

Simultaneous optimal transport

Ruodu Wang* Zhenyuan Zhang†

20th December 2024

Abstract

We propose a general framework of mass transport between non-negative vector-valued measures, which will be called simultaneous optimal transport (SOT). 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; similarly, in economic 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 conditions and duality formulas are established. More interestingly, by connecting SOT to a natural relaxation of martingale optimal transport (MOT), we introduce the MOT-SOT parity, which allows for explicit solutions of SOT in many interesting cases.

Keywords: Vector-valued measures, duality, martingale optimal transport, multivariate convex order, matching

Contents

1	Introduction	2
2	Simultaneous optimal transport	4
2.1	Simultaneous transport	4
2.2	Balanced simultaneous transport	7
2.3	Assumptions and standing notation	9
3	Existence, inequalities, and examples	9
3.1	Existence of simultaneous transports	9
3.2	Some simple lower bounds for on the minimum cost	11
3.3	Peculiarities of the simultaneous transport	12
4	General properties of simultaneous optimal transport	13
4.1	The Kantorovich formulation	14
4.2	Equivalence between Monge and Kantorovich costs	15
4.3	Connecting the balanced and unbalanced settings	15
4.4	Duality for simultaneous optimal transport	16
5	MOT-SOT parity	17
5.1	Connecting MOT and SOT	18
5.2	Optimality of simultaneous transport and examples	20
5.3	Decomposition of two-way transport	22

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

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

6	Concluding remarks	25
A	Proofs and additional results	27
A.1	Proofs of results in Section 3	27
A.2	Proofs of results in Section 4	29
A.3	Proofs of results in Section 5	35
A.4	On the Wasserstein distance between vector-valued measures	41
A.5	Dual MOT-SOT parity	43
B	A small review of optimal transport in higher dimensions	44
C	Application to a labour market equilibrium model	45

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, operations research, statistics, machine learning, and quantitative finance. A specialized treatment of optimal transport in economics is given by Galichon (2016). 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 this paper, we propose a new framework of optimal transport, which will be called *simultaneous optimal transport* (SOT). 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, with a precise formulation in Section 2) moves mass from d measures on X to d measures on Y *simultaneously*.

SOT provides powerful tools for matching problems with multiple distributional constraints. A considerable amount of new challenges and relevant applications arise, which will gradually be revealed in this paper. The new framework, being mathematically interesting itself, is motivated by several applications from economics, risk management, and stochastic modeling, which are discussed in Section 2 and Appendix C. As a primary example (details in Example 1), 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. Another natural example is refugee resettlement, where refugee families are resettled to different affiliates while fulfilling various quotas and requirements (Example 2).

We will explain below several sharp contrasts between the new and the classic frameworks, along with our contributions and results. The following points are ordered by their natural logical appearance, although the main mathematical results (Theorems 1-4) come a bit later.

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 SOT 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 SOT setting; see Section 2 for an explanation. A connection between the balanced and unbalanced settings is established in Section 4 via a continuity result (Proposition 5).

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 over-supplying 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 SOT 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 (Proposition 1), based on existing results of [Torgersen \(1991\)](#) and [Shen et al. \(2019\)](#). Several other interesting inequalities (e.g., Proposition 3) 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 SOT framework, 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 (Theorem 1).

Fifth, a duality theorem for balanced SOT is obtained in Section 4.4, which has a different form compared with the classic duality formula (Theorem 2). Using the duality result, we construct in Appendix C 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 supply and firms seek.

Sixth, and most importantly, SOT enjoys a unique connection to the active literature of martingale optimal transport (MOT) between two probability measures, that is, classic optimal transport with a martingale constraint. The study of MOT, initialized by [Beiglböck et al. \(2013\)](#) in discrete time and [Galichon et al. \(2014\)](#) in continuous time, is motivated by applications in mathematical finance, in particular, in robust option pricing. The theory is further reinforced by [Beiglböck and Juillet \(2016\)](#), [Beiglböck et al. \(2017\)](#), and [De March and Touzi \(2019\)](#), among many others; see also [Henry-Labordère \(2017\)](#) for a recent survey. In Section 5, we discover an intriguing connection between SOT and MOT, which we call the MOT-SOT parity, that connects SOT in the balanced case with a suitable relaxation of MOT (Theorem 3). This connection allows us to apply techniques from MOT to SOT, thus bridging between two rich topics. In the special case of two-way transport, i.e., simultaneous transport is possible in both forward and backward directions, the MOT component of the problem is degenerate, and the SOT problem can be completely solved (Theorem 4).

In Section 6 we conclude the paper with several other promising directions for future research and

open challenges. 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. To minimize distraction to the reader, we collect them in Appendix B with some detailed discussions. The closest to our framework is perhaps Wolansky (2021) 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 $c : X \times Y \rightarrow [0, \infty]$, the classic optimal transport problem raised by Monge asks for

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

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

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

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

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

2.1 Simultaneous transport

Throughout, components of vector-valued measures are non-negative (i.e., they are measures, not signed measures). We denote by $d \in \mathbb{N}$ the dimension of a vector-valued measure, where the more interesting case is when $d \geq 2$, and by $[d] = \{1, \dots, d\}$. We work with d -tuples of finite Borel measures $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$ on X and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d)$ on Y such that for each $j \in [d]$, $\mu_j(X) \geq \nu_j(Y) > 0$.

We propose the new framework of *simultaneous optimal transport* (SOT) 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

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

Here and throughout, equalities and inequalities are understood component-wise, and for two measures μ and ν on the same space, $\mu \geq \nu$ means that $\mu(A) \geq \nu(A)$ 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 $\kappa_{\#} \boldsymbol{\mu} \geq \boldsymbol{\nu}$, and defined as

$$\kappa_{\#} \boldsymbol{\mu}(\cdot) := \int_X \kappa(x; \cdot) \boldsymbol{\mu}(dx) \geq \boldsymbol{\nu}(\cdot). \quad (1)$$

Imagine that one would like to distribute goods from a (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-emptiness 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 of simultaneous transport problems, which sheds some light on the special structure and technical difference of our problem in contrast to the classic problem.

Example 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,

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

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

$$\sum_{j=1}^m a_j = \sum_{k=1}^{m'} a'_k = \sum_{j=1}^m b_j = \sum_{k=1}^{m'} b'_k = 1. \quad (2)$$

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 Appendix C 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 5.

Example 1 and its continuous version will serve as a primary example to facilitate the understanding of our new framework. To quantify the cost of simultaneous transports, a cost function will

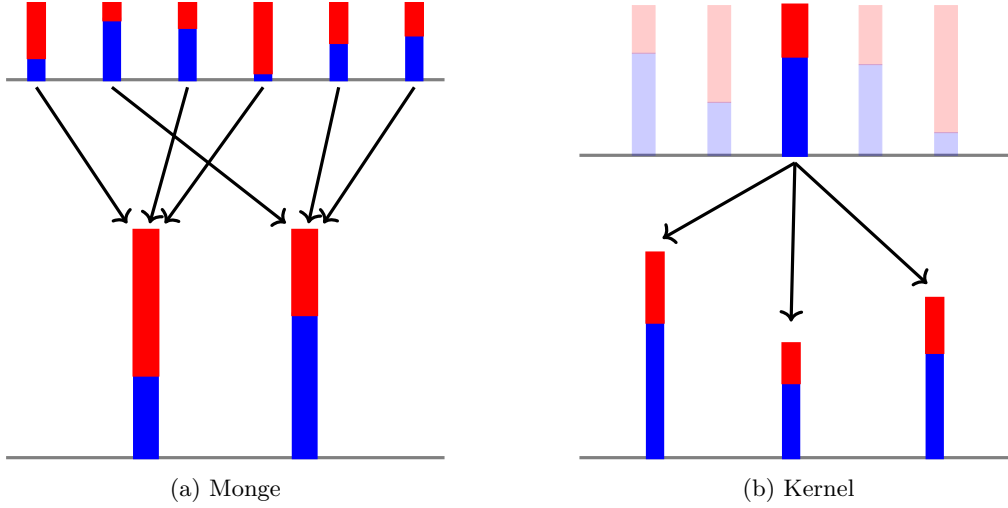


Figure 1: A showcase of simultaneous transport of supplies; red and blue represent different types of products.

be associated with the simultaneous transport problem, as in the classic formulation. Throughout, we define the normalized average measures

$$\bar{\mu} := \frac{\sum_{j=1}^d \mu_j}{\sum_{j=1}^d \mu_j(X)} \quad \text{and} \quad \bar{\nu} := \frac{\sum_{j=1}^d \nu_j}{\sum_{j=1}^d \nu_j(Y)}, \quad (3)$$

which are probability measures. In case μ_1, \dots, μ_d are themselves probability measures, $\bar{\mu}$ is their arithmetic average. Consider a measurable function $c : X \times Y \rightarrow [0, \infty]$ 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

$$\mathcal{C}_\eta(T) := \int_X c(x, T(x)) \eta(dx). \quad (4)$$

Such a reference measure η allows us the greatest generality in view of Example 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.) This is also equivalent to multiplying the cost function c by a factor $d\eta/d\bar{\mu}(x)$ and using $\bar{\mu}$ as the reference measure, but such a change may affect the continuity of the cost function.

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

$$\mathcal{C}_\eta(\kappa) := \int_{X \times Y} c(x, y) \eta \otimes \kappa(dx, dy). \quad (5)$$

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

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

as well as the optimizing transport map and kernel. 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 ∞ . In dimension $d = 1$ and when $\eta = \bar{\mu}$, this cost coincides with the classic Monge–Kantorovich costs. In case $\eta = \bar{\mu}$, we omit the subscript η in (4) and (5).

Example 2 (Refugee resettlement). Refugee resettlement is an active problem in operations research (Delacrétaz et al. (2016) and Ahani et al. (2021)). Let $\mathcal{F} = \{F^1, \dots, F^I\}$ denote the set of refugee families, where each family F^i consists of members $F^i = \{f^{i,1}, \dots, f^{i,J_i}\}$. Our goal is to resettle these refugee families to affiliates $\mathcal{L} = \{L^1, \dots, L^N\}$, such as different cities across USA. There are various quotas $\mathcal{Q} = \{Q^1, \dots, Q^K\}$ to be fulfilled by each family, such as the numbers of adults and children. Let q_k^i denote the contribution of quota k by family i . Each quota Q^k must exceed \underline{q}_k^ℓ in the affiliate L^ℓ . The constraints are twofold: each family member in a refugee family must be resettled to the same affiliation, and the quota requirements are satisfied. To each family-affiliation match is attached a quality score v_ℓ^i , such as the total employment outcome. These lead to the following integer optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_i \sum_\ell v_\ell^i z_\ell^i, \\ & \text{subject to} && z_\ell^i \in \{0, 1\} \text{ and } \sum_\ell z_\ell^i \leq 1 \text{ for all } i; \\ & && \sum_i q_k^i z_\ell^i \geq \underline{q}_k^\ell \text{ for all } \ell, k. \end{aligned} \tag{6}$$

To see this is within the SOT framework (1), we let $X = \mathcal{F}$ and $Y = \mathcal{L}$, and define measures $(\mu_1, \dots, \mu_{|\mathcal{Q}|})$ on X by $\mu_k(\{F^i\}) = q_k^i$ and $(\nu_1, \dots, \nu_{|\mathcal{Q}|})$ on Y by $\nu_k(\{L^\ell\}) = \underline{q}_k^\ell$. The cost function is $-v_\ell^i$. The condition $z_\ell^i \in \{0, 1\}$ asserts that the problem is Monge, corresponding to the fact that each family may be resettled only in one affiliate. Our formulation differs from the original formulations in Delacrétaz et al. (2016) and Ahani et al. (2021) where the “ \geq ” in (6) is “ \leq ”, thus a “dual SOT problem” unbalanced in an opposite direction.

In general, if the supports of $\bar{\mu}, \bar{\nu}$ are both finite (e.g., Example 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 $d \geq 2$ 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

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

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 1.

Example 3 (Financial cost efficiency with multiple distributional constraints). Let (X, \mathcal{F}) be a measurable space on which μ_1, \dots, μ_d are d probability measures and \mathcal{L} be the set of random variables on (X, \mathcal{F}) . Let ν_1, \dots, ν_d be d distributions on \mathbb{R} and define

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

where $L \stackrel{\text{law}}{\sim}_{\mu} \nu$ means that L has distribution ν under μ . The set $\mathcal{L}_\nu(\boldsymbol{\mu})$ represents all possible financial positions which have 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

$$\min \{ \mathbb{E}^\eta[f(L)] \mid L \in \mathcal{L}_\nu(\boldsymbol{\mu}) \}, \quad (7)$$

where $\eta \ll \bar{\mu}$ and f is a nonnegative measurable function. If the probability measure η is a pricing measure on the financial market, then the optimization problem (7) 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 $d \geq 2$, the problem becomes much more complicated, and a special case of mutually singular μ_1, \dots, μ_d is studied by Wang and Ziegel (2021) as the basic tool for representing *coherent scenario-based risk measures*.

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

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

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 1 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 4.4 are applicable.

Example 4 (Time-homogeneous Markov processes with specified marginals). Let μ_1, \dots, μ_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 μ_1, \dots, μ_T . The Markov kernels of ξ , $\kappa_t : \mathbb{R}^N \rightarrow \mathcal{P}(\mathbb{R}^N)$ 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

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

Therefore, the distribution of a time-homogeneous Markov process with marginals (μ_1, \dots, μ_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 SOT, we can study *optimal* (in some sense) time-homogeneous Markov processes. A special case of this example will be given in Proposition 2.

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 $d \geq 2$ the two problems are not equivalent. The reason that the conversion works for $d = 1$ is that the set of unbalanced transports

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

is identical to the set of transports

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

This is not necessarily true in case $d \geq 2$. 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. In other words, even if an unbalanced transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ exists, there may not exist a way to glue mass to $\boldsymbol{\nu}$ that leads to a balanced transport. 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. As we will see, the assumption $\eta \sim \bar{\mu}$ is necessary for the Kantorovich reformulation to make sense.

- i. In Sections 3 and 4.1 through 4.3, we will prove general results in the unbalanced setting;
- ii. in Sections 4.4 and 5 we work within the balanced setting;
- iii. in Section 5.3 we further require that both $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\mathcal{K}(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty; that is, we consider two-way transports.

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 Section 4 we assume that $\eta \sim \bar{\mu}$;
- iii. in Section 5 and throughout our examples we assume for simplicity that $\eta = \bar{\mu}$.

The hierarchical structure of assumptions is summarized in Table 1.

Table 1: Assumptions across sections

Section	Tuples of measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$	Reference η
3	Possibly unbalanced	$\eta \ll \bar{\mu}$
4.1-4.3	Possibly unbalanced	$\eta \sim \bar{\mu}$
4.4	Balanced	$\eta \sim \bar{\mu}$
5.1-5.2	Balanced	$\eta = \bar{\mu}$
5.3	Balanced and two-way	$\eta = \bar{\mu}$

Throughout, we consider the general setting where X and Y are Polish spaces unless otherwise stated. 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

In the study of SOT and its structure, the Radon–Nikodym derivatives of $\boldsymbol{\mu}, \boldsymbol{\nu}$ with respect to $\bar{\mu}, \bar{\nu}$ play a crucial role. For this reason, we recall (3) and introduce the shorthand notation

$$\boldsymbol{\mu}' = \frac{d\boldsymbol{\mu}}{d\bar{\mu}} \quad \text{and} \quad \boldsymbol{\nu}' = \frac{d\boldsymbol{\nu}}{d\bar{\nu}}. \quad (8)$$

We also denote by $m_{\boldsymbol{\mu}}$ and $m_{\boldsymbol{\nu}}$ the laws of $\boldsymbol{\mu}'$ under $\bar{\mu}$ and of $\boldsymbol{\nu}'$ under $\bar{\nu}$. Note that both $m_{\boldsymbol{\mu}}$ and $m_{\boldsymbol{\nu}}$ are probability measures on \mathbb{R}^d .

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. The following definition is adapted from Shen et al. (2019) where $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$ is assumed. Let us emphasize that the paper Shen et al. (2019) is only related to the existence of simultaneous transports, and is independent of everything else discussed in this paper.

Definition 1. We say that $\mu \in \mathcal{M}(X)^d$ is *jointly atomless* if there exists a random variable $\xi : X \rightarrow \mathbb{R}$ such that under $\bar{\mu}$, ξ is atomless and independent of μ' .

Remark 1. Shen et al. (2019) called the notion of joint non-atomicity in Definition 1 as “conditional non-atomicity”. We choose the term “joint non-atomicity” because this notion is indeed a collective property of (μ_1, \dots, μ_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 $d \geq 2$; see Remark 5.

Proposition 1 (Torgersen (1991); Shen et al. (2019)). *Let $\mu \in \mathcal{M}(X)^d$ and $\nu \in \mathcal{M}(Y)^d$.*

- (i) *The set $\mathcal{K}(\mu, \nu)$ is non-empty if and only if $m_\mu \succeq_{\text{icx}} m_\nu$, where \succeq_{icx} is the multivariate increasing convex order.¹*
- (ii) *Assume that μ is jointly atomless. The set $\mathcal{T}(\mu, \nu)$ is non-empty if and only if $m_\mu \succeq_{\text{icx}} m_\nu$.*

In particular, it follows from Proposition 1 that $\mathcal{K}(\mu, \nu)$ and $\mathcal{K}(\nu, \mu)$ are both non-empty if and only if $m_\mu = m_\nu$. We also note that $m_\mu \succeq_{\text{icx}} m_\nu$ implies $\mu(X) \geq \nu(Y)$ 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 $\mathcal{K}(\mu, \nu)$ under this condition.

Remark 2. In Definition 1, if $\mu(X) = \nu(Y)$, then $m_\mu \succeq_{\text{icx}} m_\nu$ is equivalent to $m_\mu \succeq_{\text{cx}} m_\nu$, where \succeq_{cx} is the multivariate convex order.²

Remark 3. To understand $m_\mu \succeq_{\text{icx}} m_\nu$ intuitively, one could look at some special cases (treated in Proposition 3.7 of Shen et al. (2019)), by assuming $\mu(X) = \nu(Y)$. Suppose that $\mathcal{K}(\mu, \nu)$ is non-empty. Then if μ has identical components, so does ν ; if μ has equivalent components, so does ν ; if ν has mutually singular components, so does μ . Moreover, if μ has mutually singular components or ν has identical components, then $\mathcal{K}(\mu, \nu)$ is non-empty.

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

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

then $\mathcal{K}(\mu, \nu)$ is non-empty.³ To see this, we may assume each μ_j is a probability measure. For $\nu := (\max_{j \in [d]} \nu_j) / ((\max_{j \in [d]} \nu_j)(Y))$, we have that $\mathcal{K}(\mu, (\nu, \dots, \nu))$ is non-empty since the constant kernel $x \mapsto \nu$ is in $\mathcal{K}(\mu, (\nu, \dots, \nu))$. Then for $\kappa \in \mathcal{K}(\mu, (\nu, \dots, \nu))$,

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

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

Remark 4. The converse of Proposition 1(ii) does not hold even if $\mu(X) = \nu(Y)$. There are examples where μ is not jointly atomless but there exists a unique Kantorovich transport that is also Monge. A trivial example could be that $d = 1$ and $\mu = \nu = \delta_0$. A more meaningful example is given by Theorem 6.1 of Wolansky (2021), which states that if the space Y is finite, the jointly atomless condition can be relaxed to the atomless condition.

We record an immediate corollary of Proposition 1 for the subsequent analysis, which can also be shown by directly using definition.

Corollary 1. *Suppose that μ, ν, η are \mathbb{R}_+^d -valued probability measures on Polish spaces such that $\mathcal{K}(\mu, \nu)$ and $\mathcal{K}(\nu, \eta)$ are non-empty. Then $\mathcal{K}(\mu, \eta)$ is non-empty.*

¹This means $\int f(\mu') d\bar{\mu} \geq \int f(\nu') d\bar{\nu}$ for all increasing convex $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that the integrals are well-defined.

²This means $\int f(\mu') d\bar{\mu} \geq \int f(\nu') d\bar{\nu}$ for all convex functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that the integrals are well-defined.

³For a collection of (signed) measures $\mu_j, j \in J$ 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.

Proposition 1 can also be applied to give a necessary condition for the existence of a time-homogeneous Markov process (see Example 4) for centered Gaussian marginals on \mathbb{R} .

Proposition 2. *Suppose that $\mu_t = \mathcal{N}(0, \sigma_t^2)$, $\sigma_t > 0$, $t = 1, \dots, T$. For the existence of a transport from $(\mu_1, \dots, \mu_{T-1})$ to (μ_2, \dots, μ_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 2 is not sufficient for $T > 3$. See Appendix A.1 for a counterexample. In the case $T = 3$, the Markov process in Proposition 2 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

$$\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa) \geq \sup_{\boldsymbol{\lambda} \in \Delta_d} \inf_{\kappa \in \mathcal{K}(\boldsymbol{\lambda} \cdot \boldsymbol{\mu}, \boldsymbol{\lambda} \cdot \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa). \quad (9)$$

In particular, if $\kappa \in \mathcal{K}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}})$ is an optimal transport from $\bar{\boldsymbol{\mu}}$ to $\bar{\boldsymbol{\nu}}$ 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 6, the inequality (9) 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 $\nu_1 \geq \nu_2$ must be transported from the parts where $\mu_1 \geq \mu_2$ (see Figure 2).

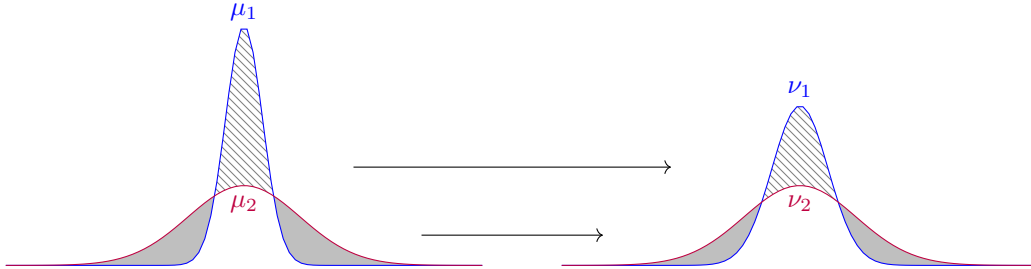


Figure 2: Part of the shaded region $(\mu_1 - \mu_2)_+$ on the left is transported to cover all of the shaded region $(\nu_1 - \nu_2)_+$ on the right; similarly, part of the gray region $(\mu_2 - \mu_1)_+$ is transported to cover all of the gray region $(\nu_2 - \nu_1)_+$.

Proposition 3. *Suppose that $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$, and for each $x \in X$, $c(x, y) = 0$ for some $y \in Y$. Then*

$$\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa) \geq \max_{i, j \in [d]} \left(\inf_{\kappa \in \mathcal{K}((\mu_i - \mu_j)_+, (\nu_i - \nu_j)_+)} \mathcal{C}_\eta(\kappa) + \inf_{\kappa \in \mathcal{K}((\mu_j - \mu_i)_+, (\nu_j - \nu_i)_+)} \mathcal{C}_\eta(\kappa) \right). \quad (10)$$

In particular, if $(\mu_i - \mu_j)_+(X) < (\nu_i - \nu_j)_+(Y)$ for some $i, j \in [d]$, then both sides of (10) are equal to ∞ .

Note that the quantities on the right-hand side of (10) arise from two separate one-dimensional transport problems. Such problems are well-studied in the optimal transport literature; see Santambrogio (2015); Villani (2003, 2009).

If μ_1, \dots, μ_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, (9) is sharp, and (10) is also sharp when $d = 2$.

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., $\mu(X) = \nu(Y)$) 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 by $\mu'_j(x), \nu'_j(y)$ 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 $\psi = (\psi_1, \dots, \psi_d)$ on Y such that

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

then all transports (should any exist) from μ to ν have the same cost

$$\int_{X \times Y} c d(\eta \otimes \kappa) = \int_X \phi d\eta + \int_Y \psi^\top d\nu, \quad (11)$$

because $\kappa_\# \mu = \nu$. 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 $\psi : Y \rightarrow \mathbb{R}$ can also be included in $c(x, y)$, by noting that

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

Moreover, (11) also hints on how a duality result would look like in this setting, which will be discussed in Section 4.4.

Example 5. Consider $X = \mathbb{R}$ on which Borel probability measures μ 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 ν be probability measures on \mathbb{R} such that $\mathcal{K}(\mu, \nu)$ 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 $\kappa \in \mathcal{K}(\mu, \nu)$,

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

Example 6. As a concrete but slightly more general example, we consider $X = Y = [0, 1]$ on which Borel probability measures μ_j, ν_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

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

For any $s \in [0, 1/2]$ and any S such that $\nu_1(S) = 1 - 2s$, 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 5, for a kernel $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ we compute its transport cost

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

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.

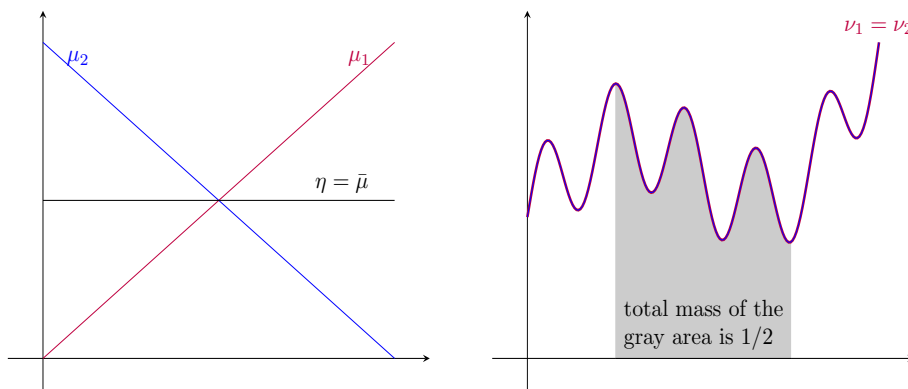


Figure 3: Densities of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ in Example 6.

Trivial as it looks, Examples 5 and 6 provide us with some interesting aspects of the SOT 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.
- ii. It is also easy to see that the equality in (9) may not hold, for example when ν_1 is uniform on $[0, 1]$. In addition, the inequality (10) becomes trivial since it gives a lower bound 0.

After developing our theory, we discuss a few more interesting examples in Section 5.3.

4 General properties of simultaneous optimal transport

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

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

The subscript η incorporates the way we calculate costs: see (4) and (5). It is immediate that

$$\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\pi) := \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c(x, y) \pi(dx, dy) = \inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}_\eta(\kappa). \quad (12)$$

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

$$\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left\{ \pi \in \mathcal{P}(X \times Y) \mid \pi(dx \times Y) = \eta(dx) \text{ and } \int_X \frac{d\boldsymbol{\mu}}{d\eta}(x) \pi(dx, dy) \geq \boldsymbol{\nu}(dy) \right\}. \quad (13)$$

In a way similar to Proposition 4, in the balanced case, i.e., $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$, we have

$$\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left\{ \pi \in \mathcal{P}(X \times Y) \mid \pi(dx \times Y) = \eta(dx) \text{ and } \int_X \frac{d\boldsymbol{\mu}}{d\eta}(x) \pi(dx, dy) = \boldsymbol{\nu}(dy) \right\}. \quad (14)$$

In particular, if $\eta = \bar{\mu}$, we denote by $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \Pi_{\bar{\mu}}(\boldsymbol{\mu}, \boldsymbol{\nu})$, and (14) reads as

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left\{ \pi \in \mathcal{P}(X \times Y) \mid \pi(dx \times Y) = \bar{\mu}(dx) \text{ and } \int_X \boldsymbol{\mu}'(x) \pi(dx, dy) = \boldsymbol{\nu}(dy) \right\}. \quad (15)$$

It seems worthwhile to explain the similarities and differences of (15) compared to the classic definition $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ in the case $d = 1$ (see (17) below). First, by summing over and normalizing the second constraint in (15), 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 (15) means that the mass sent from the contribution of μ_j covers exactly the corresponding portion of ν_j in Y .

We can reformulate (15) as

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left\{ \pi \in \mathcal{P}(X \times Y) \mid \int_{X \times Y} f(x) \pi(dx, dy) = \int_X f d\bar{\mu} \text{ and } \int_{X \times Y} \boldsymbol{\mu}'(x) g(y) \pi(dx, dy) = \int_Y g d\boldsymbol{\nu} \text{ for all measurable } f, g \right\}. \quad (16)$$

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

$$\tilde{\Pi}(\mu, \nu) := \{\pi \in \mathcal{P}(X \times Y) \mid \pi(dx \times Y) = \mu(dx) \text{ and } \pi(X \times dy) = \nu(dy)\}. \quad (17)$$

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 (16), we have for any $j \in [d]$, the measure $\mu'_j(x) \pi(dx, dy)$ 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 μ, ν , as expected from Proposition 1. It seems unlikely that $\Pi(\mu, \nu)$ can be defined in a similar way as (17) 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 1, 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 $T \in \mathcal{T}(\mu, \nu)$, we can associate a kernel $\kappa_T \in \mathcal{K}(\mu, \nu)$ defined by $\kappa_T(x; B) := \mathbb{1}_{\{T(x) \in B\}}$. In view of (4) and (5), they have the same transport cost.

Theorem 1 (Cost equality). *Let $\eta \sim \bar{\mu}$. Suppose that X, Y are compact spaces on which μ, ν are supported, μ is jointly atomless, and c is continuous. Then the transport plans and transport maps admit the same infimum cost. That is,*

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

Combining with (12) yields the following.

Corollary 2. *Consider a reference measure $\eta \sim \bar{\mu}$. Suppose that X, Y are compact spaces on which μ, ν are supported, μ is jointly atomless, and c is continuous, then Monge and Kantorovich transport costs have the same infimum value. That is,*

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

The proof of Theorem 1 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 1.

Remark 5. 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 1). In the case $d = 1$ with non-atomicity, 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 A.1).
- iii. The joint non-atomicity condition enables us to conclude Monge and Kantorovich problems have the same infimum (Corollary 2), 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 SOT in the unbalanced setting. In real applications such as the setting of Example 1, it likely holds that $\mu(X) \geq \nu(Y)$ 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 $\nu(Y) \approx 0.95 \times \mu(X)$.

As we will see in the subsequent sections, results on duality, equilibria, and the MOT-SOT parity will be obtained in the setting of balanced transport, since the balanced setting has a 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 $\nu(Y) \approx (1 - \varepsilon) \times \mu(X)$ for a small $\varepsilon > 0$, which is more realistic in applications.

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

Proposition 5. *Suppose that X, Y are compact Polish spaces, $\mu \in \mathcal{M}(X)^d$, and $d\mu/d\eta$ and c are continuous. Suppose that $(\nu^n)_{n \in \mathbb{N}} \subseteq \mathcal{M}(Y)^d$ is a sequence of measures converging weakly to $\nu \in \mathcal{M}(Y)^d$ such that $\nu^n \leq \nu$ for each $n \in \mathbb{N}$. Then*

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

Proposition 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.

4.4 Duality for simultaneous optimal transport

Consider \mathbb{R}_+^d -valued measures μ, ν on Polish spaces X, Y satisfying $\mu(X) = \nu(Y)$ (e.g., when they are probability measures), a reference probability $\eta \sim \bar{\mu}$, and μ, ν are absolutely continuous with respect to $\bar{\mu}, \bar{\nu}$ with densities μ' on X and ν' on Y respectively. Also, recall that (14) is the set of all transport plans from the vector-valued measure μ to the vector-valued measure ν .

We give a duality theorem for SOT on Polish spaces. A detailed proof will be provided in Appendix A.2.

Theorem 2 (Duality). *Suppose that X, Y are Polish spaces, $\eta \sim \bar{\mu}$ with both $d\mu/d\eta$ and $d\eta/d\bar{\mu}$ bounded continuous, and $c : X \times Y \rightarrow [0, \infty]$ is lower semi-continuous.⁴ Duality holds as*

$$\inf_{\pi \in \Pi_\eta(\mu, \nu)} \int_{X \times Y} c \, d\pi = \sup_{(\phi, \psi) \in \Phi_c} \int_X \phi \, d\eta + \int_Y \psi^\top \, d\nu, \quad (18)$$

where

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

Moreover, the infimum in (18) is attained.

Remark 6. Under more restrictive assumptions (X, Y compact and c continuous), Theorem 2 also follows from an abstract duality theorem established in Gover (2020) for Banach space-valued measures.

In the case $d = 1$ and $\eta = \mu$, this recovers Theorem 1.3 in Villani (2003) under the assumption of compactness. If $\eta = \bar{\mu}$ and μ' is upper semi-continuous, this result is a special case of the more general moment-type duality formula; see Rachev and Rüschendorf (1998).

⁴Recall 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.

If $d\eta/d\bar{\mu}$ 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

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

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 (16).

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

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

where

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

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

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

5 MOT-SOT parity

Consider $\boldsymbol{\mu} \in \mathcal{M}(X)^d$ and $\boldsymbol{\nu} \in \mathcal{M}(Y)^d$ where for simplicity X, Y are Euclidean spaces, and recall our notation $\bar{\mu}, \bar{\nu}, \boldsymbol{\mu}', \boldsymbol{\nu}', m_\mu, m_\nu$ from (3) and (8). By the disintegration theorem, there exist measures $\{\mu_{\mathbf{z}}\}_{\mathbf{z} \in \mathbb{R}_+^d}$ such that

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

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

$$\int_X f(x) \bar{\mu}(dx) = \int_{\mathbb{R}_+^d} \int_{A_{\mathbf{z}}} f(x) \mu_{\mathbf{z}}(dx) m_\mu(d\mathbf{z}).$$

Moreover, the family of measures $\{\mu_{\mathbf{z}}\}_{\mathbf{z} \in \mathbb{R}_+^d}$ is uniquely determined for m_μ -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 recall that given probability measures μ, ν , a coupling (ξ_μ, ξ_ν) of μ, ν is called a *martingale transport* if (ξ_μ, ξ_ν) forms a martingale, and we denote by $\mathcal{M}(\mu, \nu)$ the set of all such couplings, which can be further identified as stochastic kernels.

We have seen from Proposition 1 that the existence of a (balanced) simultaneous transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ is equivalent to $m_\mu \succeq_{\text{cx}} m_\nu$. This naturally gives rise to a martingale transport from m_ν to m_μ in view of Strassen's theorem (Strassen (1965)). Such a martingale transport, seen as a coupling, encodes the way we take combinations of the $(\mathbb{R}_+^d$ -valued) derivatives $\boldsymbol{\mu}'$ to form the derivatives $\boldsymbol{\nu}'$. The martingale constraint is equivalent to the constraint of mixing $\boldsymbol{\mu}'$ to get $\boldsymbol{\nu}'$. Essentially, at this step, we do not “distinguish” the points in $A_{\mathbf{z}}$ since $\boldsymbol{\mu}'$ is constant there, but treat the set $A_{\mathbf{z}}$ as a single point; the same applies to $B_{\mathbf{z}'}$. Of course, such a martingale transport may not be unique (in fact, for two-way transports it is unique). This choice of a martingale transport can be seen as the first layer of freedom for a simultaneous transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$. The second layer of freedom is how

to transport on each slice from $A_{\mathbf{z}}$ to $B_{\mathbf{z}}$, where now we do not treat them as single points, but equip the measure $\mu_{\mathbf{z}}, \nu_{\mathbf{z}}$ on them. In comparison, the martingale transport treats them as points, which can be regarded as an “integrated version”. In particular, this extends our results on the two-way transport. In some nice cases, the optimization problem reduces to classic optimization problems that admit explicit solutions.

In this section, we will often encounter transport from $(\mathbb{R}_+^d \times [0, 1], m_{\boldsymbol{\mu}} \times \tau)$ to $(\mathbb{R}_+^d \times [0, 1], m_{\boldsymbol{\nu}} \times \tau)$. To simplify formulas, we will write $\kappa^x(\cdot)$ as $\kappa(x; \cdot)$ for a stochastic kernel κ , where x often has two components.

5.1 Connecting MOT and SOT

Let τ be the Lebesgue measure on $[0, 1]$ and write $\mathcal{R} = \mathbb{R}_+^d \times [0, 1]$. The set of couplings between $(\mathcal{R}, m_{\boldsymbol{\mu}} \times \tau)$ and $(\mathcal{R}, m_{\boldsymbol{\nu}} \times \tau)$ that are backward martingale in the first marginal is given by

$$\{((X, U), (X', U')) \in \widehat{\Pi}(m_{\boldsymbol{\mu}} \times \tau, m_{\boldsymbol{\nu}} \times \tau) \mid \mathbb{E}[X|(X', U')] = X'\},$$

where $\widehat{\Pi}$ is the set of random vectors having distributions in Π . We disintegrate such couplings into stochastic kernels, and denote by $\mathbb{M}_{b,1}$ the corresponding collection of stochastic kernels. Formally, $\mathbb{M}_{b,1}$ is the set

$$\left\{ \hat{\kappa} \in \mathcal{K}(m_{\boldsymbol{\mu}} \times \tau, m_{\boldsymbol{\nu}} \times \tau) \mid \int_{\mathcal{R}} \hat{\kappa}^{(\mathbf{z}, u)}(Z' \times V) \mathbf{z}' \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}) = \int_{Z'} \mathbf{z}' \tau(V) m_{\boldsymbol{\nu}}(d\mathbf{z}') \quad \forall Z' \times V \subseteq \mathcal{R} \right\}.$$

For $\mathbf{z}, \mathbf{z}' \in \mathbb{R}_+^d$, define also the sets of stochastic kernels

$$\mathcal{K}_{\mathbf{z}} = \mathcal{K}(\mu_{\mathbf{z}}, \delta_{\mathbf{z}} \times \tau) \quad \text{and} \quad \tilde{\mathcal{K}}_{\mathbf{z}'} = \mathcal{K}(\delta_{\mathbf{z}'} \times \tau, \nu_{\mathbf{z}'}).$$

Since τ is atomless, there exist kernels $\kappa_{\mathbf{z}} \in \mathcal{K}_{\mathbf{z}}, \tilde{\kappa}_{\mathbf{z}'} \in \tilde{\mathcal{K}}_{\mathbf{z}'}$ that are backward Monge and Monge, respectively.

Theorem 3 (MOT-SOT parity). *Suppose that $\boldsymbol{\mu}(X) = \boldsymbol{\nu}(Y)$. Fix arbitrary collections of kernels $\kappa_{\mathbf{z}} \in \mathcal{K}_{\mathbf{z}}, \tilde{\kappa}_{\mathbf{z}'} \in \tilde{\mathcal{K}}_{\mathbf{z}'}$ indexed by $\mathbf{z}, \mathbf{z}' \in \mathbb{R}_+^d$, where $\kappa_{\mathbf{z}}$ is backward Monge, $\tilde{\kappa}_{\mathbf{z}'}$ is Monge, and $\mathbf{z} \mapsto \kappa_{\mathbf{z}}$ and $\mathbf{z}' \mapsto \tilde{\kappa}_{\mathbf{z}'}$ are measurable. Every $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ can be represented as*

$$\kappa^x(B) = \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\boldsymbol{\mu}'(x)}^x(\boldsymbol{\mu}'(x), du) \hat{\kappa}^{(\boldsymbol{\mu}'(x), u)}(d\mathbf{z}', du) \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(B), \quad x \in X, B \subseteq Y, \quad (19)$$

for some $\hat{\kappa} \in \mathbb{M}_{b,1}$. Conversely, given any $\hat{\kappa} \in \mathbb{M}_{b,1}$, the equation (19) defines a simultaneous transport kernel $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$.

Remark 7. Write $f : \mathcal{R} \rightarrow \mathcal{P}(Y), (\mathbf{z}', u') \mapsto \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}$. Then (19) can be written as, for $x \in X$,

$$\kappa^x = \mathbb{E}[f(\zeta, \xi)] = \mathbb{E} \left[\tilde{\kappa}_{\zeta}^{(\zeta, \xi)} \right]$$

where $(\zeta, \xi) \stackrel{\text{law}}{\sim} \int_0^1 \hat{\kappa}^{(\boldsymbol{\mu}'(x), u)} \kappa_{\boldsymbol{\mu}'(x)}^x(du)$.

Let us first explain the intuition. The uniform measure τ on $[0, 1]$ can be regarded as a parameterization space to keep the information of the space $(A_{\mathbf{z}}, \mu_{\mathbf{z}})$ when we map it to a single point on \mathbb{R}_+^d .⁵ It can be replaced by any atomless measure. The Monge property of $\kappa_{\mathbf{z}}, \tilde{\kappa}_{\mathbf{z}'}$ will allow us to reconstruct the original simultaneous transport because it guarantees that no information is lost at the step where we encode $(A_{\mathbf{z}}, \mu_{\mathbf{z}})$ using a single point.⁶ Theorem 3 says that the way we parameterize this information does not matter—it is possible to fix two collections of parameterizations

$$\begin{array}{ccc}
X & \xrightarrow{\kappa} & Y \\
\begin{array}{c} \uparrow \\ x \mapsto (\boldsymbol{\mu}'(x), \kappa_{\boldsymbol{\mu}'(x)}^x) \end{array} & \begin{array}{c} \downarrow \\ (\mathbf{z}, u) \mapsto T_{\mathbf{z}}(u) \end{array} & \begin{array}{c} \downarrow \\ (\mathbf{z}', u') \mapsto S_{\mathbf{z}'}(u') \end{array} \\
(\mathcal{R}, m_{\boldsymbol{\mu}} \times \tau) & \xrightarrow{\hat{\kappa}} & (\mathcal{R}, m_{\boldsymbol{\nu}} \times \tau) \\
& & \begin{array}{c} \uparrow \\ y \mapsto (\boldsymbol{\nu}'(y), \tilde{\kappa}_{\boldsymbol{\nu}'(y)}^y) \end{array}
\end{array}$$

Figure 4: MOT-SOT parity illustrated with commutative diagram: the double arrows connects the upper half (simultaneous transport) and lower half (martingale transport). Note that the downward arrows are not given by a map in general, which explains why Monge transport may not always exist.

a priori, as long as they have the Monge properties and are measurable. See also Figure 4 for a pictorial representation.

We also need a few technical considerations. To see that Theorem 3 actually makes sense, we need the following.

- (i) Existence of a measurable selection of $\{\kappa_{\mathbf{z}'}\}_{\mathbf{z}' \in \mathbb{R}_+^d}$ and $\{\tilde{\kappa}_{\mathbf{z}'}\}_{\mathbf{z}' \in \mathbb{R}_+^d}$ satisfying the Monge properties. When X, Y are Euclidean spaces, this is guaranteed by a measurable selection of optimal plans (Villani (2009, Corollary 5.22)) for the quadratic cost, which are given by deterministic maps; see Gangbo and McCann (1996).⁷ A sufficient condition for Polish spaces is given by Villani (2009, Theorem 5.30).
- (ii) Joint measurability of $\tilde{\kappa}_{\mathbf{z}', u'}^{\mathbf{z}', u'}(B)$ in (\mathbf{z}', u') (so that the integral in (19) makes sense). To see this, denote the transport plan corresponding to $\tilde{\kappa}_{\mathbf{z}'}$ by $\tilde{\pi}_{\mathbf{z}'}$, which is a probability measure on $\{\mathbf{z}'\} \times [0, 1] \times B_{\mathbf{z}'}$. Define $\tilde{\pi} = \int \tilde{\pi}_{\mathbf{z}'} m_{\boldsymbol{\nu}}(d\mathbf{z}')$, which is a probability measure on $\mathcal{R} \times Y$ and is well-defined by (i). Next, disintegrate $\tilde{\pi}$ in the first two coordinates, (\mathbf{z}', u') to get a family of measures $\{\tilde{\kappa}_{\mathbf{z}', u'}(\cdot)\}$ that is jointly measurable in (\mathbf{z}', u') . By uniqueness of disintegration and since $B_{\mathbf{z}'}, \mathbf{z}' \in \mathbb{R}_+^d$ are disjoint, we must have $\tilde{\kappa}_{\mathbf{z}', u'}(\cdot) = \tilde{\kappa}_{\mathbf{z}', u'}^{\mathbf{z}', u'}(\cdot)$.

Example 7. Let $X = Y = \{0, 1\}$ and $\boldsymbol{\mu}(\{0\}) = (1/3, 2/3)$, $\boldsymbol{\mu}(\{1\}) = (2/3, 1/3)$, $\boldsymbol{\nu}(\{0\}) = (1/3, 1/3)$, and $\boldsymbol{\nu}(\{1\}) = (2/3, 2/3)$. We have $m_{\boldsymbol{\mu}} = (\delta_{(4/3, 2/3)} + \delta_{(2/3, 4/3)})/2$ and $m_{\boldsymbol{\nu}} = \delta_{(1, 1)}$. The backward martingale transport from $m_{\boldsymbol{\mu}}$ to $m_{\boldsymbol{\nu}}$ is unique. Therefore by Theorem 3, the simultaneous transport from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ is unique, and we can easily check that it is given by $\kappa(0, \{0\}) = \kappa(1, \{0\}) = 1/3$ and $\kappa(0, \{1\}) = \kappa(1, \{1\}) = 2/3$.

A first immediate consequence is the following commutative relation. This can be seen as a special case of a more general commutative relation illustrated by Figure 4 below.

Corollary 3. *Let $\boldsymbol{\mu} \in \mathcal{P}(X)^d$ and $\boldsymbol{\nu} \in \mathcal{P}(Y)^d$ satisfy that $\mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is non-empty. Suppose that $\boldsymbol{\mu}$ is jointly atomless and $m_{\boldsymbol{\mu}}$ is atomless. Then there exists a backward martingale coupling between $m_{\boldsymbol{\mu}}$ and $m_{\boldsymbol{\nu}}$ that is also Monge.⁸ Moreover, if we denote by h the map that induces this Monge transport, then there exists a simultaneous transport map $f \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ satisfying*

$$\boldsymbol{\nu}'(f(x)) = h(\boldsymbol{\mu}'(x)), \quad x \in X.$$

Finally, we mention a more general version of Theorem 3 for the unbalanced setting. Apart from the constraints $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ we add an extra constraint that $\kappa \in \mathcal{K}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}})$ (which is automatically satisfied in the balanced case; see (A.15)), i.e., define

$$\tilde{\mathcal{K}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu}) \mid \kappa \in \mathcal{K}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}})\}.$$

⁵When $\boldsymbol{\mu}'$ (resp. $\boldsymbol{\nu}'$) is injective, we may remove the parameterization space on the $m_{\boldsymbol{\mu}}$ (resp. $m_{\boldsymbol{\nu}}$) side.

⁶To see the Monge property is crucial, imagine we use independent couplings for both—it will not yield the set of all simultaneous transports.

⁷More precisely, we replace τ by $[0, 1]^\ell$, where ℓ is the larger dimension of X and Y . As commented above, this will not affect the result.

⁸A coupling (ξ_1, ξ_2) is *backward martingale* if $\mathbb{E}[\xi_1 | \xi_2] = \xi_2$, that is, (ξ_2, ξ_1) forms a martingale.

There exists a parity relation between $\tilde{\mathcal{K}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and the set of stochastic kernels from $(\mathcal{R}, m_{\boldsymbol{\mu}} \times \tau)$ to $(\mathcal{R}, m_{\boldsymbol{\nu}} \times \tau)$ that is backward *submartingale* in the first marginal. The proof is very similar to the proof of Theorem 3 and we omit the details.

5.2 Optimality of simultaneous transport and examples

In this section, for simplicity we will keep the reference measure $\eta = \bar{\mu}$. The general case $\eta \ll \bar{\mu}$ follows by modifying the cost function $c(x, y)$.

Suppose that we are given $\kappa_{\mathbf{z}} \in \mathcal{K}_{\mathbf{z}}$, $\tilde{\kappa}_{\mathbf{z}'}$ $\in \tilde{\mathcal{K}}_{\mathbf{z}'}$, $\mathbf{z}, \mathbf{z}' \in \mathbb{R}_+^d$ measurable as in the setup of Theorem 3. For $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ with representation (19), let us compute the cost of the associated simultaneous transport:

$$\begin{aligned}
\mathcal{C}(\kappa) &= \int_{X \times Y} c(x, y) \kappa^x(dy) \bar{\mu}(dx) \\
&= \int_{\mathbb{R}_+^d} \int_{A_{\mathbf{z}}} \int_Y c(x, y) \kappa^x(dy) \mu_{\mathbf{z}}(dx) m_{\boldsymbol{\mu}}(d\mathbf{z}) \\
&= \int_{\mathbb{R}_+^d} \int_{A_{\mathbf{z}}} \int_Y c(x, y) \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \hat{\kappa}^{(\mathbf{z}, u)}(d\mathbf{z}', du') \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(dy) \mu_{\mathbf{z}}(dx) m_{\boldsymbol{\mu}}(d\mathbf{z}) \\
&= \int_{\mathcal{R}} \int_{\mathcal{R}} \hat{\kappa}^{(\mathbf{z}, u)}(d\mathbf{z}', du') \left(\int_{A_{\mathbf{z}}} \int_Y c(x, y) \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(dy) \mu_{\mathbf{z}}(dx) \right) m_{\boldsymbol{\mu}}(d\mathbf{z}) \\
&= \int_{\mathcal{R}} \int_{\mathcal{R}} \hat{\kappa}^{(\mathbf{z}, u)}(d\mathbf{z}', du') \left(\int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}'}} c(x, y) \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(dy) \mu_{\mathbf{z}}(dx) \right) m_{\boldsymbol{\mu}}(d\mathbf{z}), \quad (20)
\end{aligned}$$

where the third equality follows since $\boldsymbol{\mu}'(x) = \mathbf{z}$ for $x \in A_{\mathbf{z}}$. The infimum cost $\inf \mathcal{C}(\kappa)$ may be computed by taking an infimum over all $\hat{\kappa}$ while fixing $\kappa_{\mathbf{z}} \in \mathcal{K}_{\mathbf{z}}$ and $\tilde{\kappa}_{\mathbf{z}'} \in \tilde{\mathcal{K}}_{\mathbf{z}'}$ for $\mathbf{z}, \mathbf{z}' \in \mathbb{R}_+^d$. Alternatively, taking an infimum over all $\kappa_{\mathbf{z}}$, $\tilde{\kappa}_{\mathbf{z}'}$, and $\hat{\kappa}$ leads to the same value.

Note that the measure

$$\gamma_{\mathbf{z}, \mathbf{z}', u'}(V) := \int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}'}} c(x, y) \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times V) \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(dy) \mu_{\mathbf{z}}(dx)$$

satisfies $\gamma_{\mathbf{z}, \mathbf{z}', u'} \ll \tau$ for all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}_+^d$ and $u' \in [0, 1]$. By Theorem 58 in [Dellacherie and Meyer \(2011\)](#), there exists a jointly measurable function $\hat{c} : \mathcal{R}^2 \rightarrow \mathbb{R}$ such that

$$\int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}'}} c(x, y) \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(dy) \mu_{\mathbf{z}}(dx) = \hat{c}(\mathbf{z}, u, \mathbf{z}', u') \tau(du).$$

We are then left with

$$\begin{aligned}
\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\kappa) &= \inf_{\hat{\kappa} \in \mathbb{M}_{b,1}} \int_{\mathcal{R}} \int_{\mathcal{R}} \hat{\kappa}^{(\mathbf{z}, u)}(d\mathbf{z}', du') \hat{c}(\mathbf{z}, u, \mathbf{z}', u') m_{\boldsymbol{\mu}} \times \tau(d\mathbf{z}, du) \\
&= \inf \mathbb{E}[\hat{c}(Z, U; Z', U')]
\end{aligned}$$

where the last infimum is taken over all possible couplings $(Z, U; Z', U') \stackrel{\text{law}}{\sim} (m_{\boldsymbol{\mu}} \times \tau) \otimes \hat{\kappa}$ where $\mathbb{E}[Z|(Z', U')] = Z'$. This becomes an optimal transport problem on “backward martingale over the first marginal” in \mathbb{R}^{d+1} . In fact, we may reduce the dimension to $(d-1) + 1$, simply because the Radon–Nikodym derivatives sum up to a constant. The connection to MOT also explains some bizarre behaviors of SOT. For example, in Example 5, the transport cost is a constant if $\boldsymbol{\mu}$ is linear and the cost function is quadratic. This stems from the well-known fact that the quadratic cost is trivial for MOT. We next discuss a few special classes and explicitly solvable examples below.

Example 8. When $\boldsymbol{\mu}'$ is injective and $Y = \mathbb{R}$, we may pick $\kappa_{\mathbf{z}}$ to map the point $(\boldsymbol{\mu}')^{-1}(\mathbf{z})$ to τ and $\tilde{\kappa}_{\mathbf{z}'}$ the comonotone coupling between $\nu_{\mathbf{z}'}$ and τ . We arrive at

$$\hat{c}(\mathbf{z}, u, \mathbf{z}', u') = \int_{B_{\mathbf{z}'}} c((\boldsymbol{\mu}')^{-1}(\mathbf{z}), y) \tilde{\kappa}_{\mathbf{z}'}^{u'}(dy) = c((\boldsymbol{\mu}')^{-1}(\mathbf{z}), F_{\mathbf{z}'}^{\leftarrow}(u')),$$

where $F_{\mathbf{z}'}^{\leftarrow}$ is the left-continuous inverse of the measure $\nu_{\mathbf{z}'}$.

Example 9. Assume that $c(x, y)$ depends only on $\boldsymbol{\mu}'(x)$ and $\boldsymbol{\nu}'(y)$ (for instance, when both $\boldsymbol{\mu}'$ and $\boldsymbol{\nu}'$ are injective), then with $\hat{c}(\boldsymbol{\mu}'(x), \boldsymbol{\nu}'(y)) = c(x, y)$, we have for any $V \subseteq [0, 1]$,

$$\begin{aligned} \int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}'}} c(x, y) \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times V) \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')} (dy) \mu_{\mathbf{z}}(dx) \\ = \hat{c}(\mathbf{z}, \mathbf{z}') \int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}'}} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times V) \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')} (dy) \mu_{\mathbf{z}}(dx) = \hat{c}(\mathbf{z}, \mathbf{z}') \tau(V). \end{aligned}$$

This yields

$$\inf_{\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\kappa) = \inf_{\hat{\kappa} \in \mathcal{M}(m_{\boldsymbol{\nu}}, m_{\boldsymbol{\mu}})} \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^d} \hat{\kappa}^{\mathbf{z}'}(d\mathbf{z}) \hat{c}(\mathbf{z}, \mathbf{z}') m_{\boldsymbol{\nu}}(d\mathbf{z}'). \quad (21)$$

Intuitively, in the optimal transport problem, we may remove the extra dimension where τ is supported, because the cost function depends only on the \mathbb{R}_+^d -coordinate. In particular, (21) is now equivalent to an MOT problem.

In this setting, it is possible to recover an optimal simultaneous transport from an optimal martingale coupling π : let $\hat{\kappa}^{(\mathbf{z}, u)}$ follow π in the first coordinate and identity in the second, then the kernel κ defined in (19) gives an optimal simultaneous transport.

An immediate consequence of Example 9 is that MOT on compact Euclidean spaces can be realized as a special case of SOT. This is called the *MOT-SOT parity* as suggested by the title of this section.

Example 10. We may strengthen Theorem 3 when both $\boldsymbol{\mu}'(x)$ and $\boldsymbol{\nu}'(y)$ are injective, as follows. It can be checked that the same result goes through if we remove our parameterization space $([0, 1], \tau)$. In other words, there is a correspondence between simultaneous transport and backward martingale transport (on \mathbb{R}^d). Using Theorem 2.1 of Nutz et al. (2022), we thus obtain a stronger version of Proposition 1, that a Monge simultaneous transport exists when $m_{\boldsymbol{\mu}}$ is atomless.⁹ If moreover c , $(\boldsymbol{\mu}')^{-1}$, and $(\boldsymbol{\nu}')^{-1}$ are continuous and c is bounded, we conclude using Corollary 2.4 of Nutz et al. (2022) the equivalence between Monge and Kantorovich costs, complementing Theorem 1.

Remark 8. Recall that the supremum in the MOT duality formula may not always be attained. Indeed, a simple counterexample is given by Beiglböck et al. (2017), Example 8.2. By the MOT-SOT parity, the supremum in our SOT duality formula (18) is not always attained either. For details, see the dual MOT-SOT parity discussions in Section A.5. A direct construction of a counterexample can also be found in Gover (2020).

Example 11. Consider $\boldsymbol{\mu}'$ taking values on only two points, say \mathbf{z}_1 and \mathbf{z}_2 (so that $\boldsymbol{\nu}'$ takes values only on the line segment joining these two points). In this case, there is a further decomposition of the transport $\hat{\kappa} \in \mathcal{M}_{b,1}$, into a collection of independent transports from $([0, 1], \tau)$ to $([0, 1], \tau)$. This is because any such $\hat{\kappa}$ must transport a positive fraction of the measure on $\mathbf{z}' \times [0, 1]$ to $\mathbf{z}_1 \times [0, 1]$ and the rest to $\mathbf{z}_2 \times [0, 1]$. Further solutions are available if \hat{c} is nicely behaved (however, this might be difficult to achieve in general). In the special case where $\boldsymbol{\mu}'$ is also injective, every simultaneous transport has the same cost.

⁹This is also true if we only assume $\boldsymbol{\nu}'$ is injective, by removing the parameterization space on the $m_{\boldsymbol{\nu}}$ side.

Example 12. Let $d = 2$ and consider measures μ, ν on \mathbb{R} such that $d\mu_1/d\bar{\mu}(x)$ and $d\nu_1/d\bar{\nu}(y)$ are affine in x, y respectively with positive slopes (the cases with negative slopes are analogous). Assume that $c(x, y) = h(x - y)$ for some differentiable h with h' strictly convex, and such that $|c(x, y)| \leq a(x) + b(y)$ for some $a \in L^1(m_\mu)$, $b \in L^1(m_\nu)$. This SOT problem is then reduced to a MOT problem on \mathbb{R} , in the form of (21), with a martingale Spence-Mirrlees cost function. Using Theorem 1.7 in Beiglböck and Juillet (2016), the MOT problem (21) is uniquely solved by the *left-curtain transport* from m_ν to m_μ . This coupling uniquely induces a simultaneous transport from μ to ν since μ', ν' are injective. This easily generalizes to when $f_1 := d\mu_1/d\bar{\mu}$ and $g_1 := d\nu_1/d\bar{\nu}$ not being linear. For example, assuming $f_1'', f_1', g_1', c_{xxy}$ all being positive suffices.

5.3 Decomposition of two-way transport

Define an equivalence relation \simeq among \mathbb{R}_+^d -valued probability measures as follows: $\mu \simeq \nu$ if $m_\mu = m_\nu$ (or equivalently, both $\Pi(\mu, \nu)$ and $\Pi(\nu, \mu)$ are non-empty). For $P \in \mathcal{P}(\mathbb{R}^d)$, we define

$$\mathcal{E}_P = \{\mu \in \Pi(X)^d \mid m_\mu = P\},$$

the equivalence class under \simeq . The transitivity of \simeq follows from Corollary 1. We also define the minimum transport cost between μ and ν as

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

Theorem 4 (Decomposition of two-way transport). *Suppose that c is continuous. For $\mu, \nu \in \mathcal{E}_P$ and $\kappa \in \mathcal{K}(\mu, \nu)$, we have $\kappa \in \mathcal{K}(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$ for P -a.e. \mathbf{z} . Moreover, the following are equivalent:*

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

$$\mathcal{C}(\kappa) = \int_{\mathbb{R}_+^d} \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) P(d\mathbf{z}). \quad (22)$$

In particular,

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

Remark 9. That the right-hand side of (22) is indeed well-defined will be discussed in the proof using a measure selection argument.

Remark 10. If $\mu, \nu \in \mathcal{E}_P$ and μ' is injective, the proof of Theorem 4 also indicates that $\mathcal{K}(\mu, \nu)$ consists of a single element, i.e., the simultaneous transport from μ to ν is unique.

A few comments are in place. Roughly speaking, two-way transports exist if and only if each transport from μ to ν (provided it exists) can be inverted to produce a transport from ν to μ . 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 $\mathbf{z} \in \mathbb{R}_+^d$,

$$\sum_{x \in X: \mu'(x) = \mathbf{z}} \bar{\mu}(\{x\}) = \sum_{y \in Y: \nu'(y) = \mathbf{z}} \bar{\nu}(\{y\}). \quad (23)$$

Theorem 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 submodular¹⁰ 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 4. *Suppose that $X = Y = \mathbb{R}$, $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}_P$ and c is continuous and submodular. Then*

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

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

In Appendix A.4, we further discuss an application of Corollary 4 to Wasserstein distances between \mathbb{R}_+^d -valued probability measures. Each of the distances will be defined on some equivalence class \mathcal{E}_P .

As another consequence of Theorem 4, we obtain the following slightly stronger duality result. This duality formula appears in a different form compared to the one in Theorem 2, and it is similar to the duality formula in the classic setting ($d = 1$). This is due to Theorem 4, which is only possible in case of two-way transport problems. Recall that in case $d = 1$, all transport problems are two-way.

Proposition 6. *Suppose that both $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\Pi(\boldsymbol{\nu}, \boldsymbol{\mu})$ are non-empty and $c : X \times Y \rightarrow [0, \infty)$ is uniformly continuous and bounded, then duality holds as*

$$\inf_{\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c d\pi = \sup_{(\phi, \psi) \in \tilde{\Phi}_c} \int_X \phi d\bar{\mu} + \int_Y \psi d\bar{\nu}, \quad (24)$$

where

$$\tilde{\Phi}_c = \left\{ (\phi, \psi) \in L^1(\bar{\mu}) \times L^1(\bar{\nu}) \mid \phi(x) + \psi(y) \leq c(x, y) \text{ if } \boldsymbol{\mu}'(x) = \boldsymbol{\nu}'(y) \right\}.$$

Moreover, both the infimum and supremum in (24) are attained.

Using the Decomposition Theorem, we discuss a few interesting examples illustrating the peculiarities of simultaneous transport (complementing Section 3.3) on an equivalence class \mathcal{E}_P . From classic optimal transport theory ($d = 1$), we first recall the following result.

Proposition 7 (Theorem 1.17 in [Santambrogio \(2015\)](#)). *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 13 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 7 may fail to hold. A counterexample is given by μ being uniform on $[0, 1] \times \{0\}$, and ν uniformly distributed on $[0, 1] \times \{a, b\}$ where $a \neq b$, with $N = 2$ and $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$.

¹⁰A function c on $X \times Y$ is submodular if $c(x, y) + c(x', y') \leq c(x, y') + c(x', y)$ whenever $x \leq x'$ and $y \leq y'$. 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 .

Example 13. 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 $d\mu_2/d\mu_1(x, y) = 2y$ and ν_2 supported on $[0, 1] \times [2, 3]$ such that $d\nu_2/d\nu_1(x, y) = 2 - 4|y - 5/2|$.

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 5, 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.

Proposition 8. *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.*

Example 14. We discuss an example where the optimal transport may not be the linear transport given in Proposition 8. 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 4. Consider $z \in (0, 2)$, then computing the density yields that $d\mu_1/d\bar{\mu}((x, y)) = z$ 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, $d\nu_1/d\bar{\nu}((x, y)) = z$ 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 $c = h_\delta(z) \geq 0$ on which some copies of Gaussian measures are equipped. Direct computation shows that a transport from μ_z to ν_z is given by

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

This is illustrated by the following Figure 5.

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

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

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

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

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 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.

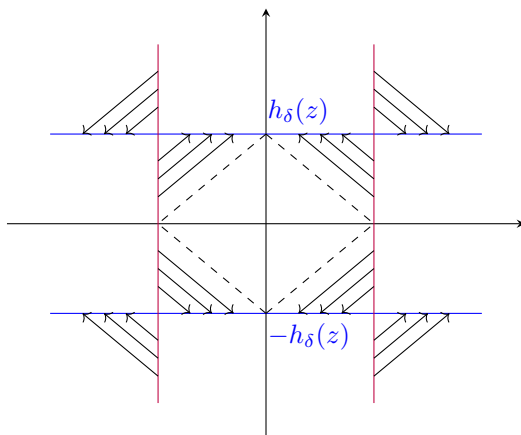


Figure 5: Optimal transport from μ_z to ν_z : the red and blue lines indicate the supports of μ_z and ν_z respectively. Black arrows indicate the transports.

6 Concluding remarks

The simultaneous optimal transport is introduced and studied in this paper. In view of the wide applications of optimal transport in economic studies, such as contract design (Ekeland (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)), there are many directions of SOT for future exploration, in addition to our motivating examples and equilibrium analysis in Section 2 and Appendix C. More broadly, optimal transport also has strong presence in robust risk assessment (Embreechts et al. (2013)), option pricing (Beiglböck et al. (2013)), machine learning (e.g., Peyré and Cuturi (2019)), operations research (e.g., Blanchet and Murthy (2019)) and statistics (e.g., Carlier et al. (2016)), which offer natural locations to look for applications of our new framework.

The framework is shown to be technically 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 directions to explore within the framework of SOT. We discuss a few directions below.

- (i) The MOT-SOT parity (Theorem 3) could potentially pave the path to many future developments of SOT. For example, some results on MOT such as complete duality may be translatable to SOT, shedding light on some of our open questions below. Computational methods for SOT may be developed based on those of MOT; see De March (2018) and Guo and Obłój (2019). Exploration along these directions is left for future study.
- (ii) 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 $d \rightarrow \infty$. 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. Assuming existence of a dominating measure and under some additional assumptions, Gover (2020) proved duality results in infinite dimension via an abstract duality theorem.
- (iii) 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 $\mu^1, \dots, \mu^n \in \mathcal{P}(X)^d$. 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

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

where $c : X^n \rightarrow \mathbb{R}$ 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 $d \geq 2$ and $n \geq 3$. A possible formulation via kernels is given by defining, for each $j \in [d]$, $\mathcal{K}(\mu_j^1, \dots, \mu_j^n) = \{\kappa : X \rightarrow \mathcal{P}(X^{n-1}) \mid \kappa_{\#} \mu_j^1 \in \Pi(\mu_j^2, \dots, \mu_j^n)\}$ and letting $\mathcal{K}(\mu^1, \dots, \mu^n) = \bigcap_{j=1}^d \mathcal{K}(\mu_j^1, \dots, \mu_j^n)$. Each $\kappa \in \mathcal{K}(\mu^1, \dots, \mu^n)$ corresponds to a multi-marginal simultaneous transport, with $n = 2$ corresponding our setting and $d = 1$ corresponding to the classic multi-marginal transport setting.

- (iv) Recall that in the Monge formulation, the objective is to minimize

$$\mathcal{C}_\eta(T) = \int_X c(x, T(x)) \eta(dx). \quad (25)$$

One may consider a nonlinear reference, i.e., η in (25) is a Choquet capacity¹¹ instead of a measure. The motivation of this formulation can be easily explained in the context of [Example 3](#), where the objective is

$$\text{to minimize } \int f(L) d\eta, \quad \text{subject to } L \in \mathcal{T}(\mu, \nu). \quad (26)$$

By taking η as a capacity, (26) 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 $\int f(L) d\eta$ is the (left) α -quantile of $f(L)$, and the problem (26) 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 $L \in \mathcal{T}(\mu, \nu)$ for a general mapping $\mathcal{R} : \mathcal{L} \rightarrow \mathbb{R}$, such as many other quantities developed in decision theory (e.g., [Hansen and Sargent \(2001\)](#); [Maccheroni et al. \(2006\)](#)). Alternatively, instead of choosing η as a capacity, μ and ν may also be chosen as tuples of capacities instead of measures.

- (v) Our optimal transport is allowed to be chosen from the entire set of transports $\mathcal{K}(\mu, \nu)$ (kernel) or $\mathcal{T}(\mu, \nu)$ (Monge). There is an active stream of research on optimal transport with constraints such as MOT and directional optimal transport (e.g., [Nutz and Wang \(2022\)](#)). Adding these constraints to the simultaneous transport gives rise to many new challenges and requires further studies.
- (vi) There are a few technical open questions related to results in this paper.
- (a) We have explained in [Remark 8](#) that the supremum may not always be attained in the duality formula (18). Establishing a complete duality remains a challenging problem, where

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

we refer to [Beiglböck et al. \(2017\)](#), [Nutz and Stebegg \(2018\)](#), and [De March and Touzi \(2019\)](#) for relevant results for MOT. Sufficient conditions for attainability were established by [Gover \(2020\)](#).

- (b) There are several places in the paper where compactness of X and Y is assumed. For instance, compactness is used in [Theorem 1](#) and [Proposition 5](#). We expect that this condition can be removed. In particular, we note that [Theorem 1](#) for $d = 1$ holds without the compactness assumption as shown by [Pratelli \(2007\)](#).

Acknowledgements

The authors thank Itai Ashlagi, Jose Blanchet, Job Boerma, Ibrahim Ekren, Henry Lam, Fabio Maccheroni, Max Nendel, Marcel Nutz, Giovanni Puccetti, Ludger Rüschendorf, Aleh Tsyvinski, and Kelvin Shuangjian Zhang for helpful comments on a previous version of the paper. We are also grateful to three anonymous referees for their helpful feedback. RW acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2018-03823, RGPAS-2018-522590).

Appendices

In the appendices, we first present proofs and some additional results in [Appendix A](#). We then collect a small review of literature on various generalizations of optimal transport in [Appendix B](#). Finally, we discuss an application of SOT duality to a labour market equilibrium model in [Appendix C](#).

A Proofs and additional results

A.1 Proofs of results in [Section 3](#)

Proof of [Proposition 1](#). 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 $\mu(X) = \nu(Y)$ is assumed. \square

Proof of [Proposition 2](#). 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 $\alpha \geq \beta \geq 1$ (case i) and decreasing log-convexity of $t \mapsto \sigma_t$ means $\alpha \leq \beta \leq 1$ (case ii).

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

- (a) $\mathcal{K}((\mu_1, \mu_2), (\mu_2, \mu_3)) \neq \emptyset$;
- (b) $\frac{d\mu_2}{d\mu_3} \Big|_{\mu_3} \preceq_{\text{cx}} \frac{d\mu_1}{d\mu_2} \Big|_{\mu_2}$;
- (c) $\frac{d\mu_3}{d\mu_2} \Big|_{\mu_2} \preceq_{\text{cx}} \frac{d\mu_2}{d\mu_1} \Big|_{\mu_1}$,

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

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

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

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

and

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

Therefore, $\mathcal{K}((\mu_1, \mu_2), (\mu_2, \mu_3)) \neq \emptyset$ is equivalent to

$$\beta^{1/2} e^{\xi^2(\frac{1}{2} - \frac{\beta}{2})} \preceq_{\text{cx}} \alpha^{1/2} e^{\xi^2(\frac{1}{2} - \frac{\alpha}{2})} \iff \beta^{-1/2} e^{\xi^2(\frac{1}{2} - \frac{1}{2\beta})} \preceq_{\text{cx}} \alpha^{-1/2} e^{\xi^2(\frac{1}{2} - \frac{1}{2\alpha})}. \quad (\text{A.1})$$

A convenient result we use here is Corollary 1.2 of [Hirsch et al. \(2011\)](#), which says that the stochastic process $((1 + 2t)^{1/2} e^{-\xi^2 t})_{t \geq 0}$ is a peacock; moreover, it is obvious that this process is non-stationary. This implies that, for $x, y \geq 1$, $\sqrt{y} e^{\xi^2(\frac{1}{2} - \frac{y}{2})} \preceq_{\text{cx}} \sqrt{x} e^{\xi^2(\frac{1}{2} - \frac{x}{2})}$ if and only if $y \leq x$. Hence, if $\alpha, \beta \geq 1$, then (A.1) is equivalent to $\beta \leq \alpha$, thus case (i), and if $\alpha, \beta \leq 1$, then (A.1) is equivalent to $\beta \geq \alpha$, 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 (A.1) is not possible. \square

The condition in Proposition 2 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 Theorem 4 in Appendix A.3, $\kappa(x; \{\pm x/2\}) = \kappa(x; \{\pm x/\sqrt{2}\}) = 1$, a contradiction.

Proof of Proposition 3. 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 $Y_1 = \{y \in Y \mid \nu'_1(y) \geq 1\} = Y_2^c$. Also fix an arbitrary $\kappa \in \mathcal{K}(\mu, \nu)$. Then for $B \subseteq Y_1$, we have

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

In fact, this holds with κ replaced by $\kappa|_{X_1}$, where we define $X_1 = \{x \in X \mid \mu'_1(x) \geq 1\} = X_2^c$. Similarly, for $B \subseteq Y_2$, we have

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

Therefore, denoting by η_1 the restriction of η on the set $\{x \in X \mid \mu'_1(x) \geq 1\}$ and $\eta_2 = \eta - \eta_1$, we obtain

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

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 $\kappa \in \mathcal{K}(\mu, \nu)$ proves (10). \square

A.2 Proofs of results in Section 4

Proof of Proposition 4. 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

$$\pi(A \times B) = \int_A \kappa(x; B) \eta(dx) \text{ for all } A \subseteq X, B \subseteq Y. \quad (\text{A.2})$$

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

$$\int_{X \times B} f(x) \pi(dx, dy) = \int_X \kappa(x; B) f(x) \eta(dx),$$

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

$$\nu_j(B) \leq \int_X \kappa(x; B) \mu_j(dx) = \int_X \kappa(x; B) \frac{d\mu_j}{d\eta}(x) \eta(dx) = \int_{X \times B} \frac{d\mu_j}{d\eta}(x) \pi(dx, dy).$$

In addition, for any $A \subseteq X$, $\pi(A \times Y) = \int_A \kappa(x; Y) \eta(dx) = \eta(A)$, 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 $\kappa : \mathbb{R} \rightarrow \mathcal{P}(Y)$ such that for $A \subseteq X$, $B \subseteq Y$,

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

which is exactly (A.2). Similarly as above, we have

$$\nu_j(B) \leq \int_{X \times B} \frac{d\mu_j}{d\eta}(x) \pi(dx, dy) = \int_X \kappa(x; B) \frac{d\mu_j}{d\eta}(x) \eta(dx) = \int_X \kappa(x; B) \mu_j(dx),$$

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

Next, we turn to the proof of Theorem 1. We first show a useful lemma, Lemma A.2 below, 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 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 A.1). 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 A.1. Let $(\Omega, \mathcal{G}, \mu)$ be a measure space. We say that (\mathcal{G}, μ) 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 $A' \subseteq A$, $A' \in \mathcal{F}$, 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 A.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 A.1. This equivalence is anticipated because, in the unconditional setting, any set being divisible (corresponding to Definition A.1) is equivalent to the existence of a continuously distributed random variable (corresponding to Definition 1); see e.g., Lemma D.1 of Vovk and Wang (2021).

Lemma A.1. *Let μ be any strictly positive convex combination of $\boldsymbol{\mu} \in \mathcal{P}(X)^d$. Then μ is jointly atomless if and only if $(\mathcal{B}(X), \mu)$ is atomless conditionally to $\sigma(d\boldsymbol{\mu}/d\mu)$.*

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). \square

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

Lemma A.2. *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 $1 \leq j \leq m$ where $m \leq d$. 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 $\mathcal{F} = \sigma(d\boldsymbol{\mu}/d\mu) = \sigma(d\mu_1/d\mu, \dots, d\mu_m/d\mu)$. 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 (μ_1, \dots, μ_m) is jointly atomless. Using Lemma A.1, (\mathcal{B}, μ) is atomless conditionally to \mathcal{F} . By definition, there exists $A' \subseteq A$, $A' \in \mathcal{B}$ such that

$$\mathbb{E}^\mu[\mathbb{1}_A|\mathcal{F}] > 0 \implies 0 < \mathbb{E}^\mu[\mathbb{1}_{A'}|\mathcal{F}] < \mathbb{E}^\mu[\mathbb{1}_A|\mathcal{F}]. \quad (\text{A.3})$$

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, (A.3) leads to

$$\mathbb{E}^{\mu_B}[\mathbb{1}_A|\mathcal{F}_B] > 0 \implies 0 < \mathbb{E}^{\mu_B}[\mathbb{1}_{A'}|\mathcal{F}_B] < \mathbb{E}^{\mu_B}[\mathbb{1}_A|\mathcal{F}_B] \quad (\text{A.4})$$

Note also that $A' \in \mathcal{B}_B$ by definition. Therefore, by treating μ_B as a probability measure on \mathcal{B}_B , (A.4) implies that (\mathcal{B}_B, μ_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 A.1 again, we conclude that $\boldsymbol{\mu}_B$ is jointly atomless. \square

Proof of Theorem 1. 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 $\frac{d\eta}{d\bar{\mu}}(x) \geq \delta$ 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 ,

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

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

$$\boldsymbol{\mu}^{i,n} := \boldsymbol{\mu}|_{K_{i,n}} \text{ and for } B \subseteq X, \boldsymbol{\nu}^{i,n}(B) := \int_{K_{i,n}} \kappa(x; B) \boldsymbol{\mu}(dx). \quad (\text{A.5})$$

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

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

It is also easy to check that $\kappa_{i,n} \in \mathcal{K}(\tilde{\boldsymbol{\mu}}^{i,n}, \tilde{\boldsymbol{\nu}}^{i,n})$. By Proposition 1, $(\tilde{\boldsymbol{\mu}}^{i,n})'|_{\tilde{\mu}^{i,n}} \succeq_{\text{cx}} (\tilde{\boldsymbol{\nu}}^{i,n})'|_{\tilde{\nu}^{i,n}}$. By Lemma A.2, $\tilde{\boldsymbol{\mu}}^{i,n}$ is jointly atomless, so that applying Proposition 1 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 $T_{i,n} : K_{i,n} \rightarrow X$ such that $\boldsymbol{\mu}^{i,n} \circ T_{i,n}^{-1} = \boldsymbol{\nu}^{i,n}$. By gluing these, we obtain a Monge transport $T_n : X \rightarrow X$. Note that $T_n \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\nu})$ since

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

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

$$\mathcal{C}_\eta(T_n) = \int_{X \times X} c(x, y) \eta \otimes \kappa_n(dx, dy) \rightarrow \int_{X \times X} c(x, y) \eta \otimes \kappa(dx, dy). \quad (\text{A.6})$$

Let us define cost functions

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

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

$$\int_{X \times X} |\bar{c}_n(x, y) - c(x, y)| \eta \otimes \kappa(dx, dy) \rightarrow 0. \quad (\text{A.7})$$

On the other hand,

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

Applying this in the second inequality below yields that

$$\begin{aligned} \mathcal{C}_\eta(T_n) &\leq \int_{X \times X} \bar{c}_n(x, y) \eta \otimes \kappa_n(dx, dy) \\ &= \sum_{i \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} \sup_{x \in K_{i,n}, y \in K_{\ell,n}} c(x, y) \int_{K_{i,n} \times K_{\ell,n}} \eta \otimes \kappa_n(dx, dy) \\ &\leq \left(1 + \frac{1}{n}\right) \sum_{i \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} \sup_{x \in K_{i,n}, y \in K_{\ell,n}} c(x, y) \int_{K_{i,n} \times K_{\ell,n}} \eta \otimes \kappa(dx, dy) \\ &= \left(1 + \frac{1}{n}\right) \int_{X \times X} \bar{c}_n(x, y) \eta \otimes \kappa(dx, dy). \end{aligned} \quad (\text{A.8})$$

Combining (A.7) and (A.8), and since $c \geq 0$, we obtain

$$\limsup_{n \rightarrow \infty} \mathcal{C}_\eta(T_n) \leq \int_{X \times X} c(x, y) \eta \otimes \kappa(dx, dy).$$

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

The liminf part is similar. We have thus proved (A.6).

In the general case where $d\eta/d\bar{\mu}$ 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 $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$,

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

This completes the proof. \square

Proof of Proposition 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 $J_{n_{k_\ell}} \rightarrow \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathcal{C}(\pi)$.

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 $d\boldsymbol{\mu}/d\eta$ is continuous, the operations defining $\Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$ is continuous with respect to weak topology in (13), thus we have $\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$. Since $c(x, y)$ is continuous, this gives that

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

Taking infimum yields that

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

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

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

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

The key to proving Theorem 2 is the following minimax theorem, which could be found in [Adams and Hedberg \(1999, Theorem 2.4.1\)](#) and sometimes referred to as Sion's minimax theorem.

Lemma A.3. *Let X be a compact Hausdorff space, Y be an arbitrary set, and $f : X \times Y \rightarrow \mathbb{R} \cup \{\infty\}$. Assume that f is lower semi-continuous in x for each fixed y , convex in x , and concave in y . Then*

$$\min_{x \in X} \sup_{y \in Y} f(x, y) = \sup_{y \in Y} \min_{x \in X} f(x, y).$$

Proof of Theorem 2. The \geq direction of (18) being obvious, we focus on the \leq part. We first assume c is bounded continuous. For a Polish space X , we denote by $C_b(X)$ the space of all bounded continuous functions on X . First observe that by definition (14), for $\pi \in \Pi_\eta(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}})$,

$$\begin{aligned} & \sup_{\substack{\phi \in C_b(X) \\ \boldsymbol{\psi} \in C_b^d(Y)}} \left\{ \int_X \phi d\eta - \int_{X \times Y} \phi(x) \pi(dx, dy) + \int_Y \boldsymbol{\psi}^\top d\boldsymbol{\nu} - \int_{X \times Y} \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) \pi(dx, dy) \right\} \\ &= \begin{cases} 0 & \text{if } \pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}); \\ \infty & \text{elsewhere.} \end{cases} \end{aligned} \tag{A.9}$$

For $\pi \in \Pi_\eta(\bar{\mu}, \bar{\nu})$, we have by using (A.9) that

$$\begin{aligned}
& \sup \left\{ \int_{X \times Y} p(x, y) \pi(dx, dy) + \int_X \phi d\eta + \int_Y \boldsymbol{\psi}^\top d\nu \mid p \in C_b(X \times Y), \right. \\
& \quad \left. \phi \in C_b(X), \boldsymbol{\psi} \in C_b^d(Y), \phi(x) + \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) \leq c(x, y) - p(x, y) \right\} \\
&= \sup \left\{ \int_{X \times Y} c(x, y) \pi(dx, dy) + \int_X \phi d\eta - \int_{X \times Y} \phi(x) \pi(dx, dy) \right. \\
& \quad \left. + \int_Y \boldsymbol{\psi}^\top d\nu - \int_{X \times Y} \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) \pi(dx, dy) \mid \phi \in C_b(X), \boldsymbol{\psi} \in C_b^d(Y) \right\} \\
&= \begin{cases} \int_{X \times Y} c(x, y) \pi(dx, dy) & \text{if } \pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}); \\ \infty & \text{elsewhere.} \end{cases} \tag{A.10}
\end{aligned}$$

Since $d\bar{\mu}/d\eta$ is bounded continuous, the set $\Pi_\eta(\bar{\mu}, \bar{\nu})$ is weakly compact. Using (A.10) and Lemma A.3, we obtain

$$\begin{aligned}
& \min_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c(x, y) \pi(dx, dy) \\
&= \min_{\pi \in \Pi_\eta(\bar{\mu}, \bar{\nu})} \sup \left\{ \int_{X \times Y} p(x, y) \pi(dx, dy) + \int_X \phi d\eta + \int_Y \boldsymbol{\psi}^\top d\nu \mid \phi \in C_b(X), \right. \\
& \quad \left. \boldsymbol{\psi} \in C_b^d(Y), p \in C_b(X \times Y), \phi(x) + \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) \leq c(x, y) - p(x, y) \right\} \\
&= \sup \left\{ \min_{\pi \in \Pi_\eta(\bar{\mu}, \bar{\nu})} \int_{X \times Y} p(x, y) \pi(dx, dy) + \int_X \phi d\eta + \int_Y \boldsymbol{\psi}^\top d\nu \mid \phi \in C_b(X), \right. \\
& \quad \left. \boldsymbol{\psi} \in C_b^d(Y), p \in C_b(X \times Y), \phi(x) + \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) \leq c(x, y) - p(x, y) \right\}.
\end{aligned}$$

By duality for classic optimal transport,

$$\begin{aligned}
& \min_{\pi \in \Pi_\eta(\bar{\mu}, \bar{\nu})} \int_{X \times Y} p(x, y) \pi(dx, dy) \\
&= \sup \left\{ \int_X \tilde{\phi} d\eta + \int_Y \tilde{\psi} d\bar{\nu} \mid \tilde{\phi} \in C_b(X), \tilde{\psi} \in C_b(Y), \tilde{\phi}(x) + \tilde{\psi}(y) \frac{d\bar{\mu}}{d\eta}(x) \leq p(x, y) \right\}.
\end{aligned}$$

Rearranging the terms we have

$$\begin{aligned}
& \min_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} c(x, y) \pi(dx, dy) \\
&= \sup \left\{ \int_X \tilde{\phi} d\eta + \int_Y \tilde{\psi} d\bar{\nu} + \int_X \phi d\eta + \int_Y \boldsymbol{\psi}^\top d\boldsymbol{\nu} \mid \tilde{\phi} \in C_b(X), \tilde{\psi} \in C_b(Y), \tilde{\phi}(x) + \tilde{\psi}(y) \leq p(x, y); \right. \\
&\quad \left. \phi \in C_b(X), \boldsymbol{\psi} \in C_b^d(Y), \phi(x) + \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) \leq c(x, y) - p(x, y) \right\} \\
&\leq \sup \left\{ \int_X \tilde{\phi} d\eta + \int_Y \tilde{\psi} d\bar{\nu} + \int_X \phi d\eta + \int_Y \boldsymbol{\psi}^\top d\boldsymbol{\nu} \mid \tilde{\phi} \in C_b(X), \tilde{\psi} \in C_b(Y), \right. \\
&\quad \left. \phi \in C_b(X), \boldsymbol{\psi} \in C_b^d(Y), \phi(x) + \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) + \tilde{\phi}(x) + \tilde{\psi}(y) \leq c(x, y) \right\} \\
&\leq \sup \left\{ \int_X \phi d\eta + \int_Y \boldsymbol{\psi}^\top d\boldsymbol{\nu} \mid (\phi, \boldsymbol{\psi}) \in C_b(X) \times C_b^d(Y), \phi(x) + \boldsymbol{\psi}(y)^\top \frac{d\boldsymbol{\mu}}{d\eta}(x) \leq c(x, y) \right\},
\end{aligned}$$

thus proving the duality formula (18) in the case where c is bounded continuous.

Consider the general case where c is lower semi-continuous, possibly taking values in $\mathbb{R} \cup \{\infty\}$. 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 $\varphi^d(\phi, \boldsymbol{\psi}) := \int_X \phi d\eta + \int_Y \boldsymbol{\psi}^\top d\boldsymbol{\nu}$. Also write $I_n(\pi) = \int_{X \times Y} c_n d\pi$. We aim to show that

$$\inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} I(\pi) \leq \sup_n \inf_{\pi \in \Pi_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})} I_n(\pi) \leq \sup_n \sup_{(\phi, \boldsymbol{\psi}) \in \Phi_{c_n}} \varphi^d(\phi, \boldsymbol{\psi}) \leq \sup_{(\phi, \boldsymbol{\psi}) \in \Phi_c} \varphi^d(\phi, \boldsymbol{\psi}). \quad (\text{A.11})$$

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

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

Letting $\varepsilon \rightarrow 0$ proves the first inequality of (A.11). Combining with the trivial bound

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

completes the proof of (18).

To show that the infimum of (18) 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 $d\boldsymbol{\mu}/d\eta$ is continuous) and

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

This shows the desired attainability. \square

A.3 Proofs of results in Section 5

Proof of Theorem 3. First we prove the easy direction. Suppose that $\kappa_{\mathbf{z}} \in \mathcal{K}_{\mathbf{z}}$, $\hat{\kappa} \in \mathbb{M}_{b,1}$, and $\tilde{\kappa}_{\mathbf{z}'} \in \tilde{\mathcal{K}}_{\mathbf{z}'}$ for all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}_+^d$. Fix a measurable set $B \subseteq Y$. Since $\kappa_{\mathbf{z}} \in \mathcal{K}_{\mathbf{z}}$, for $V \subseteq [0, 1]$,

$$\int_{A_{\mathbf{z}}} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times V) \mu_{\mathbf{z}}(dx) = \tau(V).$$

Therefore we have

$$\begin{aligned} \int_X \kappa^x(B) \boldsymbol{\mu}(dx) &= \int_X \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\boldsymbol{\mu}'(x), du) \hat{\kappa}(\boldsymbol{\mu}'(x), u)(d\mathbf{z}', du') \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(B) \boldsymbol{\mu}'(x) \bar{\mu}(dx) \\ &= \int_{\mathbb{R}_+^d} \int_{A_{\mathbf{z}}} \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \hat{\kappa}(\mathbf{z}, u)(d\mathbf{z}', du') \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(B) \mathbf{z} \mu_{\mathbf{z}}(dx) m_{\boldsymbol{\mu}}(d\mathbf{z}) \\ &= \int_{\mathcal{R}} \int_{\mathcal{R}} \hat{\kappa}(\mathbf{z}, u)(d\mathbf{z}', du') \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(B) \mathbf{z} \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}). \end{aligned}$$

Since $\hat{\kappa} \in \mathbb{M}_{b,1}$, it holds for $Z' \subseteq \mathbb{R}_+^d$ and $V \subseteq [0, 1]$,

$$\int_{\mathcal{R}} \hat{\kappa}(\mathbf{z}, u)(Z' \times V) \mathbf{z} \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}) = \int_{Z'} \mathbf{z}' \tau(V) m_{\boldsymbol{\nu}}(d\mathbf{z}').$$

This gives

$$\begin{aligned} \int_X \kappa^x(B) \boldsymbol{\mu}(dx) &= \int_{\mathcal{R}} \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(B) \mathbf{z}' \tau(du') m_{\boldsymbol{\nu}}(d\mathbf{z}') \\ &= \int_{\mathcal{R}} \int_{B_{\mathbf{z}'}} \mathbb{1}_B(y) \mathbf{z}' \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(dy) \tau(du') m_{\boldsymbol{\nu}}(d\mathbf{z}'). \end{aligned}$$

Using $\tilde{\kappa}_{\mathbf{z}'} \in \tilde{\mathcal{K}}_{\mathbf{z}'}$, we have that for $B \subseteq Y$,

$$\int_{[0,1]} \tilde{\kappa}_{\mathbf{z}'}^{(\mathbf{z}', u')}(B) \tau(du') = \nu_{\mathbf{z}'}(B).$$

We conclude that

$$\int_X \kappa^x(B) \boldsymbol{\mu}(dx) = \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \mathbb{1}_B(y) \mathbf{z}' \nu_{\mathbf{z}'}(dy) m_{\boldsymbol{\nu}}(d\mathbf{z}') = \boldsymbol{\nu}(B).$$

Consider now kernels $\kappa_{\mathbf{z}} \in \mathcal{K}_{\mathbf{z}}$, $\tilde{\kappa}_{\mathbf{z}'} \in \tilde{\mathcal{K}}_{\mathbf{z}'}$, $\mathbf{z}, \mathbf{z}' \in \mathbb{R}_+^d$ and $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ as fixed, where $\kappa_{\mathbf{z}}$ is backward Monge and $\tilde{\kappa}_{\mathbf{z}'}$ is Monge. Denote by $T_{\mathbf{z}}$ the inverse of $\kappa_{\mathbf{z}}$ which can be chosen as any Monge transport from $([0, 1], \tau)$ to $(A_{\mathbf{z}}, \mu_{\mathbf{z}})$. More precisely, we have for any $B \subseteq Y$,

$$\int_X \kappa^x(B) \bar{\mu}(dx) = \int_{\mathcal{R}} \kappa^{T_{\mathbf{z}}(u)}(B) \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}) \quad (\text{A.12})$$

In this case, we may compose the kernels to get a kernel

$$\hat{\kappa}(\mathbf{z}, u)(D) := \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \kappa^{T_{\mathbf{z}}(u)}(dy) \tilde{\kappa}_{\mathbf{z}'}^y(D) m_{\boldsymbol{\nu}}(d\mathbf{z}'), \quad (\text{A.13})$$

as illustrated by Figure 4.

To show $\hat{\kappa} \in \mathbb{M}_{b,1}$, it suffices to show that for $Z' \subseteq \mathbb{R}_+^d$ and $V \subseteq [0, 1]$,

$$(i) \int_{\mathcal{R}} \hat{\kappa}(\mathbf{z}, u)(Z' \times V) \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}) = \tau(V) m_{\boldsymbol{\nu}}(Z');$$

$$(ii) \int_{\mathcal{R}} \hat{\kappa}^{(\mathbf{z}, u)}(Z' \times V) \mathbf{z} \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}) = \int_{Z'} \mathbf{z}' \tau(V) m_{\boldsymbol{\nu}}(d\mathbf{z}').$$

To prove (i), we first claim that for $B \subseteq Y$,

$$\nu_{\mathbf{z}'}(B) = \int_X \kappa^x(B \cap B_{\mathbf{z}'}) \bar{\mu}(dx). \quad (\text{A.14})$$

This is a direct consequence of the uniqueness of disintegration and for $B \subseteq Y$,

$$\begin{aligned} \bar{\nu}(B) &= \int_{\mathbb{R}_+^d} \int_{A_{\mathbf{z}}} \kappa^x(B) \mu_{\mathbf{z}}(dx) m_{\boldsymbol{\mu}}(d\mathbf{z}) \\ &= \int_{\mathbb{R}_+^d} \int_{A_{\mathbf{z}}} \int_{\mathbb{R}_+^d} \kappa^x(B \cap B_{\mathbf{z}'}) m_{\boldsymbol{\nu}}(d\mathbf{z}') \mu_{\mathbf{z}}(dx) m_{\boldsymbol{\mu}}(d\mathbf{z}) \\ &= \int_{\mathbb{R}_+^d} \int_X \kappa^x(B \cap B_{\mathbf{z}'}) \bar{\mu}(dx) m_{\boldsymbol{\nu}}(d\mathbf{z}'). \end{aligned} \quad (\text{A.15})$$

It follows that using (A.14) in the second equality, (A.12) in the third, (A.13) in the fourth, that

$$\begin{aligned} &\tau(E_1)Q(E_2) \\ &= \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(E_1 \times E_2) \nu_{\mathbf{z}'}(dy) m_{\boldsymbol{\nu}}(d\mathbf{z}') \\ &= \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(E_1 \times E_2) \int_X \kappa^x(dy \cap B_{\mathbf{z}'}) \bar{\mu}(dx) m_{\boldsymbol{\nu}}(d\mathbf{z}') \\ &= \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(E_1 \times E_2) \int_{\mathcal{R}} \kappa^{T_{\mathbf{z}}(u)}(dy \cap B_{\mathbf{z}'}) \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}) m_{\boldsymbol{\nu}}(d\mathbf{z}') \\ &= \int_{\mathcal{R}} \hat{\kappa}^{(\mathbf{z}, u)}(E_1 \times E_2) \tau(du) m_{\boldsymbol{\mu}}(d\mathbf{z}). \end{aligned}$$

To show (ii), we first note that by definition of $\tilde{\kappa}_{\mathbf{z}'}$, for all \mathbf{z}' and $V \subseteq [0, 1]$,

$$\tau(V) = \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(\mathbf{z}', V) \bar{\nu}(dy).$$

Therefore, since $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$, for $Z' \subseteq \mathbb{R}_+^d$ and $V \subseteq [0, 1]$, we have

$$\begin{aligned} \int_{Z'} \mathbf{z}' \tau(V) m_{\boldsymbol{\nu}}(d\mathbf{z}') &= \int_{Z'} \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(\mathbf{z}', V) \bar{\nu}(dy) \mathbf{z}' m_{\boldsymbol{\nu}}(d\mathbf{z}') \\ &= \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \mathbb{1}_{\{\boldsymbol{\nu}'(y) \in Z'\}} \tilde{\kappa}_{\mathbf{z}'}^y(\mathbf{z}', V) \bar{\nu}(dy) \mathbf{z}' m_{\boldsymbol{\nu}}(d\mathbf{z}') \\ &= \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(Z' \times V) \boldsymbol{\nu}(dy) m_{\boldsymbol{\nu}}(d\mathbf{z}') \\ &= \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(Z' \times V) \int_X \kappa^x(dy) \boldsymbol{\mu}'(x) \bar{\mu}(dx) m_{\boldsymbol{\nu}}(d\mathbf{z}'). \end{aligned}$$

By (A.12) and (A.13), we obtain for an arbitrary $Z' \subseteq \mathbb{R}_+^d$ that

$$\begin{aligned}
& \int_{Z'} \mathbf{z}' \tau(V) m_\nu(d\mathbf{z}') \\
&= \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \tilde{\kappa}_{\mathbf{z}'}^y(Z' \times V) \int_{\mathcal{R}} \kappa^{T_{\mathbf{z}}(u)}(dy) \mathbf{z}' \tau(du) m_\mu(d\mathbf{z}) m_\nu(d\mathbf{z}') \\
&= \int_{\mathcal{R}} \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \kappa^{T_{\mathbf{z}}(u)}(dy) \tilde{\kappa}_{\mathbf{z}}^y(Z' \times V) m_\nu(d\mathbf{z}') \mathbf{z}' \tau(du) m_\mu(d\mathbf{z}) \\
&= \int_{\mathcal{R}} \hat{\kappa}^{(\mathbf{z}, u)}(Z' \times V) \mathbf{z}' \tau(du) m_\mu(d\mathbf{z}).
\end{aligned}$$

This finishes the proof that $\hat{\kappa} \in \mathbb{M}_{b,1}$. Next we show that after composing these kernels we get back κ , i.e., for $\boldsymbol{\mu}'(x) = \mathbf{z}$ and $B \subseteq Y$, that

$$\kappa^x(B) = \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \hat{\kappa}^{(\mathbf{z}, u)}(d\mathbf{z}', du') \mathbb{1}_{\{S_{\mathbf{z}'}(u') \in B\}}. \quad (\text{A.16})$$

Since $T_{\mathbf{z}}$ and $\kappa_{\mathbf{z}}$ forms inverses of each other, we have

$$\begin{aligned}
\kappa^x(B) &= \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \kappa^{T_{\mathbf{z}}(u)}(B) \\
&= \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \kappa^{T_{\mathbf{z}}(u)}(dy) m_\nu(d\mathbf{z}') \mathbb{1}_{\{y \in B\}}.
\end{aligned}$$

Similarly, since $S_{\mathbf{z}'}$ and $\tilde{\kappa}_{\mathbf{z}'}$ are inverses of each other, it holds that

$$\mathbb{1}_{\{y \in B\}} = \tilde{\kappa}_{\mathbf{z}'}^y(\mathbf{z}', (S_{\mathbf{z}'})^{-1}(B)).$$

Therefore, using (A.13) in the last step yields

$$\begin{aligned}
& \kappa^x(B) \\
&= \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \kappa^{T_{\mathbf{z}}(u)}(dy) \tilde{\kappa}_{\mathbf{z}'}^y(\mathbf{z}', (S_{\mathbf{z}'})^{-1}(B)) m_\nu(d\mathbf{z}') \\
&= \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \int_{B_{\mathbf{z}'}} \kappa^{T_{\mathbf{z}}(u)}(dy) \tilde{\kappa}_{\mathbf{z}'}^y(\{\mathbf{z}'\} \times du') \mathbb{1}_{\{S_{\mathbf{z}'}(u') \in B\}} m_\nu(d\mathbf{z}') \\
&= \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \int_{\mathbb{R}_+^d} \int_{B_{\mathbf{z}'}} \mathbb{1}_{\{S_{\mathbf{z}'}(u') \in B\}} \kappa^{T_{\mathbf{z}}(u)}(dy) \tilde{\kappa}_{\mathbf{z}'}^y(d\mathbf{z}', du') m_\nu(d\bar{\mathbf{z}}') \\
&= \int_{\mathcal{R}} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \hat{\kappa}^{(\mathbf{z}, u)}(d\mathbf{z}', du') \mathbb{1}_{\{S_{\mathbf{z}'}(u') \in B\}},
\end{aligned}$$

proving (A.16), hence concluding the proof. \square

Proof of Corollary 3. The existence of a backward martingale Monge coupling follows from Proposition 1 and Theorem 2.1 of Nutz et al. (2022). We let $\hat{\kappa}$ in (19) be induced by the Monge map h in the \mathbb{R}_+^d dimension and identity in the $[0, 1]$ dimension. Thus, there exists $\kappa \in \mathcal{K}(\boldsymbol{\mu}, \boldsymbol{\nu})$ given by (19) such that

$$\kappa^x(B) = \int_{[0,1]} \kappa_{\boldsymbol{\mu}'(x)}^x(\boldsymbol{\mu}'(x), du) \tilde{\kappa}_{h(\boldsymbol{\mu}'(x))}^{(h(\boldsymbol{\mu}'(x)), u)}(B). \quad (\text{A.17})$$

Since $\boldsymbol{\mu}$ is jointly atomless, each $\mu_{\mathbf{z}}$ is atomless, hence we may pick $\kappa_{\mathbf{z}}$ that is Monge for $\mathbf{z} \in \mathbb{R}_+^d$. As a consequence, κ is given by a composition of two Monge maps, hence is Monge. Denote by f the map that induces κ . It is then immediate from (A.17) that $\boldsymbol{\nu}'(f(x)) = h(\boldsymbol{\mu}'(x))$. \square

Proof of Theorem 4. Assume that $m_{\boldsymbol{\mu}} = m_{\boldsymbol{\nu}}$ and $c(x, y)$ is continuous. Then the martingale transport is unique, so that any $\hat{\kappa} \in \mathcal{M}_{b,1}$ must be the identity in the first coordinate. The transport cost (20) then simplifies into

$$\begin{aligned} \mathcal{C}(\kappa) &= \int_{\mathcal{R}} \int_{[0,1]} \hat{\kappa}^{(\mathbf{z},u)}(\{\mathbf{z}\} \times du') \left(\int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}}} c(x, y) \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \tilde{\kappa}_{\mathbf{z}}^{(\mathbf{z},u')}(\mathrm{d}y) \mu_{\mathbf{z}}(\mathrm{d}x) \right) m_{\boldsymbol{\mu}}(\mathrm{d}\mathbf{z}) \\ &= \int_{\mathbb{R}_+^d} \int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}}} c(x, y) \int_{[0,1]} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \hat{\kappa}^{(\mathbf{z},u)}(\{\mathbf{z}\} \times du') \tilde{\kappa}_{\mathbf{z}}^{(\mathbf{z},u')}(\mathrm{d}y) \mu_{\mathbf{z}}(\mathrm{d}x) m_{\boldsymbol{\mu}}(\mathrm{d}\mathbf{z}), \end{aligned}$$

so that

$$\begin{aligned} \inf_{\hat{\kappa} \in \mathcal{M}_{b,1}} \mathcal{C}(\kappa) &= \int_{\mathbb{R}_+^d} \left(\inf_{\hat{\kappa} \in \mathcal{M}_{b,1}} \int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}}} c(x, y) \int_{[0,1]} \int_{[0,1]} \kappa_{\mathbf{z}}^x(\{\mathbf{z}\} \times du) \right. \\ &\quad \left. \hat{\kappa}^{(\mathbf{z},u)}(\{\mathbf{z}\} \times du') \tilde{\kappa}_{\mathbf{z}}^{(\mathbf{z},u')}(\mathrm{d}y) \mu_{\mathbf{z}}(\mathrm{d}x) \right) m_{\boldsymbol{\mu}}(\mathrm{d}\mathbf{z}) \\ &\geq \int_{\mathbb{R}_+^d} \left(\inf_{\kappa \in \mathcal{K}(\mu_{\mathbf{z}}, \nu_{\mathbf{z}})} \int_{A_{\mathbf{z}}} \int_{B_{\mathbf{z}}} c(x, y) \kappa^x(\mathrm{d}y) \mu_{\mathbf{z}}(\mathrm{d}x) \right) m_{\boldsymbol{\mu}}(\mathrm{d}\mathbf{z}) \quad (\text{A.18}) \\ &= \int_{\mathbb{R}_+^d} \mathcal{I}_c(\mu_{\mathbf{z}}, \nu_{\mathbf{z}}) P(\mathrm{d}\mathbf{z}). \end{aligned}$$

We next show the inequality in (A.18) is in fact an equality, so that (i) is equivalent to (iii). Recall from the disintegration theorem that the map

$$\mathbb{R}_+^d \rightarrow \mathcal{P}(X) \times \mathcal{P}(Y), \quad \mathbf{z} \mapsto (\mu_{\mathbf{z}}, \nu_{\mathbf{z}})$$

is measurable. By Corollary 5.22 in Villani (2009) and since c is continuous, 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

$$\pi := \int_{\mathbb{R}_+^d} \pi_{\mathbf{z}} m_{\boldsymbol{\mu}}(\mathrm{d}\mathbf{z}).$$

It is straightforward to check using (15) that $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$. Alternatively, using the kernel formulation, this means there exists a well-defined 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 $\nu_{\mathbf{z}}$. Therefore, (A.18) is an equality, and 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). \square

Proof of Proposition 6. To show the \geq direction, consider any $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and any $(\phi, \psi) \in \tilde{\Phi}(c)$. Recall from Theorem 4 that

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

It then holds that

$$\int_X \phi \mathrm{d}\bar{\boldsymbol{\mu}} + \int_Y \psi \mathrm{d}\bar{\boldsymbol{\nu}} = \int_{X \times Y} \phi(x) + \psi(y) \pi(\mathrm{d}x, \mathrm{d}y) \leq \int_{X \times Y} c \mathrm{d}\pi.$$

This proves the \geq in (24).

Using (22) and the classic duality, it suffices to prove

$$\begin{aligned} & \sup \left\{ \int_X \phi \, d\bar{\mu} + \int_Y \psi \, d\bar{\nu} \mid (\phi, \psi) \in \tilde{\Phi}_c \right\} \\ & \geq \int_{\mathbb{R}_+^d} \sup \left\{ \int_X \phi_{\mathbf{z}} \, d\mu_{\mathbf{z}} + \int_Y \psi_{\mathbf{z}} \, d\nu_{\mathbf{z}} \mid \phi_{\mathbf{z}}(x) + \psi_{\mathbf{z}}(y) \leq c(x, y) \right\} m_{\boldsymbol{\mu}}(d\mathbf{z}). \end{aligned}$$

By Theorem 1.39 of Santambrogio (2015), the suprema on the right-hand side are attained for bounded continuous functions $\phi_{\mathbf{z}}, \psi_{\mathbf{z}}$. By Theorem 18.19 of Aliprantis and Border (2006), there exists a measurable selection $\mathbf{z} \rightarrow (\phi_{\mathbf{z}}, \psi_{\mathbf{z}})$ where each $(\phi_{\mathbf{z}}, \psi_{\mathbf{z}})$ is a maximizer. We define $\phi(x) = \phi_{\boldsymbol{\mu}'(x)}(x)$ and $\psi(y) = \psi_{\boldsymbol{\nu}'(y)}(y)$. Since $\mathbf{z} \mapsto \phi_{\mathbf{z}}(x)$ is measurable and $x \mapsto \phi_{\mathbf{z}}(x)$ is continuous, we have $(\mathbf{z}, x) \mapsto \phi_{\mathbf{z}}(x)$ is jointly measurable, and hence ϕ, ψ are measurable. Moreover, $(\phi, \psi) \in L^1(\bar{\mu}) \times L^1(\bar{\nu})$ since c is bounded. Evidently, $(\phi, \psi) \in \tilde{\Phi}_c$. It also follows from the disintegration theorem that

$$\int_X \phi \, d\bar{\mu} = \int_{\mathbb{R}_+^d} \int_X \phi_{\mathbf{z}} \, d\mu_{\mathbf{z}} m_{\boldsymbol{\mu}}(d\mathbf{z}).$$

This proves the desired inequality and hence (24).

By Theorem 4, the infimum in (24) is attained. Our construction of the maximizers ϕ, ψ above also implies that the supremum is attained. \square

Proof of Proposition 8. 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

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

By proper transformations we may without loss of generality assume that μ_1 and ν_1 are standard Gaussian, which we denote by χ . We then have

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

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)

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

Taking logarithm we obtain

$$(\mathbf{Z} - \mathbf{m})^\top \Sigma^{-1} (\mathbf{Z} - \mathbf{m}) - \mathbf{Z}^\top \mathbf{Z} + \log \det \Sigma \stackrel{\text{law}}{=} (\mathbf{Z} - \mathbf{n})^\top \Omega^{-1} (\mathbf{Z} - \mathbf{n}) - \mathbf{Z}^\top \mathbf{Z} + \log \det \Omega.$$

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

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

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 ,

$$|\det(I - 2t(\Sigma^{-1} - I))| = |\det(I - 2t(\Omega^{-1} - I))|. \quad (\text{A.19})$$

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

$$(\Sigma^{-1}\mathbf{m})^\top(I - 2t(\Sigma^{-1} - I))^{-1}(\Sigma^{-1}\mathbf{m}) = (\Omega^{-1}\mathbf{n})^\top(I - 2t(\Omega^{-1} - I))^{-1}(\Omega^{-1}\mathbf{n}). \quad (\text{A.20})$$

From [\(A.19\)](#), 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 [\(A.20\)](#), 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 $k \geq 2$,

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

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

$$((PE)^{-1}\mathbf{m})^\top(D')^k((PE)^{-1}\mathbf{m}) = (Q^{-1}\mathbf{n})^\top(D')^k(Q^{-1}\mathbf{n}). \quad (\text{A.21})$$

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_ℓ the corresponding eigenspaces with dimensions d_1, \dots, d_ℓ . The system of equations [\(A.21\)](#) 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, [\(A.21\)](#) 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. \square

A natural question to ask is whether [Proposition 8](#) extends to dimensions $d > 2$. In this case, computation of Laplace transforms yields that instead of the relation [\(A.19\)](#) above, we have for all $\mathbf{t} = \{t_j\}_{2 \leq j \leq d}$ 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 for any $2 \leq j \leq d$, $\Sigma_j^{-1} = P\Omega_j^{-1}P^{-1}$. This is related to the simultaneous similarity of matrices problem, which was solved in [Friedland \(1983\)](#) in the complex case. [Friedland \(1983\)](#) proved that given some mild conditions on the characteristic polynomial

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

there are only finitely many orbits of tuples of symmetric matrices (A_1, \dots, A_d) under the action of simultaneous conjugation by an orthogonal matrix. 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 8](#). 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.

A.4 On the Wasserstein distance between vector-valued measures

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 $p \geq 1$.

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

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

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

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

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

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

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

The following consequence of [Theorem 4](#) provides a collection \mathcal{E} of \mathbb{R}_+^d -valued probability measures $\mathcal{E} \subseteq \mathcal{P}(X)_{p, \rho}^d$ such that for any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{E}$, $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{W}_p(\boldsymbol{\nu}, \boldsymbol{\mu}) < \infty$. We recall the equivalence relation \simeq from [Section 5.3](#).

Proposition A.1. *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 of Proposition A.1. By [Theorem 4](#), we have

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

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

The upshot of Proposition A.1 is that, for μ, ν belonging to the same equivalence class we can define the Wasserstein distance

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

By the Decomposition Theorem, for $\mu, \nu \in \mathcal{E}_P$, we have

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

The following corollary then follows from standard results on the analysis on the space of random variables taking values in a Polish space; see Crauel (2002).

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

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

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

we have by taking infimum that

$$\mathcal{W}_p(\mu, \nu)^p \geq \frac{1}{d} \sum_{j=1}^d \mathcal{W}_p(\mu_j, \nu_j)^p. \quad (\text{A.22})$$

It is also straightforward to see that (A.22) is not an equality in Example 14.

In the case $d = 1$, the metric \mathcal{W}_p metrizes the weak topology if the space X is compact. On the other hand, if $d > 1$, in general the convergence of $\mathcal{W}_p(\mu^{(n)}, \mu)$ to zero does not imply the weak convergence of each component of $\mu^{(n)}$ to that of μ , even if $\mu^{(n)}, \mu \in \mathcal{E}_P$ for $n \in \mathbb{N}$ and X is compact.

Example A.1. As a sanity check, let us consider the case where $\mu_1 = \dots = \mu_d$ and $\nu_1 = \dots = \nu_d$. Then according to discussions in Section 3.3, the optimal transport from μ_1 to ν_1 is also optimal from μ to ν . This means

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

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 A.2. Suppose that $\mu \in \mathcal{P}(\mathbb{R}^d)$, define $T(x) = ax + b$ for some $a > 0$, $b \in \mathbb{R}$ and $\nu := \mu \circ T^{-1}$. Consider the convex cost $c(x, y) = |x - y|^p$, $p \geq 1$. Then since the linear transformation is comonotone, the associated kernel $\kappa_T \in \mathcal{K}(\bar{\mu}, \bar{\nu})$ is an optimal transport from $\bar{\mu}$ to $\bar{\nu}$. By arguments in Section 3.3, κ_T is also optimal from μ to ν . In particular, (A.22) is an equality. Moreover, by the arguments in Section 3.3, in case μ_1, \dots, μ_d have disjoint supports, (A.22) is also an equality.

A.5 Dual MOT-SOT parity

Duality for MOT was first established by [Beiglböck et al. \(2013\)](#) in the following form. Given probability measures μ, ν on \mathbb{R} with $\mu \preceq_{\text{cx}} \nu$ and an upper semi-continuous cost function c , it holds

$$\begin{aligned} & \inf_{\pi \in \mathcal{M}(\mu, \nu)} \int c(x, y) \pi(dx, dy) \\ &= \sup \left\{ \int \phi d\mu + \int \psi d\nu \mid \phi(x) + \psi(y) + h(x)(y - x) \leq c(x, y) \right\}, \end{aligned} \quad (\text{A.23})$$

where it is also noted that the supremum may not always be attained; see also [Beiglböck et al. \(2017\)](#). Our goal in this section is to connect the dual problems in (18) and (A.23) when the primal problems are connected via the MOT-SOT parity.

Let us consider two measures P, Q supported on $[0, 2]$ with mean 1 (this extends naturally to compactly supported measures), with $P \succeq_{\text{cx}} Q$. Let $c(x, y)$ be a continuous cost function. We next construct measures μ, ν on $[0, 1]$ so that the corresponding SOT problem is connected to the MOT problem with marginals P, Q . Let F, G be cdfs for P, Q and assume they are continuously invertible.¹³ Let τ be the Lebesgue measure on $[0, 1]$ and define $d\mu_1 = F^{-1}d\tau$, $d\mu_2 = (2 - F^{-1})d\tau$, $d\nu_1 = G^{-1}d\tau$, $d\nu_2 = (2 - G^{-1})d\tau$. In this case, μ' and ν' are injective. By (18) and Example 9, the dual SOT problem solves

$$\begin{aligned} & \sup \left\{ \int \phi(x) dx + \int \psi_1(y) G^{-1}(y) dy + \int \psi_2(y) (2 - G^{-1}(y)) dy \mid \right. \\ & \quad \left. \phi(x) + \psi_1(y) F^{-1}(x) + \psi_2(y) (2 - F^{-1}(x)) \leq c(F^{-1}(x), F^{-1}(y)) \right\} \\ &= \sup \left\{ \int \phi(x) dx + \int \psi_1(y) G^{-1}(y) dy + \int \psi_2(y) (2 - G^{-1}(y)) dy \mid \right. \\ & \quad \left. \phi(F(x)) + \psi_1(G(y)) x + \psi_2(G(y)) (2 - x) \leq c(x, y) \right\} \\ &= \sup \left\{ \int \phi(F^{-1}(x)) dx + \int \psi_1(G^{-1}(y)) G^{-1}(y) dy + \int \psi_2(G^{-1}(y)) (2 - G^{-1}(y)) dy \mid \right. \\ & \quad \left. \phi(x) + (\psi_1(y) - \psi_2(y)) x + 2\psi_2(y) \leq c(x, y) \right\} \\ &= \sup \left\{ \int \phi(x) P(dx) + \int \psi_1(y) y Q(dy) + \int \psi_2(y) (2 - y) Q(dy) \mid \right. \\ & \quad \left. \phi(x) + (\psi_1(y) - \psi_2(y)) x + 2\psi_2(y) \leq c(x, y) \right\} \\ &= \sup \left\{ \int \phi(x) P(dx) + \int \psi(y) Q(dy) \mid \phi(x) + \psi(y) + (y - x) h(y) \leq c(x, y) \right\}, \end{aligned}$$

using change of variables. This is precisely the (pointwise) MOT duality (A.23). The dual MOT-SOT parity can be phrased as follows: if $(\hat{\phi}, \hat{\psi}, h)$ is a dual optimizer for MOT and (ϕ, ψ_1, ψ_2) for SOT, then

$$\hat{\phi}(z) = \phi(F^{-1}(z)) \quad \text{and} \quad \hat{\psi}(z') = \phi_1(G^{-1}(z')) z' - \phi_2(G^{-1}(z')) z' + 2\psi_2(G^{-1}(z')).$$

¹³These regularity conditions on P, Q do not affect the non-attainability of the supremum in (A.23). Indeed, it is the irreducibility of the martingale coupling that matters.

B A small review of optimal transport in higher dimensions

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. The closest to our setting is [Wolansky \(2021\)](#) and [Gover \(2020\)](#) in point (vi) below.

- (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

$$\int_{X_1 \times \dots \times X_d} c(x_1, \dots, x_d) \pi(dx_1, \dots, dx_d)$$

among measures $\pi \in \mathcal{P}(X_1 \times \dots \times X_d)$ with marginals μ_1, \dots, μ_d . A duality formula can be established. However, the existence of a Monge transport is a more delicate problem for dimension $d \geq 3$. 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 $c(x_1, \dots, x_d) = (x_1 + \dots + x_d)^2$ is obtained by [Wang and Wang \(2016\)](#) under some conditions on (μ_1, \dots, μ_d) .

- (ii) More generally, [Rüschendorf \(1991\)](#) considered the multivariate marginal problem. For a collection \mathcal{E} of subsets of $\{1, \dots, n\}$, consider the set of measures on $X_1 \times \dots \times X_n$ that have fixed projections onto each $\prod_{j \in J} X_j$, $J \in \mathcal{E}$. The existence of such measures is a non-trivial task. A duality formula in a more general context was established earlier by [Rüschendorf \(1984\)](#). For more recent results, see [Gladkov et al. \(2019, 2021\)](#) for the special case where \mathcal{E} consists of subsets of cardinality k , $k \leq n$. This problem is also connected to Monge–Kantorovich problem with linear constraints.
- (iii) [Bacon \(2020\)](#) 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 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d)$, 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 1](#)). The existence of a transport is guaranteed and duality is obtained. [Bacon \(2020\)](#) also investigated an extension of the Wasserstein distances.
- (iv) Some earlier studies are in a similar direction as [Bacon \(2020\)](#). To list a few, in [Chen et al. \(2018a,b\)](#) and [Ryu et al. \(2018\)](#), the notion of “vector-valued optimal transport” was proposed. Inspired by the dynamic formulation of classic optimal transport with the L^2 cost, they took the Benamou–Brenier perspective and formulated an optimal transport problem between vector-valued measures using divergences in a network flow problem. Similarly as [Bacon \(2020\)](#), 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 $\int_X \rho(x, x_0) \|\eta\| (dx) < \infty$, where $\|\eta\|$ is the total variation norm of η . This problem deals with

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

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) A recent monograph [Wolansky \(2021\)](#) and PhD thesis [Gover \(2020\)](#) 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, most of their results concern duality formulas, existence of dual optimizers, and the structure (e.g., existence and uniqueness) of the *optimal* multi-partition, where Y is a finite set and under certain atomless assumptions.¹⁴ 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 \(2021\)](#) and [Gover \(2020\)](#) are Proposition 1 and Theorem 2, where our results offer more generality.

C Application to 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.¹⁵ First, we state the relevant version of the duality formula in Theorem 2. 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 $\eta \sim \bar{\boldsymbol{\mu}}$. Assume that X and Y are compact, and $g : X \times Y \rightarrow [-\infty, \infty)$ is upper semi-continuous. The duality formula, with a maximization in place of a minimization in (18), is

$$\sup_{\pi \in \Pi_{\eta}(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{X \times Y} g \, d\pi = \inf_{(\phi, \boldsymbol{\psi}) \in \Phi_g} \int_X \phi \, d\eta + \int_Y \boldsymbol{\psi}^\top \, d\boldsymbol{\nu}, \quad (\text{A.24})$$

where

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

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 $\boldsymbol{\mu}' = d\boldsymbol{\mu}/d\eta$, 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

¹⁴which 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.

¹⁵We refer to [Galichon \(2016\)](#), [Bacon et al. \(2023\)](#), and the references therein for variations of labour market equilibrium models using an optimal transport approach.

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 $\int g \, d\pi$.

Take two arbitrary functions $w : X \rightarrow \mathbb{R}$ and $\mathbf{p} : Y \rightarrow \mathbb{R}^d$. 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

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

which follows from the definition of π .

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

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

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

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

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,

$$\int_{X \times Y} g \, d\pi \geq \int_X w \, d\eta + \int_Y \mathbf{p}^\top d\boldsymbol{\nu}. \quad (\text{A.25})$$

Since (a) implies

$$w(x) + \mathbf{p}(y) \cdot \boldsymbol{\mu}'(x) \geq g(x, y) \quad (\text{A.26})$$

for all $x \in X$ and $y \in Y$, integrating (A.26) with respect to π gives

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

and hence, (A.25) has to hold as an equality, and this implies the duality (A.24). 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.

References

- Adams, D. R. and Hedberg, L. I. (1999). *Function Spaces and Potential Theory* (Vol. 314). Springer Science & Business Media.
- Ahani, N., Andersson, T., Martinello, A., Teytelboym, A., and Trapp, A. C. (2021). Placement optimization in refugee resettlement. *Operations Research*, **69**(5), 1468–1486.

- Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis. A Hitchhiker's Guide*. Third Edition. Springer, Berlin.
- 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. (2020). Multi-species Optimal Transportation. *Journal of Optimization Theory and Applications*, **184**(2), 315–337.
- Bacon, X., Carlier, G., and Nazaret, B. (2023). A spatial Pareto exchange economy problem. *Applied Mathematics & Optimization*, **87**(3), 45.
- 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. *Journal of Political Economy*, to appear.
- 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.
- De March, H. (2018). Entropic approximation for multi-dimensional martingale optimal transport. *arXiv preprint arXiv:1812.11104*.
- De March, H. and Touzi, N. (2019). Irreducible convex paving for decomposition of multidimensional martingale transport plans. *The Annals of Probability*, **47**(3), 1726–1774.

- Delacrétaz, D., Kominers, S. D., and Teytelboym, A. (2016). Refugee resettlement. *University of Oxford Department of Economics Working Paper*.
- Delbaen, F. (2021). Common monotonicity and time-consistency for Lebesgue-continuous monetary utility functions. *Finance and Stochastics*, **25**, 597–614.
- Dellacherie, C. and Meyer, P. A. (2011). *Probabilities and potential, B: Theory of Martingales*. Elsevier.
- 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.
- Gover, S. (2020). *Duality Theorems and Vector Measures in Optimal Transportation Theory* (Doctoral dissertation). Technion—Israel Institute of Technology.
- Guo, G. and Oblój, J. (2019). Computational methods for martingale optimal transport problems. *The Annals of Applied Probability*, **29**(6), 3311–3347.
- Hansen, L. P. and Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review*. **91**(2), 60–66.
- Henry-Labordère, P. (2017). *Model-free Hedging: A Martingale Optimal Transport Viewpoint*. Chapman and Hall/CRC.
- Hirsch, F., Profeta, C., Roynette, B., and Yor, M. (2011). *Peacocks and Associated Martingales, with Explicit Constructions*. Springer Science & Business Media.
- Joe, H. (2014). *Dependence Modeling with Copulas*. London: Chapman & Hall.
- Mathai, A. M. and Provost, S. B. (1992). *Quadratic Forms in Random Variables: Theory and Applications*. Dekker.

- 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 Stebegg, F. (2018). Canonical supermartingale couplings. *The Annals of Probability*, **46**(6), 3351–3398.
- Nutz, M. and Wang, R. (2022). The directional optimal transport. *Annals of Applied Probability*, **32**(2), 1400–1420.
- Nutz, M., Wang, R., and Zhang, Z. (2022). Martingale Transports and Monge Maps. *To appear in Annals of Applied Probability*.
- 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 Poincaré (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. (1984). On the minimum discrimination theorem. *Statistics and Decisions*, **1**, 263–283.
- Rüschendorf, L. (1991). Bounds for distributions with multivariate marginals. *Lecture Notes-Monograph Series Vol. 19, Stochastic Orders and Decision under Risk*, 285–310.
- 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.
- Strassen, V. (1965). The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, **36**(2), 423–439.

- 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. (2021). *Optimal Transport: A Semi-Discrete Approach* (Vol. 37). Walter de Gruyter GmbH & Co KG.