^d_+} \int_{B_{\mathbf{z}}} \mathbb{1}_B(y)\frac{\mathrm{d}\nu_j}{\mathrm{d}\bar{\nu}}(y)\nu_{\mathbf{z}}(\mathrm{d}y)P(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbb{R}^d_+} z_j \int_{B_{\mathbf{z}}} \mathbb{1}_B(y) \int_{A_{\mathbf{z}}} \kappa(x; \mathrm{d}y)\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbb{R}^d_+} \int_{A_{\mathbf{z}}} z_j\kappa(x; B \cap B_{\mathbf{z}})\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbb{R}^d_+} \int_{A_{\mathbf{z}}} \frac{\mathrm{d}\mu_j}{\mathrm{d}\mu}(x)\kappa(x; B)\mu_{\mathbf{z}}(\mathrm{d}x)P(\mathrm{d}\mathbf{z}) \\
&= \int_X \kappa(x; B)\mu_j(\mathrm{d}x).
\end{aligned}$$

This proves that $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$. □

Proof of Theorem 6.4. We first show that the right-hand side of (20) is well-defined. Recall from the disintegration theorem that the map

$$\mathbb{R}^d_+ \to \mathcal{P}(X) \times \mathcal{P}(Y), \; \mathbf{z} \mapsto (\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$$

is measurable. By Corollary 5.22 in Villani (2009),[15] there exists a measurable map $\mathbf{z} \mapsto \pi_{\mathbf{z}}$ such that for each $\mathbf{z}$, $\pi_{\mathbf{z}}$ is an optimal transport plan from $\mu_{\mathbf{z}}$ to $\nu_{\mathbf{z}}$. We then define the average measure

$$\pi := \int_{\mathbb{R}^d_+} \pi_{\mathbf{z}} P(\mathrm{d}\mathbf{z}).$$

Using similar arguments as in Lemma D.5 as well as (12), we see that $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$. Alternatively, using the kernel formulation, this means there exists a stochastic kernel $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ such that $\kappa \in \mathcal{K}(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$ is an optimal transport from $\mu_{\mathbf{z}}$ to $\nu_{\mathbf{z}}$.[16] Moreover, since $\mathbf{z} \mapsto \pi_{\mathbf{z}}$ is measurable, so is $\mathbf{z} \mapsto \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$.

It follows from the previous paragraph and Lemma D.4 that the kernel κ, such that $\kappa \in \mathcal{K}(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$ is an optimal transport from $\mu_{\mathbf{z}}$ to $\nu_{\mathbf{z}}$, is exactly an optimal transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$. This implies that

$$\mathcal{I}_c(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int_{\mathbb{R}^d_+} \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) P(\mathrm{d}\mathbf{z}).$$

In particular, this gives the equivalence of (i) and (iii).

Note that for $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$,

$$\begin{aligned} \mathcal{C}_{\bar{\mu}}(\kappa) &= \int_X \int_Y c(x,y)\kappa(x,\mathrm{d}y)\bar{\mu}(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d_+} \int_{A_{\mathbf{z}}} \int_Y c(x,y)\kappa(x,\mathrm{d}y)\mu_{\mathbf{z}}(\mathrm{d}x) P(\mathrm{d}\mathbf{z}) \geqslant \int_{\mathbb{R}^d_+} \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) P(\mathrm{d}\mathbf{z}), \end{aligned}$$

where equality holds if and only if

$$\mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) = \int_{A_{\mathbf{z}}} \int_Y c(x,y)\kappa(x,\mathrm{d}y)\mu_{\mathbf{z}}(\mathrm{d}x).$$

That is, κ is optimal from $\mu_{\mathbf{z}}$ to $\nu_{\mathbf{z}}$ for P-a.s. $\mathbf{z}$. This gives the equivalence of (ii) and (iii). □

D.3 Proof of Proposition 6.15 and related discussions

Proof of Proposition 6.15. We first note that, since $\mu_1 \sim \bar{\mu}$ and $\nu_1 \sim \bar{\nu}$, by Lemma 3.5 of Shen et al. (2019), $\boldsymbol{\mu} \simeq \boldsymbol{\nu}$ is equivalent to

$$\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_1}, \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}\right)|_{\mu_1} \overset{\text{law}}{=} \left(\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_1}, \frac{\mathrm{d}\nu_2}{\mathrm{d}\nu_1}\right)|_{\nu_1}.$$

By an isomorphism argument as in Section C.2, we may without loss of generality assume that μ_1 and ν_1 are standard Gaussian, which we denote by χ. We then have

$$\left(\frac{\mathrm{d}\mu_2}{\mathrm{d}\chi}\right)|_{\chi} \overset{\text{law}}{=} \left(\frac{\mathrm{d}\nu_2}{\mathrm{d}\chi}\right)|_{\chi}.$$

[15] Here we use our assumption that c is continuous.

[16] Note that this does not follow from Lemma D.5 where we assumed a priori that κ is a well-defined stochastic kernel, which is a non-trivial fact following from the measure selection theorem.

Suppose that $\mu_2 = N(\mathbf{m}, \Sigma)$ and $\nu_2 = N(\mathbf{n}, \Omega)$. Plugging in the densities we obtain (where $\mathbf{Z}$ is a standard Gaussian random vector)

$$\begin{aligned}&\sqrt{\frac{1}{\det \Sigma}} \exp\left(-\frac{1}{2}((\mathbf{Z}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{Z}-\mathbf{m}) - \mathbf{Z}^\top \mathbf{Z})\right)\\ \overset{\text{law}}{=} &\sqrt{\frac{1}{\det \Omega}} \exp\left(-\frac{1}{2}((\mathbf{Z}-\mathbf{n})^\top \Omega^{-1}(\mathbf{Z}-\mathbf{n}) - \mathbf{Z}^\top \mathbf{Z})\right).\end{aligned}$$

Taking logarithm we obtain

$$\begin{aligned}&(\mathbf{Z}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{Z}-\mathbf{m}) - \mathbf{Z}^\top \mathbf{Z} + \log\det\Sigma\\ \overset{\text{law}}{=} \ &(\mathbf{Z}-\mathbf{n})^\top \Omega^{-1}(\mathbf{Z}-\mathbf{n}) - \mathbf{Z}^\top \mathbf{Z} + \log\det\Omega.\end{aligned}$$

Using (5) in Good and Welch (1963) we can compute the Laplace transforms, so that for all t,

$$\begin{aligned}&\frac{\exp(-2(t\Sigma^{-1}\mathbf{m})^\top (I - 2t(\Sigma^{-1}-I))^{-1}(t\Sigma^{-1}\mathbf{m}))}{|\det(I-2t(\Sigma^{-1}-I))|^{1/2}}\\ &\quad\times \exp\left(t(\mathbf{m}^\top \Sigma \mathbf{m} + \log\det\Sigma)\right)\\ =&\frac{\exp(-2(t\Omega^{-1}\mathbf{n})^\top (I - 2t(\Omega^{-1}-I))^{-1}(t\Omega^{-1}\mathbf{n}))}{|\det(I-2t(\Omega^{-1}-I))|^{1/2}}\\ &\quad\times \exp\left(t(\mathbf{n}^\top \Omega \mathbf{n} + \log\det\Omega)\right).\end{aligned}$$

After squaring both sides, we may recognize either side as a product of a rational function in t and an exponential of a rational function in t (see e.g., Mathai and Provost (1992), Theorem 3.2a.2). The rational functions on both sides must coincide. Thus, for all t,

$$|\det(I - 2t(\Sigma^{-1}-I))| = |\det(I-2t(\Omega^{-1}-I))|. \tag{39}$$

Taking logarithm of the rest we see that the Taylor coefficients around $t = 0$ of $-2(t\Sigma^{-1}\mathbf{m})^\top (I - 2t(\Sigma^{-1}-I))^{-1}(t\Sigma^{-1}\mathbf{m})$ and $t(\mathbf{m}^\top\Sigma\mathbf{m} + \log\det\Sigma)$ separate. This yields

$$\begin{aligned}&(\Sigma^{-1}\mathbf{m})^\top (I-2t(\Sigma^{-1}-I))^{-1}(\Sigma^{-1}\mathbf{m})\\ &= (\Omega^{-1}\mathbf{n})^\top (I-2t(\Omega^{-1}-I))^{-1}(\Omega^{-1}\mathbf{n}).\end{aligned} \tag{40}$$

From (39), we have that the characteristic polynomials of Σ and Ω coincide. Since both of them are symmetric and positive definite, they have the same eigenvalues counted with multiplicity. Writing $\Sigma^{-1} = PDP^{-1}$ and $\Omega^{-1} = QD'Q^{-1}$ with P, Q orthogonal, we have that there is an elementary permutation matrix E such that $D = ED'E^{-1}$. This gives $\Sigma^{-1} = (PEQ^{-1})\Omega^{-1}(PEQ^{-1})^{-1}$. Plugging this into the (40), we have for all t,

$$\begin{aligned}&((PEQ^{-1})^{-1}\Sigma^{-1}\mathbf{m})^\top (I-2t(\Omega^{-1}-I))^{-1}((PEQ^{-1})^{-1}\Sigma^{-1}\mathbf{m})\\ &= (\Omega^{-1}\mathbf{n})^\top (I-2t(\Omega^{-1}-I))^{-1}(\Omega^{-1}\mathbf{n}).\end{aligned}$$

By expanding the term $(I-2t(\Omega^{-1}-I))^{-1}$ and comparing the coefficients in the expansion, we have for any $k \geqslant 2$,

$$((PEQ^{-1})^{-1}\mathbf{m})^\top \Omega^{-k}((PEQ^{-1})^{-1}\mathbf{m}) = \mathbf{n}^\top \Omega^{-k}\mathbf{n}.$$

Since $\Omega^{-1} = QD'Q^{-1}$, we have

$$((PE)^{-1}\mathbf{m})^\top (D')^k((PE)^{-1}\mathbf{m}) = (Q^{-1}\mathbf{n})^\top (D')^k(Q^{-1}\mathbf{n}). \tag{41}$$

Since Ω is positive definite, D' is diagonal and has positive entries along the diagonal. Denote $\lambda_1,\dots,\lambda_\ell$ the distinct eigenvalues (or distinct diagonal entries) of D' and $S_1,\dots,S_\ell$ the corresponding eigenspaces with dimensions $d_1,\dots,d_\ell$. The system of equations (41) then becomes ℓ linearly independent equations since the rank of the Vandermonde matrix formed by diagonal entries of D' is at most ℓ. In this way, (41) reduces to ℓ restrictions that the lengths of the vectors $(PE)^{-1}\mathbf{m}$ and $Q^{-1}\mathbf{n}$ are the same on each S_ℓ. Hence, there exists an orthogonal matrix O consisting of ℓ blocks on the subspaces S_ℓ, each of which is an element in $\mathcal{O}(d_\ell)$ (the set of orthogonal matrices of dimension d_ℓ), such that $Q^{-1}\mathbf{n} = O(PE)^{-1}\mathbf{m}$. Thus $\mathbf{n} = QO(PE)^{-1}\mathbf{m} = (QOQ^{-1})(PEQ^{-1})^{-1}\mathbf{m}$. Since D' is a multiple of identity on each S_ℓ, it commutes with O on each block, hence D' commutes with O. Therefore, the matrix

$$(PEQ^{-1})^{-1}\Sigma^{-1}(PEQ^{-1}) = \Omega^{-1} = QD'Q^{-1}$$

commutes with QOQ^{-1}. We conclude that

$$\Omega^{-1} = (QO(PE)^{-1})^{-1}\Sigma^{-1}(QO(PE)^{-1}).$$

That is, there exists a matrix $M := QO(PE)^{-1}$ such that $\Omega^{-1} = M^{-1}\Sigma^{-1}M$ and $\mathbf{n} = M\mathbf{m}$. Therefore, the linear map M transports μ_2 to ν_2. Since M is orthogonal, it also transports $\chi = \mu_1$ to $\chi = \nu_1$. This concludes the proof. □

A natural question to ask is whether Proposition 6.15 extends to dimensions $d > 2$. In this case, computation of Laplace transforms yields that instead of the relation (39) above, we have for all $\mathbf{t} = \{t_j\}_{2\leqslant j\leqslant d}$ that

$$\left|\det\left(I - 2\sum_{j=2}^{d} t_j(\Sigma_j^{-1} - I)\right)\right| = \left|\det\left(I - 2\sum_{j=2}^{d} t_j(\Omega_j^{-1} - I)\right)\right|$$

and our goal is to provide an orthogonal matrix P such that $\Sigma^{-1} = P\Omega^{-1}P^{-1}$. This is related to the simultaneous similarity of matrices problem, which was solved in Friedland (1983) in the complex case. Friedland (1983) proved that there are only finitely many orbits of the d tuples of symmetric matrices $(A_1,\dots,A_d)$ under the action of simultaneous conjugation by an orthogonal matrix, given some mild conditions on the characteristic polynomial

$$p(\lambda, x) := \det\left(\lambda I - \sum_{j=1}^{d} A_j x^j\right).$$

An open problem was raised whether the same holds for real-valued matrices in Friedland (1983). A counterexample was provided later in Sergeichuk (1998) with matrices that are not positive definite. In addition, note that to apply to our situation, we need a single orbit instead of a finite number of them. Nevertheless, we are not aware of counterexamples in the case $d > 2$ to Proposition 6.15. If two-way transports exist between tuples of Gaussian measures while no linear transport exists, it is interesting to know what such a transport looks like.

E A small review of some related literature

As mentioned in the introduction, we briefly survey a few directions on generalizing the Monge-Kantorovich optimal transport problem in higher dimensions present in the existing literature:

i. The multi-marginal optimal transport problem is a generalization of the classic Monge-Kantorovich transport problem concerning couplings of more than two marginals. For example, the objective of the Kantorovich version of such problems is to minimize

$$\int_{X_1\times\cdots\times X_d} c(x_1,\ldots,x_d)\pi(\mathrm{d}x_1,\ldots,\mathrm{d}x_d)$$

among measures $\pi \in \mathcal{P}(X_1 \times \cdots \times X_d)$ with marginals $\mu_1,\ldots,\mu_d$. A duality formula can be established. However, the existence of a Monge transport is a more delicate problem for dimension $d \geqslant 3$. This problem has applications in physics and economics. See Pass (2015) and Santambrogio (2015) for a review and Rachev and Rüschendorf (1998) for a rich treatment. A solution for the minimization problem with $c(x_1,\ldots,x_d) = (x_1 + \cdots + x_d)^2$ is obtained by Wang and Wang (2016) for some specific choices of $(\mu_1,\ldots,\mu_d)$.

ii. The (n,k) multi-stochastic Monge-Kantorovich problem raised recently in Gladkov et al. (2019) and Gladkov et al. (2021) generalizes the multi-marginal transport problem. For $1 \leqslant k < n$ they considered measures on $X_1\times\cdots\times X_n$ that have fixed projections onto each $X_{i_1} \times \cdots \times X_{i_k}$ where $1 \leqslant i_1 < \cdots < i_k \leqslant n$. The existence of such measures is a non-trivial task. Assuming existence, Gladkov et al. (2019) and Gladkov et al. (2021) also established the theories of duality and cyclical monotonicity. This problem is also connected to Monge-Kantorovich problem with linear constraints.

iii. Bacon (2019) generalized the classic Monge-Kantorovich transport problem to multiple measures, with both transports and transfers allowed, with the name "vector-valued optimal transport". Given probability measures $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_d)$ and $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_d)$, one is allowed to transport not only from each μ_j to ν_j, but also from each μ_j to $\nu_{j'}$ where $j \neq j'$ (this is called a transfer), but the costs may be different. That is, the cost function is matrix-valued with d^2 components and the goal is to minimize the total cost (such a setting does not apply to our main motivating example in Example 2.1). The existence of a transport is guaranteed and duality is obtained. Bacon (2019) also investigated an extension of the Wasserstein distances.

iv. Some earlier studies are in a similar direction as Bacon (2019). To list a few, in Chen et al. (2018a,b) and Ryu et al. (2018), the notion of "vector-valued optimal transport" was proposed. They combined the scalar transport with a network flow problem, formulated using divergences. Similarly as Bacon (2019), both transports and transfers are allowed. In addition, numerical algorithms are available and applications to image processing are discussed.

v. More recently, Ciosmak (2021) proposed a generalization of the Kantorovich-Rubinstein transport problem to higher dimensions, with the name "optimal transport for vector measures". Consider a metric space (X, ρ) and a signed measure η on X such that $\eta(X) = 0$ and there exists $x_0 \in X$ such that $\int_X \rho(x, x_0) \|\eta\| (\mathrm{d}x) < \infty$, where $\|\eta\|$ is the total variation norm of η. This problem deals with

$$\inf_{\pi : P_1\pi - P_2\pi = \eta} \int_{X\times X} \rho(x, y) \|\pi\| (\mathrm{d}x, \mathrm{d}y)$$

where π is an $\mathbb{R}^d$-valued measure, and P_1, P_2 are projections onto the first two coordinates. Existence of π is guaranteed. The Kantorovich-Rubinstein duality formula is extended.

vi. In a recent monograph, Wolansky (2020) discussed the notions of vector-valued transport and optimal multi-partitions. This is similar to our work as such vector-valued transports are indeed simultaneous transports. However, the focus is on the case where the support of $\bar{\nu}$ is finite.[17] Most of the results concern duality formulas and the structure (e.g., existence and uniqueness) of the *optimal* multi-partition, where Y is a finite set and under certain assumptions. A different notion of Wasserstein distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ was formulated by choosing both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ as the measures at origin, defined as

$$\mathcal{V}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) := \left(\inf_{\boldsymbol{\eta} \in \mathcal{M}(X)^d} \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\eta})^p + \mathcal{W}_p(\boldsymbol{\nu}, \boldsymbol{\eta})^p \right)^{1/p}.$$

An application to learning theory is also discussed. The only mathematical overlaps between our paper and Wolansky (2020) are Proposition 3.4 and Theorem 5.1, where our results offer more generality.

[17] which explains the name "multi-partitions". Due to the nature of the problem, it seems mathematically difficult to approximate the general theory by the special case where Y is discrete.