Transport type metrics on the space of probability measures involving singular base measures*

Luca Nenna†   Brendan Pass‡

September 15, 2022

Abstract

We develop the theory of a metric, which we call the ν-based Wasserstein metric and denote by W_ν, on the set of probability measures $\mathcal{P}(X)$ on a domain $X \subseteq \mathbb{R}^m$. This metric is based on a slight refinement of the notion of generalized geodesics with respect to a base measure ν and is relevant in particular for the case when ν is singular with respect to m-dimensional Lebesgue measure; it is also closely related to the concept of linearized optimal transport. The ν-based Wasserstein metric is defined in terms of an iterated variational problem involving optimal transport to ν; we also characterize it in terms of integrations of classical Wasserstein metric between the conditional probabilities when measures are disintegrated with respect to optimal transport to ν, and through limits of certain multi-marginal optimal transport problems. We also introduce a class of metrics which are dual in a certain sense to W_ν, defined relative to a fixed based measure μ, on the set of measures which are absolutely continuous with respect to a second fixed based measure σ.

As we vary the base measure ν, the ν-based Wasserstein metric interpolates between the usual quadratic Wasserstein metric (obtained when ν is a Dirac mass) and a metric associated with the uniquely defined generalized geodesics obtained when ν is sufficiently regular (eg, absolutely continuous with respect to Lebesgue). When ν concentrates on a lower dimensional submanifold of $\mathbb{R}^m$, we prove that the variational problem in the definition of the ν-based Wasserstein metric has a unique solution. We also establish geodesic convexity of the usual class of functionals, and

of the set of source measures μ such that optimal transport between μ and ν satisfies a strengthening of the generalized nestedness condition introduced in [16]. We finally introduce a slight variant of the dual metric mentioned above in order to prove convergence of an iterative scheme to solve a variational problem arising in game theory.

Contents

1 Introduction

Given two probability measures μ_0 and μ_1 on a convex, bounded domain $X \subseteq \mathbb{R}^m$, the Wasserstein distance between them is defined as the infimal value in the Monge-Kantorovich optimal transport problem; that is,

$$W_2(\mu_0, \mu_1) := \sqrt{\inf_{\gamma \in \Pi(\mu_0,\mu_1)} \int_{X \times X} |x_0 - x_1|^2 d\pi(x_0, x_1)} \tag{1}$$

where the infimum is over the set $\Pi(\mu_0, \mu_1)$ of joint measures γ on $X \times X$ whose marginals are μ_0 and μ_1.

Among the many important properties of the Wasserstein distance (reviewed in [19] [20] and [21] for example) is the fact that it is a metric on the set $\mathcal{P}(X)$ of probability measures on X. In turn, the geodesics induced by this metric, known as *displacement interpolants* and introduced in [14],

play key roles in many problems, both theoretical and applied. Variants of displacement interpolants, known as *generalized geodesics* (introduced in [1], see Definition 2 below), are a natural and important tool in the analysis of problems involving a fixed base measure $\nu \in \mathcal{P}(X)$. For example, a variety of problems in game theory, economics and urban planning consist in minimizing functionals on $\mathcal{P}(X)$ involving optimal transport to a fixed ν. In particular, we mention the optimal transport based formulation of Cournot-Nash equilibria in game theory of Blanchet-Carlier, in which ν parameterizes a population of players and one searches for their equilibrium distribution of strategies, parameterized by $\mu \in \mathcal{P}(X)$ [3] [2] [4]. When ν is absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^m$ on X, these interpolants are uniquely defined, and are in fact geodesics for a metric under which $\mathcal{P}(X)$ is isomorphic to a subset of the Hilbert space $L^2(X, \mathbb{R}^n)$; that is, under this isomorphism, generalized geodesics are mapped to line segments.

On the other hand, it is natural in certain problems to consider *singular* base measures[1], and our goal here is to initiate the development of a framework to study these. We study a metric on an appropriate subset of $\mathcal{P}(X)$ for each choice of base measure ν, which we call the *ν-based Wasserstein metric* (see Definition 3); essentially, this metric arises from minimizing the average squared distance among all couplings of μ_0 and μ_1 corresponding to generalized geodesics. It is also very closely related to the concept of linear optimal transport, introduced in [22][2]. Linear optimal transport was mainly used in [22] for image comparison problems, and the base measure ν was often taken to be discrete. In that work, and, to the best of our knowledge, in the considerable literature on linear optimal transport following from it, the theoretical properties of the metric, especially for base measures supported on lower dimensional submanifolds, to which we pay special attention below, have largely been left undeveloped.

Geodesics with respect to the ν-based Wasserstein metric will always be generalized geodesics with respect to ν, but, for different structures of ν, very different geometry is induced. In particular, when ν is absolutely continuous, we obtain the Hilbert space geometry discussed above, whereas when ν is a Dirac mass, we obtain the Wasserstein metric, independently of the y parameterizing $\nu = \delta_y$ within this class. We pay special attention to the cases in between, in particular when ν concentrates on a lower dimensional submanifold of $\mathbb{R}^m$, in which case the problem has a natural interplay with the unequal

---

[1] For instance, in game theory problems derived from spatial economics, ν may represent a population of players, parametrized by their location $y \in \mathbb{R}^2$ [4]; it is often the case that the population is essentially concentrated along a one dimensional subset, such as a major highway or railroad.

[2] In fact, in the case we are most interested in, when there exists a unique optimal transport between ν and each of the measures to be compared, the ν-based Wasserstein metric coincides with the metric derived from solving the linear optimal transport problem, denoted by $d_{LOT,\nu}$ in [22]. They differ slightly for more general measures; in this case, the ν-based Wasserstein metric is in fact only a semi-metric.

3

dimensional optimal transport problem explored in [8], [9], [16] and [17]. In the particular case when ν concentrates on a line segment, we show that our metric coincides with the layerwise-Wasserstein metric (see (10)) introduced in [12] to analyze anisotropic data such as plants' root shapes.

We establish three equivalent characterizations of the ν-based Wasserstein metric, roughly speaking:

1. as an optimal transport problem restricted to couplings which are correlated along ν (we take this as the definition);

2. by optimally coupling conditional probabilities of μ_0 and μ_1 after disintegrating with respect to optimal transport to ν;

3. as limits of multi-marginal optimal transport between μ_0, μ_1 and ν.

In many cases of interest, we establish uniqueness of the corresponding geodesics (although uniqueness of generalized geodesics for regular base measures was established in [1], they are generally not unique when the base measure is singular). We also study geodesic convexity of several functionals which play a key role in optimal transport research; many of these functionals were originally introduced by McCann [14], who established their displacement convexity. Using the standard terminology (see Section 5 for the definition of these terms) convexity along W_ν geodesics of potential energies, interaction energies and the Wasserstein distance to ν follow immediately from known results (namely, their convexity along any generalized geodesic, established in [1]), whereas convexity along W_ν geodesics of the internal energy, under certain conditions, requires a new proof (convexity of the internal energy is known to hold along some, but not all, generalized geodesics [1]). We note that this applies in particular to the layerwise-Wasserstein distance, yielding a far reaching improvement to Corollary 4.2 in [12]. We also show that when ν concentrates on a lower dimensional submanifold, the set of measures μ for which the model $(|x-y|^2, \mu, \nu)$ satisfies a strengthening of the generalized nested condition (see Definition 21) is geodesically convex (we recall that nestedness, introduced in [9] and its higher dimensional generalization from [16], are important properties of unequal dimensional optimal transport problems; when present, they greatly simplify analysis of these problems).

We also introduce a class of metrics which is in a certain sense dual to the ν-based Wasserstein metric, relevant in the case when the measure μ on $X \subset \mathbb{R}^m$ is fixed, and one would like to interpolate between measures on a fixed, lower dimensional submanifold Y (see the precise definition (13)). This is often the case in a variety of applications. The seemingly natural choice, generalized geodesics with base μ, which essentially interpolates between Kantorovich potentials, namely the optimal dual variable of the optimal transport problem, on the X side, does not generally result in interpolants supported on Y. Here, we instead compare and interpolate between Kantorovich potentials on the Y side in order to compare and interpolate between measures on

4

Y. This is in a certain sense complementary, or dual, to our original metric on $\mathcal{P}(X)$ (which involves comparing potentials on Y, in order to compare measures on X – with an additional embedded optimization problem, since potentials on Y do not uniquely determine measures supported on X). We, finally, show how the ideas introduced here can be applied to prove convergence of computational methods to find equilibria in some game theoretic models. In particular we identify conditions under which equilibria are fixed points of a contractive mapping, implying uniqueness of the equilibrium (although this is easily deduced by other methods as well) and, perhaps more importantly, that it can be computed by iterating the mapping. This iteration had already been introduced as a method of computation by Blanchet-Carlier when $m = 1$, but without a proof of convergence, and in higher dimensions, with a proof of convergence but for simpler interaction terms [3]. Here, we prove that the relevant mapping is a contraction with respect to a variant of the dual metric described above[3].

The manuscript is organized as follows. In the next section we introduce the ν-based Wasserstein metric, establish several basic properties of it, and also introduce our class of dual metrics. In Section 3, we recall relevant facts about unequal dimensional optimal transport, and prove a new lemma on the structure of optimal plans which will be crucial in subsequent sections. In the fourth section, we identify conditions under which the variational problem arising in the definition of the ν-based Wasserstein metric has a unique solution, and establish a result on the structure of geodesics for the ν-based Wasserstein metric. We use this structure in Section 5 to establish geodesic convexity results. In the sixth section, we identify conditions under which certain game theoretic equilibria can be characterized by fixed points of a contractive mapping with respect to an appropriate metric.

2   Definition and basic properties

Throughout this paper, for a given set S, a semi-metric will be a function $d : S \times S \to [0, \infty)$ satisfying:

- $d(x, y) = 0$ if and only if $x = y$ (separation).

- $d(x, y) = d(y, x)$ for all $x, y \in S$ (symmetry).

  A semi-metric d will be called a metric if it also satisfies:

- $d(x, y) \leq d(x, z) + d(y, z)$ for all $x, y, z \in S$ (triangle inequality).

We will often refer to $d(x, y)$ as the distance between x and y.

---

[3]In fact, the actual metric used in the proof differs slightly from the dual metric in general, although they coincide under certain conditions.

2.1 Background on optimal transport

In what follows, $\nu \in \mathcal{P}(X)$ will be a fixed reference measure on a convex, bounded[4] domain $X \subseteq \mathbb{R}^m$. Given measures $\mu_0, \mu_1 \in \mathcal{P}(X)$, we denote by $\Pi(\nu, \mu_i)$ the set of probability measures on $X \times X$ whose marginals are ν and μ_i, for $i = 0, 1$ and by $\Pi(\nu, \mu_0, \mu_1)$ the set of probability measures on $X \times X \times X$ with ν, μ_0 and μ_1 as marginals. For a measure $\gamma \in \mathcal{P}(X \times X \times X)$, we will denote its first, second and third marginals by γ_y, γ_{x_0} and γ_{x_1}, respectively. Similarly, we will denote by $\gamma_{yx_0}, \gamma_{yx_1}$ and $\gamma_{x_0x_1}$ its projections onto the appropriate product $X \times X$; for example, γ_{yx_0} is the push forward of γ under the mapping $(y, x_0, x_1) \mapsto (y, x_0)$.

The optimal transport problem between μ_i and ν, for a general continuous cost function $c : \overline{X} \times \overline{X} \to \mathbb{R}$ is to minimize over $\pi \in \Pi(\mu_i, \nu)$

$$\int_{X \times X} c(x_i, y) d\pi(x_i, y). \tag{2}$$

We will most often be interested in the quadratic cost function, $c(x_i, y) = |x_i - y|^2$, where $|\cdot|$ denotes here the standard euclidean norm. Notice that in the quadratic case, (2) is the problem arising in (1). In addition, (2) has a well known dual problem; the minimum in (2) is equal to the maximum of

$$\int_X u(x) d\mu_i(x) + \int_X v(y) d\nu(y) \tag{3}$$

among pairs of functions $(u, v) \in L^1(\mu_i) \times L^1(\nu)$ with $u(x) + v(y) \leq c(x, y)$. It is well known that maximizers to (3) exist and they may be taken to be c-conjugate; that is,

$$u(x) = v^c(x) := \min_{y \in X}[c(x, y) - v(y)], \quad v(y) = u^c(y) := \min_{x \in X}[c(x, y) - u(x)].$$

We let $\Pi_{opt}(\nu, \mu_i)$ be the set of optimal couplings between ν and μ_i with respect to optimal transport for the quadratic cost function; that is:

$$\Pi_{opt}(\nu, \mu_i) := \operatorname{argmin}_{\pi \in \Pi(\nu, \mu_i)} \int_{X \times X} |x_i - y|^2 d\pi(x_i, y). \tag{4}$$

Definition 1. *A* Wasserstein geodesic *between μ_0 and μ_1 (also known as a* displacement interpolant*, using the terminology originally introduced by Mc-Cann [14]) is a curve μ_t in $\mathcal{P}(X)$ of the form $\mu_t = ((1 - t)e_0 + te_1)_{\#}\pi$ for some $\pi \in \Pi_{opt}(\mu_0, \mu_1)$, where for $i = 0, 1$, $e_i : X \times X \to X$ is defined by $e_i(x_0, x_1) = x_i$.*

---

[4]We have chosen to work on a bounded set X here mostly out of technical convenience and to keep the presentation simple. We expect that most results can be extended to unbounded domains under appropriate hypotheses (for instance, decay conditions on the measures).

Recall that, in general, a curve z_t parametrized by $t \in [0,1]$ in a metric space (Z,d) is a *minimizing geodesic* if $d(z_s, z_t) = |s - t|d(z_0, z_1)$ for all $s, t \in [0,1]$. It is well known that the Wasserstein geodesic defined above is in fact a geodesic in this sense for the Wasserstein metric.

We now give the definition of generalized geodesics from [1].

Definition 2. *A generalized geodesic with base measure ν from μ_0 to μ_1 is a curve μ_t in $\mathcal{P}(X)$ of the form $\mu_t = ((1-t)e_0 + te_1)_{\#}\gamma$ for some $\gamma \in \mathcal{P}(X \times X \times X)$ where $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$ for $i = 0, 1$.*

We note that we are especially interested here in the case where ν is singular with respect to Lebesgue measure. Even in this case, $\Pi_{opt}(\nu, \mu_i)$ will very often consist of a single probability measure; in fact, by Brenier's theorem this is the case as soon as μ_i is absolutely continuous with respect to Lebesgue measure [5]. It will turn out that our definition below is not a metric on all of $\mathcal{P}(X)$, but is when restricted to the set $\mathcal{P}_\nu^u(X)$ of probability measures on X for which the solution to the optimal transport problem to ν is unique; that is, the set such that $\Pi_{opt}(\nu, \mu)$ is a singleton.

Before closing this section, we recall the well known notion of disintegration of measures. Given $\pi_i \in \Pi(\nu, \mu_i)$, we disintegrate with respect to ν. There exists a measurable collection μ_i^y of probability measures on X, indexed by y, such that
$$\pi_i(x, y) = \nu(y) \otimes \mu_i^y(x);$$
that is, for every bounded continuous function g on $X \times X$, we have
$$\int_{X \times X} g(x,y)d\pi_i(x,y) = \int_X \left[\int_X g(x,y)d\mu_i^y(x) \right] d\nu(y).$$

We will sometimes call μ_i^y the *conditional probability* of π_i given y.

2.2 The ν-based Wasserstein metric

We now define our metric with base point ν as follows.

Definition 3. *Let $\nu \in \mathcal{P}(X)$. For $\mu_0, \mu_1 \in \mathcal{P}(X)$, we define the ν-based Wasserstein metric as*
$$W_\nu(\mu_1, \mu_0) := \sqrt{\inf_{\gamma \in \Gamma} \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma(y, x_0, x_1)}, \tag{5}$$

where $\Gamma := \{\gamma \in \mathcal{P}(X \times X \times X) \mid \gamma_{yx_i} \in \Pi_{opt}(\mu_i, \nu), i = 0, 1\}$.

Remark 4. *As we will see below, the name ν-based Wasserstein metric is a slight abuse of terminology, since W_ν is only a semi-metric on $\mathcal{P}(X)$ in general, although it is a metric on an appropriate subset.*

Note that the glueing lemma (see [19][Lemma 5.5], for example) implies the existence of a $\gamma \in \Pi(\nu, \mu_0, \mu_1)$ such that $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$ for $i = 0, 1$; therefore, W_ν is well defined. Standard arguments imply the existence of a minimizing γ in (5).

Remark 5. *This definition is closely related to the concept of linear optimal transport, introduced in [22]. The difference is that in linear optimal transport, a* fixed *optimal transport $\pi_{\mu_i} \in \Pi_{opt}(\nu, \mu_i)$ is selected for each μ_i (see equation (3) in [22]), whereas in the definition of W_ν one minimizes over the entire set $\Pi_{opt}(\nu, \mu_i)$. For $\mu_i \in \mathcal{P}_\nu^u(X)$, the two concepts clearly coincide, and it is on this set that W_ν yields a metric (see Lemma 6 below). Outside of this set, W_ν still yields a semi-metric, whereas linear optimal transport might be better described as defining a metric on the selected $\pi_{\mu_i} \in \Pi_{opt}(\nu, \mu_i)$, since it is dependent on these choices; this idea will be developed briefly in Remark 8 below, but first let us rigorously establish that W_ν is indeed a semi-metric.*

The proof of the following Lemma is very similar to the proof that the classical Wasserstein metric is in fact a metric (see for example, [19][Proposition 5.1]). Recall that a semi-metric satisfies the symmetry and identity of indiscernibles axioms in the definition of a metric, but does not satisfy the triangle inequality.

Lemma 6. W_ν *is a semi-metric on $\mathcal{P}(X)$. It is a metric on $\mathcal{P}_\nu^u(X)$.*

Proof. It is immediate that $W_\nu(\mu_0, \mu_1) \geq 0$, with equality if and only if $\mu_0 = \mu_1$, and that $W_\nu(\mu_0, \mu_1) = W_\nu(\mu_1, \mu_0)$. Therefore, W_ν is a semi-metric.

It remains to show that W_ν is a metric on $\mathcal{P}_\nu^u(X)$; we must only verify the triangle inequality. Let μ_0, μ_1 and μ_2 belong to $\mathcal{P}_\nu^u(X)$. Let $\gamma_1 \in \Pi(\nu, \mu_0, \mu_1)$ and $\gamma_2 \in \Pi(\nu, \mu_0, \mu_2)$ be optimal couplings in (5); that is $W_\nu(\mu_0, \mu_1) = \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma_1(y, x_0, x_1)$ and $W_\nu(\mu_0, \mu_2) = \int_{X \times X \times X} |x_0 - x_2|^2 d\gamma_2(y, x_0, x_2)$. Now note that both $(\gamma_1)_{yx_0}$ and $(\gamma_2)_{yx_0}$ are both optimal transports between ν and μ_0; by the uniqueness assumption, we therefore have $(\gamma_1)_{yx_0} = (\gamma_2)_{yx_0}$. The glueing lemma (the version in [20], Lemma 7.6, is sufficiently general) then implies the existence of a measure $\gamma \in \mathcal{P}(X \times X \times X \times X)$ such that $\gamma_{yx_0x_1} = \gamma_1$ and $\gamma_{yx_0x_2} = \gamma_2$. We note that $\gamma_{yx_1x_2}$ satisfies $(\gamma_{yx_1x_2})_{yx_1} = \gamma_{yx_1} = (\gamma_1)_{yx_1} \in$

$\Pi_{opt}(\nu, \mu_1)$ and $(\gamma_{yx_1x_2})_{yx_2} = \gamma_{yx_2} = (\gamma_2)_{yx_2} \in \Pi_{opt}(\nu, \mu_2)$. Therefore, we have

$$
\begin{aligned}
W_\nu(\mu_1, \mu_2) &\leq \sqrt{\int_{X \times X \times X} |x_1 - x_2|^2 d\gamma_{yx_1x_2}(y, x_1, x_2)} \\
&= \sqrt{\int_{X \times X \times X \times X} |x_1 - x_2|^2 d\gamma(y, x_0, x_1, x_2)} \\
&= \|x_1 - x_2\|_{L^2(\gamma)} \\
&\leq \|x_1 - x_0\|_{L^2(\gamma)} + \|x_2 - x_0\|_{L^2(\gamma)} \\
&= \sqrt{\int_{X \times X \times X \times X} |x_1 - x_0|^2 d\gamma(y, x_0, x_1, x_2)} + \sqrt{\int_{X \times X \times X \times X} |x_0 - x_2|^2 d\gamma(y, x_0, x_1, x_2)} \\
&= \sqrt{\int_{X \times X \times X} |x_1 - x_0|^2 d\gamma_1(y, x_0, x_1)} + \sqrt{\int_{X \times X \times X} |x_0 - x_2|^2 d\gamma_2(y, x_0, x_2)} \\
&= W_\nu(\mu_0, \mu_1) + W_\nu(\mu_0, \mu_2)
\end{aligned}
$$

$\square$

The following example confirms that the triangle inequality can fail if we do not restrict to $\mathcal{P}_\nu^u(X)$.

Example 7. *(Failure of the triangle inequality outside of $\mathcal{P}_\nu^u(X)$) Let $X \subseteq \mathbb{R}^2$, and take $\nu = \frac{1}{2}[\delta_{(1,0)} + \delta_{(-1,0)}]$, $\mu_0 = \frac{1}{2}[\delta_{(0,1)} + \delta_{(0,-1)}]$, $\mu_1 = \frac{1}{2}[\delta_{(\varepsilon,1)} + \delta_{(-\varepsilon,-1)}]$, $\mu_2 = \frac{1}{2}[\delta_{(-\varepsilon,1)} + \delta_{(\varepsilon,-1)}]$, for some $\varepsilon > 0$. Then any measure $\pi \in \Pi(\nu, \mu_0)$ is optimal between ν and μ_0. The only optimal transport plan between ν and μ_1, on the other hand, maps $(1,0)$ to $(\varepsilon, 1)$ and $(-1,0)$ to $(-\varepsilon, -1)$. Similarly, the only optimal transport plan between ν and μ_2, maps $(1,0)$ to $(\varepsilon, -1)$ and $(-1,0)$ to $(-\varepsilon, 1)$. We therefore compute that*

$$W_\nu(\mu_0, \mu_1) = \varepsilon = W_\nu(\mu_0, \mu_2)$$

but

$$W_\nu(\mu_1, \mu_2) = 2.$$

Therefore, $W_\nu(\mu_1, \mu_2) > W_\nu(\mu_0, \mu_1) + W_\nu(\mu_0, \mu_2)$ for $\varepsilon < 1$, and so the triangle inequality fails.

Remark 8. *As an alternative to our definition one may equivalently define first a metric on optimal couplings and then optimize over all such couplings. More precisely, let $\Pi_{opt}(\nu) = \cup_{\mu \in \mathcal{P}(X)} \Pi_{opt}(\nu, \mu)$ be the subset of optimal couplings in $\mathcal{P}(X \times X)$ whose first marginal is ν. It is then straightforward to show that*

$$\tilde{W}_\nu(\pi_0, \pi_1) = \sqrt{\inf_{\gamma \in \mathcal{P}(X \times X \times X)|\gamma_{yx_i} = \pi_i, i=0,1} \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma(y, x_0, x_1)} \quad (6)$$

is a metric on $\Pi_{opt}(\nu)$, which we will refer to as the ν-based Wasserstein metric on couplings. Moreover, the ν-based Wasserstein metric then has the following characterization:

$$W_\nu(\mu_0, \mu_1) = \inf_{\pi_i \in \Pi_{opt}(\nu, \mu_i), i=0,1} \tilde{W}_\nu(\pi_0, \pi_1).$$

In addition, we note that the linearized optimal transport framework defined in [22] is essentially metric (6) between couplings.

We pause now to describe the ν-based Wasserstein metric for several simple examples of choices for ν.

Example 9. *We recall the main example used to introduce generalized geodesics in Section 9.2 of [1]. Let ν be absolutely continuous with respect to m-dimensional Lebesgue measure on X. Then by Brenier's theorem [5] there exist unique optimal couplings of the form $(Id, \tilde{T}_i)_{\#}\nu$ in $\Pi_{opt}(\nu, \mu_i)$, and therefore the only measure γ with $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$ for $i = 0, 1$ is $\gamma = (Id, \tilde{T}_0, \tilde{T}_1)_{\#}\nu$. We then have*

$$W_\nu(\mu_0, \mu_1) = \int_X |\tilde{T}_0(y) - \tilde{T}_1(y)|^2 d\nu(y)$$

so that the metric space $(\mathcal{P}(X), W_\nu)$ is isometric to a subset of the Hilbert space $L^2(\nu)$. Geodesics for this metric take the form $t \mapsto \left(tT_1 + (1-t)T_0 \right)_{\#}\nu$; these are the standard generalized geodesics found in, for example, Definition 7.31 of [19].

Example 10. *At the other extreme, suppose $\nu = \delta_y$ is a Dirac mass. Then for any coupling $\pi \in \Pi(\mu_0, \mu_1)$, the measure $\gamma = \delta_y \otimes \pi$ has $\gamma_{yx_i} = \delta \otimes \mu_i \in \Pi_{opt}(\nu, \mu_i)$. Since*

$$\int_{X \times X \times X} |x_0 - x_1|^2 d\gamma(y, x_0, x_1) = \int_{X \times X} |x_0 - x_1|^2 d\pi(x_0, x_1)$$

we have

$$W_\nu^2(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{X \times X} |x_0 - x_1|^2 d\pi(x_0, x_1)$$

which is exactly the standard quadratic Wasserstein metric.

In this paper, we will be especially interested in the cases in between these extremes, when ν is singular with respect to Lebesgue measure but not a Dirac mass. One of the simplest such cases is the following example.

Example 11. *Suppose that ν concentrates on a line segment and is absolutely continuous with respect to one dimensional Hausdorff measure. It then turns out that the ν-based Wasserstein metric coincides with the layerwise-Wasserstein metric introduced in [12]. The proof of this fact is slightly more involved than the previous two examples, and is included as a separate proposition below (Proposition 16 – note that the definition of the layerwise-Wasserstein metric is recalled in equation (10) below as well).*

As we show below, W_ν is also related to the following multi-marginal optimal transport problem. Fix $\varepsilon > 0$ and set

$$\text{MM}_\nu^\varepsilon(\mu_0, \mu_1) := \inf_{\gamma \in \Pi(\nu, \mu_0, \mu_1)} \int_{X \times X \times X} [\varepsilon|x_0 - x_1|^2 + |x_0 - y|^2 + |x_1 - y|^2] d\gamma(y, x_0, x_1).$$
(7)

The following result establishes two different characterizations of the ν-based Wasserstein metric.

Theorem 12. *The following holds*

1. $W_\nu^2(\mu_1, \mu_0) = \inf_{\pi_i \in \Pi_{opt}(\mu_i, \nu), i=0,1} \int_X W_2^2(\mu_0^y, \mu_1^y) d\nu(y)$, *where μ_i^y is the conditional probability given y of the optimal coupling $\pi_i = \nu(y) \otimes \mu_i^y(x) \in \Pi_{opt}(\nu, \mu_i)$ between ν and μ_i.*

2. *Furthermore, any weak limit point $\overline{\gamma}$ as $\varepsilon \to 0$ of minimizers γ_ε of the multi-marginal problem (7) is an optimal coupling between μ_0 and μ_1 for the problem (5) defining W_ν.*

Proof. The first part is almost immediate: for fixed $\pi_i \in \Pi_{opt}(\mu_i, \nu)$, if $\gamma \in \mathcal{P}(X^3)$ is disintegrated with respect to ν, $\gamma(y, x_0, x_1) = \gamma^y(x_0, x_1) \otimes \nu(y)$, then $\gamma_{x_i y} = \pi_i$ is equivalent to the conditional probability γ^y on $X \times X$ having the conditional probabilities μ_i^y of the $\pi_i(x_i, y) = \mu_i^y(x_i) \otimes \nu(y)$ as marginals for almost every fixed y. We can therefore rewrite the integral in (5) as $\int_Y \int_{X \times X} |x_0 - x_1|^2 d\gamma^y(x_0, x_1) d\nu(y)$. Lemma 12.4.7 in [1] then implies that we can choose a Borel family $\gamma^y \in \mathcal{P}(X \times X)$, where for ν almost every y, γ^y is an optimal coupling between the μ_i^y. The choice $\gamma(y, x_0, x_1) = \gamma^y(x_0, x_1) \otimes \nu(y)$ then minimizes the integral in (5), which yields the desired formula.

Turning to the second point, let γ be any competitor in the definition of W_ν; that is, assume $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$ for $i = 0, 1$. Since this clearly implies $\gamma \in \Pi(\nu, \mu_0, \mu_1)$, optimality of γ_ε in the multi-marginal problem yields

$$\int_{X \times X \times X} [\varepsilon|x_0 - x_1|^2 + |x_0 - y|^2 + |x_1 - y|^2] d\gamma_\varepsilon(y, x_0, x_1)$$
$$\leq \int_{X \times X \times X} [\varepsilon|x_0 - x_1|^2 + |x_0 - y|^2 + |x_1 - y|^2] d\gamma(y, x_0, x_1).$$
(8)

Taking the limit as $\varepsilon \to 0$ gives

$$\int_{X \times X \times X} [|x_0 - y|^2 + |x_1 - y|^2] d\overline{\gamma}(y, x_0, x_1) \leq \int_{X \times X \times X} [|x_0 - y|^2 + |x_1 - y|^2] d\gamma(y, x_0, x_1),$$

or

$$\int_{X \times X} |x_0 - y|^2 d\overline{\gamma}_{yx_0}(y, x_0) + \int_{X \times X} |x_1 - y|^2 d\overline{\gamma}_{yx_1}(y, x_1)$$
$$\leq \int_{X \times X} |x_0 - y|^2] d\gamma_{yx_0}(y, x_0) + \int_{X \times X} |x_1 - y|^2 d\gamma_{yx_1}(y, x_1)$$

which immediately implies the optimality of the two-fold marginals of $\overline{\gamma}$ in (4), $\overline{\gamma}_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$ for $= 0, 1$.

Furthermore, the optimality of the two-fold marginals of γ, $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$ means that $\int_{X \times X \times X} |x_i - y|^2 d\gamma(y, x_0, x_1) \leq \int_{X \times X \times X} |x_i - y|^2 d\gamma_\varepsilon(y, x_0, x_1)$; combined with (8), this implies that we must have, for all ε,

$$\int_{X \times X \times X} |x_0 - x_1|^2 d\gamma_\varepsilon(y, x_0, x_1) \leq \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma(y, x_0, x_1).$$

Passing to the limit gives

$$\int_{X \times X \times X} |x_0 - x_1|^2 d\overline{\gamma}(y, x_0, x_1) \leq \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma(y, x_0, x_1).$$

Since this holds for every γ with $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$, it implies the desired result.

$\square$

Adapting the techniques in the proof of the second part of the preceding theorem, we can also establish the following Γ-convergence result relating MM_ν^ε and W_ν.

Proposition 13. *The functional on $\mathcal{P}(X) \times \mathcal{P}(X)$ defined by*

$$(\mu_0, \mu_1) \mapsto F_\nu^\varepsilon(\mu_0, \mu_1) := \frac{1}{\varepsilon} \operatorname{MM}_\nu^\varepsilon(\mu_0, \mu_1) - \frac{1}{\varepsilon} W_2^2(\nu, \mu_0) - \frac{1}{\varepsilon} W_2^2(\nu, \mu_0)$$

Γ-converges to $(\mu_0, \mu_1) \mapsto W_\nu^2(\mu_0, \mu_1)$ as $\varepsilon \to 0$ with respect to the product on $\mathcal{P}(X) \times \mathcal{P}(X)$ of the weak topology on $\mathcal{P}(X)$ with itself

Proof. Fix μ_0 and μ_1 in $\mathcal{P}(X)$; to establish Γ-convergence, we must:

1. Show that whenever $\mu_0^\varepsilon, \mu_1^\varepsilon$ converge weakly to μ_0, μ_1 as $\varepsilon \to 0$ we have

$$W_\nu^2(\mu_0, \mu_1) \leq \liminf_{\varepsilon \to 0} F_\nu^\varepsilon(\mu_0^\varepsilon, \mu_1^\varepsilon)$$

2. Show that there exists $\mu_0^\varepsilon, \mu_1^\varepsilon$ converging weakly to μ_0, μ_1 as $\varepsilon \to 0$ such that

$$W_\nu^2(\mu_0, \mu_1) \geq \limsup_{\varepsilon \to 0} F_\nu^\varepsilon(\mu_0^\varepsilon, \mu_1^\varepsilon)$$

We tackle 1 first. Let $\mu_0^\varepsilon, \mu_1^\varepsilon$ converge weakly to μ_0, μ_1. Note that for all ε, letting γ_ε be optimal in (7) with marginals μ_0^ε and μ_1^ε, we have $\int_{X \times X \times X} |x_i - y|^2 d\gamma_\varepsilon(x_0, x_1, y) \geq W_2^2(\nu, \mu_i^\varepsilon)$, and so

$$F_\nu^\varepsilon(\mu_0^\varepsilon, \mu_1^\varepsilon) \geq \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma_\varepsilon(x_0, x_1, y). \tag{9}$$

Now, let γ_0 be a weak* limit point of the γ_ε. By Theorem 5.20 in [21][5], γ_0 solves (7) with $\varepsilon = 0$; that is, multi-marginal optimal transport with cost $c_0(x_0, x_1, y) = |x_0 - y|^2 + |x_1 - y|^2$ and marginals μ_0, μ_1 and ν. This means that $\gamma_{x_i y}$ must be optimal for optimal transport with the quadratic cost between μ_i and ν, for $i = 0, 1$, respectively. Therefore, γ_0 is a competitor in the problem (5) used to define W_ν and so

$$\int_{X \times X \times X} |x_0 - x_1|^2 d\gamma_0(x_0, x_1, y) \geq W_\nu^2(\mu_0, \mu_1).$$

Passing to the limit in (9), we get

$$\liminf_{\varepsilon \to 0} F_\nu^\varepsilon(\mu_0^\varepsilon, \mu_1^\varepsilon) \geq \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma_0(x_0, x_1, y) \geq W_\nu^2(\mu_0, \mu_1)$$

along each subsequence for which γ_ε converges. This establishes property 1. Property 2 is simpler; we may take $\mu_i^\varepsilon = \mu_i$ for $i = 0, 1$. To see this, note that choosing γ in (7) with optimal two-fold marginals, $\gamma_{x_i y} \in \Pi_{opt}(\mu_i, \nu)$ so that $\int_{X \times X \times X} |x_i - y|^2 d\gamma(x_0, x_1, y) = W_2^2(\mu_0, \mu_1)$, causes the last terms in F_ν^ε to cancel, we get

$$F_\nu^\varepsilon(\mu_0, \mu_1) \leq \int_{X \times X \times X} |x_0 - x_1|^2 d\gamma(x_0, x_1, y).$$

As this holds for each γ with $\gamma_{x_i y} \in \Pi_{opt}(\mu_i, \nu)$, we can minimize over all such γ to obtain

$$F_\nu^\varepsilon(\mu_0, \mu_1) \leq W_\nu^2(\mu_0, \mu_1).$$

Therefore,

$$\limsup_{\varepsilon \to 0} F_\nu^\varepsilon(\mu_0, \mu_1) \leq W_\nu^2(\mu_0, \mu_1)$$

as desired. $\qquad\square$

Remark 14. *In this paper, we will pay special attention to the case where ν is concentrated on a lower dimensional submanifold, parametrized by $f : Y \to X$, where $Y \subset \mathbb{R}^n$, with $n < m$, and f is a smooth injection. In this case, by a slight abuse of notation, we will often consider ν to be a probability measure on Y, and the quadratic optimal transport problem in the definition (4) is equivalent to optimal transport between X and Y with cost function $c(x, y) = |x - f(y)|^2$, or, equivalently, $c(x, y) = -x \cdot f(y)$.*

We next consider the case when ν concentrates on a line; we show below that in this case, W_ν corresponds to the layerwise-Wasserstein metric from [12], whose definition is recalled below.

---

[5] Although Theorem 5.20 in [21] is stated only for two marginal problems, the same result may be proven for multi-marginal problems in exactly the same way, using the equivalence between c-cyclical monotonicity and optimality established in [11].

Definition 15. *The layerwise-Wasserstein metric between measures $\mu_0, \mu_1 \in \mathcal{P}(X)$ is given by*

$$d_{LW}^2(\mu_0, \mu_1) := W_2^2(\mu_0^V, \mu_1^V) + \int_0^1 W_2^2(\tilde{\mu}_0^l, \tilde{\mu}_1^l)dl \qquad (10)$$

where the $\mu_i^V = \left(x = (x^1, x^2, ..., x^m) \mapsto x^1 \right)_{\#} \mu_i$ are the vertical marginals of the μ_i, $\tilde{\mu}_i$ are rescaled versions of the μ_i, defined by $\tilde{\mu}_i = (F_{\mu_i}, Id)_{\#}\mu_i$ where F_{μ_i} is the cumulative distribution function of μ_i^V, with the identity mapping Id applied on the last $m-1$ coordinates (so that $(F_{\mu_i}, Id)(x^1, x^2, ..., x^m) = (F_{\mu_i}(x^1), x^2, ..., x^m))$ and $\tilde{\mu}_i = \mu_i^l \otimes dl$ is disintegrated with respect to its uniform vertical marginal (see [12] for a more detailed description).

Proposition 16. *Suppose ν is concentrated on the line segment $\{(t, 0, 0, ...0) : t \in \mathbb{R}\}$ and is absolutely continuous with respect to 1-dimensional Hausdorff measure. Then W_ν is equal to the layerwise Wasserstein metric.*

Proof. Using the framework described in Remark 14 with $n = 1$ and $f(y) = (y, 0, 0, ..., 0)$, we consider optimal transport between μ_i on X and ν on $Y \subseteq \mathbb{R}$ with cost function $|x_i - f(y)|^2 = (x_i^1 - y)^2 + (x_i^2)^2 + ...(x_i^m)^2$, or, equivalently, the cost function $c(x_i, y) = (x_i^1 - y)^2$. This is an *index cost* in the terminology of [16]; that is, $c(x_i, y) = \tilde{b}(I(x_i), y)$ depends on $x_i \in \mathbb{R}^m$ only through the lower dimensional variable $I(x_i) = x_i^1$. For such costs, the optimal transport problem can be solved semi-explicitly, essentially by solving the optimal transport problem between the one dimensional variables $I(x_i) = x_i^1$ and y. We described the solution below.

The level set $(T_i)^{-1}(y)$ of the optimal transport map $T_i : X \to Y$ consists of the hyperplane $\{(z_i(y), x_i^2, ..., x_i^m)\}$ where the fixed $z_i(y)$ is chosen so that

$$\mu_i(\{(x_i^1, x_i^2, ...x_i^m) : x_i^1 \leq z_i(y)\}) = \nu(-\infty, y). \qquad (11)$$

By the first part of Theorem 12, the optimal arrangement γ in (5) then pairs the conditional probability μ_0^y on $\{z_0(y)\} \times \mathbb{R}^{m-1}$ with the corresponding conditional probability μ_1^y on $\{z_1(y)\} \times \mathbb{R}^{m-1}$.

By Theorem 12, we then have

$$W_\nu^2(\mu_0, \mu_1) = \int_Y W_2^2(\mu_0^y, \mu_1^y)d\nu(y)$$

Note that $|(z_0(y), x_0^2, ..., x_0^m) - (z_1(y), x_1^2, ..., x_1^m)|^2 = |(x_0^2, ..., x_0^m) - (x_1^2, ..., x_1^m)|^2 + (z_0(y) - z_1(y))^2$, so that $W_2^2(\mu_0^y, \mu_1^y) = W_2^2(\hat{\mu}_0^y, \hat{\mu}_1^y) + (z_0(y) - z_1(y))^2$, where the measures $\hat{\mu}_i^y := \left((z_i(y), x_i^2, ..., x_i^m) \mapsto (x_i^2, ..., x_i^m) \right)_{\#} \mu_i^y$ are measures on $\mathbb{R}^{m-1}$. Letting F be the quantile function of ν, and changing variables via $l = F(y)$, we have

$$\begin{aligned} W_\nu^2(\mu_0, \mu_1) &= \int_Y [W_2^2(\hat{\mu}_0^y, \hat{\mu}_1^y) + (z_0(y) - z_1(y))^2]d\nu(y) \\ &= \int_0^1 [W_2^2(\hat{\mu}_0^{F^{-1}(l)}, \hat{\mu}_1^{F^{-1}(l)}) + (z_0(F^{-1}(l)) - z_1(F^{-1}(l))^2]dl \end{aligned}$$

14

Note that the second term is exactly the Wasserstein metric between the first marginals of μ_0 and μ_1. Since the conditional probabilities $\tilde{\mu}_i^l$ are both supported on $\{l\} \times \mathbb{R}^{m-1}$, and $\left((l, x_i^2, ..., x_i^m) \mapsto (x_i^2, ..., x_i^m)\right)_{\#} \tilde{\mu}_i^l = \hat{\mu}_i^{F^{-1}(l)}$, we have $W_2^2(\tilde{\mu}_i^l, \tilde{\mu}_i^l) = W_2^2(\hat{\mu}_0^{F^{-1}(l)}, \hat{\mu}_1^{F^{-1}(l)}) + (l-l)^2 = W_2^2(\hat{\mu}_0^{F^{-1}(l)}, \hat{\mu}_1^{F^{-1}(l)})$. The last line above is then exactly the definition of the layerwise-Wasserstein metric. $\qquad\square$

As was noted in [12], in two dimensions the layerwise-Wasserstein metric corresponds to the Knothe-Rosenblatt rearrangement [13, 18] and so we immediately obtain the following.

Corollary 17. *Let $m = 2$. Then, under the assumptions in the preceding Proposition, the optimal rearrangement γ in (5) satisfies $\gamma_{x_0 x_1} = (Id, G)_{\#}\mu_0$, where G is the Knothe-Rosenblatt rearrangement.*

Proof. For each y, the level set of the optimal map $T^{-1}(y)$ is the line segment $\{(z_i(y), x_i^2)\}$ where (11) becomes:

$$\mu_i(\{(x_i^1, x_i^2) : x_i^1 \le z_i(y)\}) = \nu(-\infty, y). \qquad (12)$$

Thus, any matching γ with $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$ must couple y with $\{(z_i(y), x_i^2)\}$; it follows that $\gamma_{x_0 x_1}$ must couple $\{(z_0(y), x_0^2)\}$ to $\{(z_1(y), x_1^2)\}$. Since (12) implies $\mu_0(\{(x_0^1, x_0^2) : x_0^1 \le z_0(y)\}) = \mu_1(\{(x_1^1, x_1^2) : x_1^1 \le z_1(y)\})$, this means that $\gamma_{x_0 x_1}$ is montone in the first coordinate. The optimal way to couple the conditional probabilites, concentrated on $\{(z_0(y), x_0^2)\}$ and $\{(z_1(y), x_1^2)\}$ is then monotonic; this exactly characterizes the Knothe-Rosenblatt rearrangement. $\qquad\square$

Remark 18. *One can recover a similar characterization of the Knothe-Rosenblatt rearrangement in higher dimensions by looking at iterated optimization problems along an orthogonal basis. This corresponds to the limit of a multi-marginal problem where the interactions are weighted iteratively, as we show in Appendix A. This is closely related to the main result in [6], where the Knothe-Rosenblatt rearrangement is characterized as a limit of optimal transport maps with anisotropic costs.*

2.3 Dual metrics

We now define a class of metrics on a lower dimensional space which are *dual* to W_ν in a certain sense; more precisely, they use Kantorovich potentials $v(y)$ on the variable y arising from optimal transport to a fixed reference measure $\mu(x)$ to compare measures on the y variable, whereas W_ν uses in part the Kantorovich potential $v(y)$ for optimal transport from a reference measure $\nu(y)$ on y to measures on x in order to compare those free measures.

Fix reference measures $\mu \in \mathcal{P}(X)$ and $\sigma \in \mathcal{P}(Y)$ on bounded domains $X \subseteq \mathbb{R}^m$ and $Y \subseteq \mathbb{R}^n$, with $n \le m$, absolutely continuous with respect to m

and n dimensional Lebesgue measure, respectively. We let $c : \overline{X} \times \overline{Y} \to \mathbb{R}$ be a cost function which is C^1 up to the boundary and satisfies the *twist condition*; that is, for each fixed $x \in X$, the mapping from $\overline{Y}$ to $\mathbb{R}^m$ defined by

$$y \mapsto D_x c(x, y)$$

is injective. Note that $c \in C^1(\overline{X} \times \overline{Y})$ implies that c is Lipschitz, since $\overline{X}$ and $\overline{Y}$ are compact.

We let $\mathcal{P}_{ac,\sigma}(Y)$ be the set of probability measures on Y which are absolutely continuous with respect to σ. For measures $\nu_0, \nu_1 \in \mathcal{P}_{ac,\sigma}(Y)$, we define

$$W^*_{\mu,\sigma,c,p}(\nu_0, \nu_1) := ||Dv_0 - Dv_1||_{L^p(\sigma)} \tag{13}$$

for some $p \in [1, \infty]$ where v_i is the c-concave Kantorovich potential (that is, c-conjugate solutions to (3) corresponding to the optimal transport problem between μ and ν_i with cost function c).

Proposition 19. $W^*_{\mu,\sigma,c,p}$ *is a metric on* $\mathcal{P}_{ac,\sigma}(Y)$.

Proof. Since the L^p norm clearly induces a metric, all that needs to be proven is that for each measure $\nu \in \mathcal{P}_{ac,\sigma}(Y)$, the gradient Dv of the Kantorovich potential is in $L^p(\sigma)$, and that the mapping $\nu \mapsto Dv$ is a bijection. Since the cost function is Lipschitz, a now standard argument originating in [15], implies that the potential v is Lipschitz; absolute continuity of ν with respect to σ (and therefore Lebesgue measure) ensures that it exists σ almost everywhere, by Rademacher's theorem. Furthermore, since v is semi-convex and Lipschitz, the gradient Dv is also bounded, and so $Dv \in L^p(\sigma)$ for all $p \geq 1$.

Absolute continuity of ν with respect to σ, and therefore Lebesgue measure, ensures that Dv is uniquely determined by ν [19, Proposition 1.15].

On the other hand, it is well known that the twist condition ensures that the unique optimizer to the optimal transport problem between μ and ν concentrates on a graph, $y = T(x)$, where T is uniquely determined by the gradient of the potential $u(x)$ via the equation $Du(x) = D_x c(x, T(x))$. Since $\nu = T_\# \mu$, we have that ν is uniquely determined by T and therefore u. The c-concave Kantorovich potential $v(y) = \inf_{x \in X}[c(x, y) - u(x)]$ is then clearly in one to one correspondence with $u(x) = \inf_{y \in Y}[c(x, y) - v(y)]$ and, therefore, it uniquely determines ν. Thus ν and Dv are in one-to-one correspondence, as desired.

$\square$

To help explain the motivation behind this metric, it is useful to first consider the $m = n$ case with $c(x, y) = |x - y|^2$. The metric $W^*_{\mu,\sigma,c,2}$ is then similar to the ν-based Wasserstein metric. Recall from Example 9 that in this case $W_\nu(\mu_0, \mu_1)$ between two measures in $\mathcal{P}(X)$ coincides with the L^2 metric on the optimal maps from ν to the μ_i; on the other hand, $W^*_{\mu,\sigma,c,2}(\nu_0, \nu_1)$ for two measures in $\mathcal{P}(Y)$ coincides with the L^2 metric on the optimal maps from the ν_i to μ. That is, $W^*_{\mu,\sigma,c,2}(\nu_0, \nu_1)$ compares gradients of potential on the

opposite, rather than same, side of the problem from the reference measure, and so is in some sense dual to W_ν in Example 9.

The case where $m > n$ is an attempt to generalize this idea in the same way that the ν-based Wasserstein metric generalizes the L^2 metric in Example 9. We emphasize that the motivating example is when $c = |x - f(y)|^2$ is quadratic between x and $f(y)$ in a lower dimensional submanifold, using the convention in Remark 14. In this case, $W^*_{\mu,\sigma,c,p}$ is dual to the ν based Wasserstein metric, in the sense that it relies on the Kantorovich potential on the target, rather than source, side.

The metric $W^*_{\mu,\sigma,c,p}$ arises naturally in certain variational problems among measures on Y involving optimal transport to the fixed measure μ on X, which is more naturally described by the potentials v on Y than u on X. We expand on this viewpoint in section 6 below, where we establish a fixed point characterization of solutions to a class of these problems. In the proof, we actually use a metric which corresponds to using the mass splitting k, introduced in [16] and recalled in equation (21) below, rather than potentials (these are the same when the problem satisfies the nestedness condition introduced in [16], as is the case for the optimal ν under appropriate conditions identified in [17] and recalled in (23) below, but not in general).

3 Unequal dimensional optimal transport

When the base measure ν concentrates on a lower dimensional submanifold, the definition (5) of the ν-Wasserstein metric relies on unequal dimensional optimal transport. We briefly review the known theory on this topic here, and establish a lemma which we will need later on.

For $c(x, y) = -x \cdot f(y)$ as in Remark 14, $y \in Y$ and $p \in -P_y$, where $-P_y := \{-x \cdot Df(y) \in X : x \in X\}$, we define, as in [16], noting that $D_y c(x, y) = -x \cdot Df(y)$

$$X_1(y, p) := \{x \in X : -x \cdot Df(y) = p\}. \tag{14}$$

Note that, since our cost is linear in X, each $X_1(y, p)$ is an affine $(m - n)$-dimensional submanifold of X, provided that $Df(y)$ has full rank. For a fixed y, each $X_1(y, p)$ is parallel.

The following result shows that when we disintegrate the optimizer with respect to ν, the conditional probabilities are absolutely continuous with respect to $(m - n)$-dimensional Hausdorff measure, and provides a formula for their densities.

Theorem 20. *Consider optimal transport between probability measures μ on $X \subseteq \mathbb{R}^m$ and ν on $Y \subseteq \mathbb{R}^n$, where $m > n$, both absolutely continuous with respect to Lebesgue measure, with densities $\overline{\mu}(x)$ and $\overline{\nu}(y)$, respectively, and cost function $c(x, y) = -x \cdot f(y)$ where $f \in C^2(Y)$ is injective and non-degenerate, that is $Df(y)$ has full rank everywhere.*

Then there is a unique optimal measure $\pi = (Id, T)_{\#}\mu$ in (2), which concentrates on the graph of a function $T : X \to Y$. The Kantorovich potential $v(y)$ is differentiable for almost every y, and at each such point $T^{-1}(y)$ is concentrated on the affine manifold $X_1(y, Dv(y))$. The conditional probability $\mu^y(x)$ corresponding to the disintegration $\mu = \mu^y \otimes \nu(y)$ of μ with respect to T is absolutely continuous with respect to $(m - n)$ - dimensional Hausdorff measure on $X_1(y, Dv(y))$ for almost every y. Furthermore, there is a subset $X(v^c) \subset X$ of full volume such that this conditional probability is concentrated on $T^{-1}(y) \cap X(v^c)$ with density $\overline{\mu}^y(x)$ with respect to $(m - n)$ – dimensional Hausdorff measure given by

$$\overline{\mu}^y(x) = \overline{\mu}(x)/[\overline{\nu}(y)JT^y(x)] \tag{15}$$

where $JT^y(x) = \frac{\sqrt{\det[Df(y)Df^T(y)]}}{\det[-x\cdot D^2f(y)-D^2v(y)]}$ is the n-dimensional Jacobian of the optimal transport map defined by $T(x) = (f)^{-1}(-Dv^c(x))$, evaluated at any point $x \in T^{-1}(y)$.

Proof. The existence and uniqueness of an optimal transport map T is well known. The remainder of the proof is based on the proof of Theorem 2 in [16]. We repeat the preliminary setup of that argument for the reader's convenience. Letting $u(x) = v^c(x)$ and $v(y) = u^c(y)$ be the Kantorovich potentials, it is well known that u is semi-concave and therefore its differential $Du(x)$ exists almost everywhere and is locally of bounded variation (see Theorem 6.8 in [10]). Theorem 6.13 in [10] then implies that for each integer $i > 0$ there exists a set Z_i of Lebesgue measure less than $1/i$, and a continuously differentiable map H_i with $DH_i = Du$ and $D^2H_i = D^2u$ outside of Z_i. The Z_i may be taken to be nested, that is, $Z_{i+1} \subset Z_i$, and we set $Z_\infty = \cap_{i=1}^\infty Z_i$ and $X(v^c) = dom(D^2u) \setminus Z_\infty$.

Now set $\tilde{Y} = dom(D^2v)$ and $\tilde{X} = T^{-1}(\tilde{Y})$. We note that $u(x) + v(y) - c(x, y) \leq 0$, with equality at $(x, T(x))$. For $x \in \tilde{X}$, differentiating with respect to $y = T(x) \in \tilde{Y}$ yields $Dv(y) = x \cdot Df(y)$, or $Dv(T(x)) = x \cdot Df(T(x))$. If $x \in dom(DT)$, differentiating with respect to x then yields:

$$[D^2v(T(x)) - x^t D^2 f(y)]DT(x) = Df(T(x)).$$

As the right hand side has full rank by assumption, this means that the $m \times n$ matrix $DT(x)$ must also have full rank, and so its Jacobian $JT(x) = det^{1/2}[DT^t(x)DT(x)] > 0$ must be non-zero. We then set $\tilde{X}_i = \{x \in \tilde{X} \setminus Z_i : |x| \leq i, JT(x) > 1/i\}$ and $X_i = \tilde{X}_i \setminus \tilde{X}_{i-1}$, so that $\tilde{X} \setminus Z_\infty = \cup_{i=1}^\infty X_i$ and the union is pairwise disjoint,

Now, equation (30) in [16] implies that for any $h \in L^\infty(\mathbb{R}^m)$ we have

$$\int_{X_i} h(x)\overline{\mu}_i(x)dx = \int_Y \overline{\nu}(y)\frac{1}{\overline{\nu}(y)}dy \int_{T_i^{-1}(y)\cap X_i} \frac{h(x)\overline{\mu}_i(x)}{JT_i^y(x)}d\mathcal{H}^{m-n}(x) \tag{16}$$

$$= \int_Y \overline{\nu}(y)\frac{1}{\overline{\nu}(y)}dy \int_{T_i^{-1}(y)\cap X(v^s)} \frac{h(x)\overline{\mu}_i(x)}{JT_i^y(x)}d\mathcal{H}^{m-n}(x) \tag{17}$$

18

One can then sum on i to obtain

$$\int_X h(x)\overline{\mu}(x)dx = \int_Y \overline{\nu}(y)dy \int_{T^{-1}(y)\cap X(v^s)} \frac{h(x)\overline{\mu}(x)}{\overline{\nu}(y)JT^y(x)} d\mathcal{H}^{m-n}(x). \qquad (18)$$

As this holds for every $h \in L^\infty(\mathbb{R}^m)$, it implies the desired disintegration. $\quad\square$

3.1 Nestedness and its high dimensional generalization

In what follows, in Section 6 in particular, we will sometimes consider optimal transport between μ and ν with respect to a general cost, $c(x,y)$ on $X \times Y \subseteq \mathbb{R}^m \times \mathbb{R}^n$, with $m \geq n$. In this case, we will denote the analogue of $X_1(y,p)$ by

$$X_1^c(y,p) := \{x \in X : D_y c(x,y) = p\}. \qquad (19)$$

Assuming that c is non-degenerate, so that the $m \times n$ matrix $D_{xy}^2 c(x,y)$ has full rank n, the implicit function theorem implies that $X_1^c(y,p)$ is an $m-n$ dimensional submanifold of X. If we denote by $v(y)$ the Kantorovich potential, we have

$$c(x,y) - v(y) \geq u(x) = \inf_{y' \in Y}(c(x,y') - v(y'))$$

with equality γ almost everywhere. For (x,y) such that equality holds, this implies that $y' \mapsto c(x,y') - v(y')$ is minimized at $y' = y$.

At any such point where v is differentiable, we therefore obtain $Dv(y) = D_y c(x,y)$, so that

$$\partial^c v(y) := \{x \in X : c(x,y) - v(y) = \inf_{y' \in Y}(c(x,y') - v(y'))\} \subseteq X_1^c(y,Dv(y)).$$

The generalized nestedness condition, introduced in [16] is that this inclusion is actually an equality.

Definition 21. *We say that* (c,μ,ν) *satisfies the* generalized nestedness condition *if for all* y *at which* v *is differentiable, we have*

$$\partial^c v(y) = X_1^c(y,Dv(y))$$

where v *is the c-conjugate Kantorovich potential for transport between* u *and* v.

The importance of this condition, uncovered in [16], is as follows. Solving the optimal transport problem (2) is essentially equivalent to solving the dual problem (3). In turn, it is shown in [16] that the solution (v^c, v) to (3) can be recovered by solving a partial differential equation for v, which is in general *non-local* and therefore extremely challenging. Under the generalized nestedness condition, however, it turns out that this equation reduces to a local equation, and therefore becomes much more tractable.

The origin of the term generalized nestedness comes from the case when $Y = (\underline{y}, \overline{y})$ is one dimensional, in which case it can be shown that generalized nestedness is essentially equivalent to nestedness of the corresponding superlevel sets. More precisely, we define:

$$X^c_{\geq}(y, p) := \{x \in X : D_y c(x, y) \geq p\}. \tag{20}$$

Assuming that μ and ν are both absolutely continuous, with densities bounded above and below, for each fixed y, we note that the measure $\mu(X^c_{\geq}(y, k))$ of the super level set $X^c_{\geq}(y, k)$ is a decreasing function of k. There will then be a unique $k = k(y)$ satisfying the population balancing condition:

$$\mu(X^c_{\geq}(y, k)) = \nu((\underline{y}, y)). \tag{21}$$

The definition of nestedness found in [9] is then the following.

Definition 22. *Assume that Y is one dimensional. We say the model (c, μ, ν) is nested if*

$$X^c_{\geq}(y, k(y)) \subseteq X^c_{\geq}(y', k(y'))$$

whenever $y \leq y'$.

The relevance of the nestedness condition is that it allows one to almost explicitly solve (2) [9]. One can attempt to define a mapping $T : X \to Y$ that maps each $x \in X^c_1(y, k(y))$ to y. In general, this may not result in a well defined mapping, since if $X^c_1(y, k(y))$ intersects $X^c_1(y', k(y'))$ for $y \neq y'$, the procedure attempts to map each x in the intersection to *both* y and y'. It turns out that the nestedness condition ensures that this does not happen; furthermore, not only is the mapping T described above well defined for nested models, but the measure $(Id, T)_{\#}\mu$ is the solution to the Kantorovich problem (2) and the $k(y)$ defined above coincides with the derivative $v'(y)$ of the Kantorovich potential.

Before closing this section, we illustrate the nestedness condition with the following simple example from [9]. Several other examples may be found in [9] [8] [7].

Example 23. *Let $c(x, y) = -x \cdot (\cos(y), \sin(y)) = -x_1 \cos(y) - x_2 \sin(y)$, μ be uniform on the subset $X = \{x : 0 < r < 1, 0 < \theta < \overline{\theta}\} \subset \mathbb{R}^2$, of the unit ball in $\mathbb{R}^2$, expressed in polar coordinates, and ν be uniform on $Y = (0, \overline{\theta})$. This corresponds to to optimal transport with the quadratic cost between X and the corresponding circular arc, parametrized by the angle y. The solution clearly maps each x radially outward to $x/|x|$, so the optimal map is $y = T(x) = \arctan(x_2/x_1)$. We will illustrate the nestedness concept by considering two cases; in the first case, the model is nested, and so the procedure described above successfully constructs the optimal mapping. In the second, the model is not nested, and the procedure fails to produce a well defined mapping.*

    ***Case 1:** Suppose that $\overline{\theta} < \pi$. Then $X^c_1(y, k) = \{X \in X : x \cdot (\sin(y), -\cos(y)) = k\}$ is a line segment parallel to $(\cos(y), \sin(y))$ and $X^c_{\geq}(y, k)$ is the portion of*

X lying below this line segment. The balancing condition (21) is satisfied for $k = k(y) = 0$, in which case $X_1^c(y, k)$ passes through the origin and the mass of the wedge $X_{\geq}^c(y, k) = \{x : 0 < r < 1, 0 < \theta < y\}$ satisfies

$$\mu(X_{\geq}^c(y, k)) = y/\bar{\theta} = \nu((0, y)).$$

These regions $X_{\geq}^c(y, k(y)) = X_{\geq}^c(y, 0)$ are nested, $X_{\geq}^c(y, 0) = \{x : 0 < r < 1, 0 < \theta < y\} \subset \{x : 0 < r < 1, 0 < \theta < y'\} = X_{\geq}^c(y', 0)$ for $y < y'$, and indeed, mapping each $x \in X_1^c(y, 0)$ to y coincides with optimal map, while setting $v'(y) = k(y)$ yields the Kantorovich potential.

    Case 2: Suppose now that $\pi < \bar{\theta} < 2\pi$. Then for $y < \bar{\theta} - \pi$, note that the line segment $X_1^c(y, 0)$ through the origin now intersects X in the quadrant $\{x_1, x_2 < 0\}$ and so $X_{\geq}^c(y, 0) = \{x : 0 < r < 1, 0 < \theta < y\} \cup \{x : 0 < r < 1, \pi + y < \theta < \bar{\theta}\}$; we therefore have $\mu(X_{\geq}^c(y, 0)) = y/\bar{\theta} + \bar{\theta} - \pi - y > y/\bar{\theta} = \nu((0, y))$. The $k = k(y)$ satisfying (21) must therefore be larger than 0. Assigning each x in $X_{\leq}^c(y, k(y))$ to y then clearly does not correspond to the optimal map, and so the result in [9] implies that the model cannot be nested. Alternatively, this can be seen directly by noting that $X_{\geq}^c(y, k(y))$ contains some point in the quadrant $\{x_1, x_2 < 0\}$ for $y < \bar{\theta} - \pi$, whereas for $\bar{\theta} - \pi < y' < \pi$, $k(y') = 0$ as before, so $X_{\geq}^c(y', k(y')) = \{x : 0 < r < 1, 0 < \theta < y\}$ is contained in the region where $x_2 > 0$.

3.2   Sufficient conditions for nestedness

In our previous work [17], motivated by problems in economics and game theory, we considered the problem of minimizing among probability measures in $\mathcal{P}([\underline{y}, \overline{y}])$ a class of functionals including:

$$\nu \mapsto T_c(\mu, \nu) + \int \overline{\nu}(y) \log(\overline{\nu}(y)) dy, \tag{22}$$

for a fixed measure μ on the bounded domain $X \subseteq \mathbb{R}^m$, absolutely continuous with respect to Lebesgue measure, and a fixed domain $Y := (\underline{y}, \overline{y}) \subset \mathbb{R}$. Here $T_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) d\pi(x, y)$ is the cost of optimal transport between μ and ν with respect to the cost function $c(x, y)$ and $\overline{\nu}$ is the density of $d\nu(y) = \overline{\nu}(y) dy$ with respect to Lebesgue measure (if ν is not absolutely continuous, the functional is taken to be infinite).

    In [17], we showed that the solution to (22) is nested provided that:

$$\sup_{y_1 \in Y, y_0 \leq y \leq y_1} \frac{D_{\mu}^{min}(y_0, y_1, k(y_0))}{y_1 - y_0} - M_c \frac{e^{-M_c y}}{e^{M_c \overline{y}} - e^{M_c \underline{y}}} - 1 < 0 \tag{23}$$

for all $y_0 \in Y$ Here, $k(y_0)$ is defined by (21),

$$D_{\mu}^{min}(y_0, y_1, k_0) = \mu(X_{\geq}(y_1, k_{max}(y_0, y_1, k_0)) \setminus X_{\geq}^c(y_0, k_0))$$

21

is the minimal mass difference where $k_{max}(y_0, y_1, k_0) = \sup\{k : X_\geq(y_0, k_0) \subseteq X_\geq(y_1, k)\}$ and $M_c := \sup_{(x,y_0,y_1) \in (\overline{X} \times \overline{Y} \times \overline{Y})} \frac{|c(x,y_0) - c(x,y_1)|}{|(x,y_0) - (x,y_1)|}$ is the Lipschitz norm with respect to variations in y (note that the condition has been reformulated slightly from [17], to match the choice of functional in (22); in [17] more general functionals were studied).

4 Uniqueness of optimal couplings and structure of geodesics when ν concentrates on a lower dimensional submanifold

We note that geodesics for the metric W_ν are generalized geodesics with base ν, according to the general definition in [1]. When ν is sufficiently regular (for instance, absolutely continuous with respect to Lebesgue measure, as in Example 9), generalized geodesics are *uniquely* determined by μ_0 and μ_1. In this case, $\Pi_{opt}(\nu, \mu_i)$ consists of a single measure concentrated on the graph of a function $\tilde{T}_i : X \mapsto X$ pushing ν forward to μ_i; $\gamma = (Id, \tilde{T}_0, \tilde{T}_1)_{\#}\nu$ is then the unique measure satisfying $\gamma_{yx_i} \in \Pi_{opt}(\nu, \mu_i)$, and therefore the unique coupling minimizing the integral in the definition of $W_\nu(\mu_0, \mu_1)$. If ν is singular, this reasoning does not apply; existence of minimizers in (5) follows immediately from standard continuity-compactness results, but they may be non-unique in general.

Here we specialize to the case when ν concentrates on some lower dimensional submanifold. Given Proposition 16, we can regard this as a sort of non-linear generalization of the layerwise-Wasserstein metric. In this setting, if the μ_i are absolutely continuous with respect to Lebesgue, we obtain a uniqueness result as a consequence of Theorem 12 and Lemma 26. We also characterize geodesics with respect to our metric is this setting.

Theorem 24. *Assume that both μ_i are absolutely continuous with respect to Lebesgue measure and that ν is supported on a smooth n-dimensional submanifold of X, and is absolutely continuous with respect to n-dimensional Hausdorff measure. Then the optimal coupling γ in (5) is unique and its two-fold marginal $\gamma_{x_0 x_1}$ is induced by a mapping $G : X \to X$ pushing μ_0 forward to μ_1. Letting T_i denote the optimal maps from μ_i to ν, the restriction of G to $T_i^{-1}(y)$ is the optimal mapping pushing μ_0^y forward to μ_1^y for a.e. y, where, as in Theorem 12, $\mu_i(x) = \mu_i^y(x) \otimes \nu(y)$. This mapping takes the form $\nabla\varphi^y : X_1(y, Dv_0(y)) \to X_1(y, Dv_1(y))$, where $\varphi^y : X_1(y, Dv_0(y)) \to \mathbb{R}$ is a convex function defined on the $(m-n)-$ dimensional affine subset $X_1(y, Dv_0(y))$.*

Proof. Brenier's theorem implies that the optimal $\pi_i = (Id, T_i)_{\sharp}\mu_i$ between μ_i and ν is unique and concentrated on the graph of a function T_i for $i = 0, 1$ [5]; the conditional probabilities μ_i^y are concentrated on $T_i^{-1}(y)$ which are clearly pairwise disjoint. Using the first characterization in Theorem 12, it is enough to show that for almost every y, there is a unique optimal coupling be-

tween the $\mu^i(y)$ and that coupling is concentrated on the graph of a function over $\mu^0(y)$. Furthermore, [16] implies that for almost every y the conditional probabilities μ_i^y are concentrated on $m - n$ dimensional affine subspaces $X_1(y, Dv_i(y)) = \{x : x \cdot Df(y) = Dv_i(y)\}$ where $f : Y \to X$ locally parameterizes the submanifold supporting ν, as in Remark 14 and v_i is the Kantorovich potential and, by Theorem 20, each μ_i^y is absolutely continuous with respect to $m - n$ dimensional Hausdorff measure. The subspaces $X_1(y, Dv_0(y))$ and $X_1(y, Dv_1(y))$ containing $T_0^{-1}(y)$ and $T_1^{-1}(y)$ respectively, are parallel. We decompose each $x = x^N + x^\perp$, where x^N is in the $m - n$ dimensional null space $null(Df(y))$ of $Df(y)$ and $x^\perp$ in its n-dimensional orthogonal complement, $null(Df(y))^\perp$, on which the linear map $Df(y)$ is injective. Then for $x_i \in X_1(y, Dv_i(y))$, $x_i^\perp$ must be the unique element of $null(Df(y))^\perp$ such that $(x_i^\perp)^T Df(y) = Dv(y_i)$; that is, $x_i^\perp$ is *constant* on $X_1(y, Dv_i(y))$.

The quadratic cost between $X_1(y, Dv_0(y))$ and $X_1(y, Dv_1(y))$ then takes the form $|x_0 - x_1|^2 = |x_0^N - x_1^N|^2 + C$, where $C = |x_0^\perp - x_1^\perp|^2$ is constant throughout $X_1(y, Dv_0(y)) \times X_1(y, Dv_1(y))$ and therefore doesn't affect optimal transport. The optimal transport problem between μ_0^y and μ_1^y is then equivalent to optimal transport with respect to the quadratic cost on $null(Df(y))$, or equivalently, $\mathbb{R}^{m-n}$. The existence, uniqueness and structure (as the gradient of a convex potential) of the optimal mapping between them then follows from another application of Brenier's theorem. Thus, for almost every y, the subset where the optimal map fails to be uniquely defined has $m - n$ Lebesgue measure 0, yielding the desired result. $\qquad\square$

We now turn to the study of minimizing W_ν geodesics.

Proposition 25. *Suppose that $\pi_0, \pi_1 \in \Pi_{opt}(\nu)$. Then $\pi_t = ((1-t)e_0 + te_1, e_2)_{\#}\gamma$ is a minimizing geodesic for the ν-based Wasserstein metric on couplings defined by (6), for each optimal γ in (6), where $e_0(x_0, x_1, y) = x_0$, $e_1(x_0, x_1, y) = x_1$, and $e_2(x_0, x_1, y) = y$,*

Suppose that $\mu_0, \mu_1 \in \mathcal{P}_\nu^u$. Then the generalized geodesic $\mu_t = ((1-t)e_0 + te_1)_{\#}\gamma$ is a minimizing geodesic for the ν based Wasserstein metric for each optimal γ in (5), provided that $\mu_t \in \mathcal{P}_\nu^u$ for all $t \in [0, 1]$.

Proof. We tackle the first statement first. As in the proof of Lemma 9.2.1 in [1], $\pi_t \in \Pi_{opt}(\nu)$. It is clear that $\gamma_t = (e_0, (1-t)e_0 + te_1, e_2)_{\#}\gamma$ has two-fold marginals $\gamma_{x_0 y} = \pi_0$ and $\gamma_{x_t y} = \pi_t$. Therefore,

$$
\begin{aligned}
\tilde{W}_\nu^2(\pi_0, \pi_t) &\leq \int_{X \times X \times X} |x_0 - x_t|^2 d\gamma_t(x_0, x_t, y) \\
&= \int_{X \times X \times X} |x_0 - [(1-t)x_0 + tx_1]|^2 d\gamma(x_0, x_1, y) \\
&= \int_{X \times X \times X} t^2 |x_0 - x_1|^2 d\gamma(x_0, x_1, y) = t^2 \tilde{W}_\nu^2(\pi_0, \pi_1)
\end{aligned}
$$

So $\tilde{W}_\nu(\pi_0, \pi_t) \leq t\tilde{W}_\nu(\pi_0, \pi_1)$. A very similar argument implies $\tilde{W}_\nu(\pi_t, \pi_s) \leq (s-t)\tilde{W}_\nu(\pi_0, \pi_1)$ for $s > t$, and $\tilde{W}_\nu(\pi_s, \pi_1) \leq (1-s)\tilde{W}_\nu(\pi_0, \pi_1)$. Adding these three inequalities yields

$$
\begin{aligned}
\tilde{W}_\nu(\pi_0, \pi_t) + \tilde{W}_\nu(\pi_t, \pi_s) + W_\nu(\pi_s, \pi_1) &\leq t\tilde{W}_\nu(\pi_0, \pi_1) + (s-t)\tilde{W}_\nu(\pi_0, \pi_1) + (1-s)\tilde{W}_\nu(\pi_0, \pi_1) \\
&= \tilde{W}_\nu(\pi_0, \pi_1)
\end{aligned}
$$

Since two applications of the triangle inequality imply the opposite inequality, $\tilde{W}_\nu(\pi_0, \pi_1) \leq \tilde{W}_\nu(\pi_0, \pi_t) + \tilde{W}_\nu(\pi_t, \pi_s) + W_\nu(\pi_s, \pi_1)$, we must have equality throughout; in particular $\tilde{W}_\nu(\pi_t, \pi_s) = (s-t)\tilde{W}_\nu(\pi_0, \pi_1)$, and so π_t is a minimizing geodesic for the ν-based Wasserstein metric on couplings.

The proof of the second assertion is very similar and is omitted. The only subtely is that in general it is not clear to us that μ_t remains in the set $\mathcal{P}^u_\nu$ on which W_ν is a metric, so that property is added as an assumption. $\qquad\square$

Under much stronger conditions on the marginals μ_0 and μ_1 and the reference measure ν, we are able to use Theorem 24 to show that the geodesic for the W_ν metric between μ_0 and μ_1 is unique.

Lemma 26 (Structure of geodesics). *Let μ_0, μ_1 and ν be as in Theorem 20. Then the geodesic μ_t between μ_0 and μ_1 is uniquely defined. The Kantorovich potential for optimal transport between μ_t and ν is $v_t = tv_1 + (1-t)v_0$, where $v_i : Y \to \mathbb{R}$ is the Kantorovich potential between μ_i and ν, for $i = 0, 1$. For almost every y, the conditional probability μ_t^y, concentrated on $X_1(y, Dv_t(y))$, is the unique displacement interpolant between μ_0^y and μ_1^y. It is absolutely continuous with respect to $(m-n)$-dimensional Hausdorff measure; denoting its density by $\overline{\mu}_t^y$, and the density of μ_t with respect to $\mathcal{L}^m$ by $\overline{\mu}_t$, we have, $\overline{\mu}_t^y(x_t) = \overline{\mu}_t(x_t)/(JT_t^y(x_t)\overline{\nu}(y))$ where $JT_t^y(x_t) := \frac{\sqrt{\det(Df^T Df(y))}}{\det[-D^2 v_t(y) - x_t \cdot D^2 f(y)]}$ is the Jacobian of the optimal map T_t from μ_t to ν, evaluated at $x_t \in T_t^{-1}(y) \subseteq X_1(y, Dv_t(y))$.*

The Brenier map between μ_0^y and μ_t^y is then given by $\nabla \varphi_t^y = (1-t)I + t\nabla\varphi^y$, where $\nabla\varphi^y : T_0^{-1}(y) \to T_1^{-1}(y)$ is the Brenier map between μ_0^y and μ_1^y.

Proof. Constructing v_t and each μ_t^y as in the statement of the Lemma, it is clear from Theorem 20 that v_t is the Kantorovich potential for optimal transport between $\mu_t(x) = \int_Y \mu_t^y(x)d\nu(y)$ and ν, and that $\pi_t = \nu(y) \otimes \mu_t^y(x)$ is the unique optimal transport plan between μ_t and ν. The first part of Theorem 12 then implies that, for any s, t

$$
W_\nu^2(\mu_s, \mu_t) = \int_X W_2^2(\mu_t^y, \mu_s^y)d\nu(y) = (t-s)^2 \int_X W_2^2(\mu_0^y, \mu_1^y)d\nu(y) = (t-s)^2 W_\nu^2(\mu_0, \mu_1)
$$
(24)

since each μ_t^y is a displacement interpolant (that is, a Wasserstein geodesic) between μ_0^y and μ_1^y. This implies that μ_t is a W_ν geodesic.

Furthermore, the uniqueness of the displacement interpolants μ_t^y together implies that any other curve $\tilde{\mu}_t \in \mathcal{P}^u_\nu(X)$, disintegrated with respect to ν

as $\tilde{\mu}_t^y(x) \otimes \nu(y)$, cannot satisfy (24) for all s, t (as otherwise $\tilde{\mu}_t^y$ would be a Wasserstein geodesic between μ_0^y and μ_1^y for ν all y, which would violate uniqueness). This yields the desired uniqueness of the W_ν geodesic.

The absolute continuity of μ_t^y with respect to $(m-n)$-dimensional Hausdorff measure on the affine set $X_1(y, Dv_t(y))$ follows from standard arguments, since it is the displacement interpolant of two absolutely continuous measures on parallel affine sets. Absolute continuity of μ_t and the formula relating $\overline{\mu}_t$ and $\overline{\mu}_t^y$ then follows from Theorem 20. Finally, the asserted structure $\nabla \varphi_t^y = (1-t)I + t\nabla\varphi^y$ of the Brenier map between between μ_0^y and μ_t^y then follows by standard optimal transport arguments applied to the parallel affine subspaces $X_1(y, Dv_0(y))$ and $X_1(y, Dv_1(y))$. $\qquad\square$

5　Convexity properties

We now turn our attention to certain convexity properties of W_ν geodesics. We begin by studying the convexity of various functionals along the geodesics, and then show that the set of measures satisfying a slight strengthening of the generalized nested condition is geodesically convex.

5.1　Geodesic convexity of functionals

We consider the functionals

$$\mu \quad \mapsto \quad W_2^2(\mu, \nu) \tag{25}$$

$$\mu \quad \mapsto \quad \int_X V(x)d\mu(x) \tag{26}$$

$$\mu \quad \mapsto \quad \int_{X^2} W(x-z)d\mu(x)d\mu(z) \tag{27}$$

and

$$\mu \mapsto \begin{cases} \int U(\overline{\mu}(x))dx & \text{if } d\mu = \overline{\mu}(x)dx \text{ is a.c. wrt Lebesgue measure on } X \\ +\infty & \text{otherwise.} \end{cases} \tag{28}$$

Convexity of each of these functionals along more classical interpolants has been studied extensively. Under suitable conditions, specified below, functionals (26) – (28) were first shown to be convex along Wasserstein geodesics (that is, displacement convex) by McCann [14]. Functionals (25) – (27) were shown to be convex along *every* generalized geodesic with base measure ν in [1]. Functional (28), on the other hand, is not convex along every generalized geodesics, but it is shown in Proposition 9.3.1 of [1] that for each $\mu_0, \mu_1 \in \mathcal{P}(X)$, there exists a generalized geodesic joining μ_0 to μ_1 along which it is convex. In particular, when the base measure ν is absolutely continuous with respect to Lebesgue measure, and so the generalized geodesic between any μ_0 and μ_1 is

unique, (28) is convex along every generalized geodesics, but this is often not true for singular base measures.

We will show that these functionals are geodesically convex with respect to the metric W_ν (under the usual conditions, specified below). Explicitly, geodesic convexity means that $\mathcal{F}(\mu_t)$ is a convex function of $t \in [0, 1]$, where $\mathcal{F} : \mathcal{P}(X) \to \mathbb{R}$ is any of (27), (28), (26) or (25). We will restrict our attention to W_ν geodesics μ_t of the form in Proposition 25; that is, $\mu_t = (te_0 + (1-t)e_1)_{\#}\gamma$, where γ is optimal in (5), under the assumption that each $\mu_t \in \mathcal{P}_\nu^u(X)$, since in more general situations the existence of a minimizing W_ν geodesic in the metric space $\mathcal{P}_\nu^u$ joining μ_0 and μ_1 is not clear. We are particularly interested in the case when the endpoints μ_0 and μ_1 and reference measure ν satisfy the assumptions in Theorem 24, under which the existence of a unique geodesic follows from Lemma 26.

Since the geodesics we consider for W_ν are always generalized geodesics, the following result follows immediately from Lemma 9.2.1 and Propositions 9.3.2 and 9.3.5 [1].

Proposition 27. *The following functionals are convex along all minimizing W_ν geodesics of the form in Proposition 25, under the corresponding conditions:*

1. *(25) is geodesically 1-convex for any ν.*

2. *(26) is geodesically convex if V is convex. It is geodesically strictly convex if V is strictly convex.*

3. *(27) is geodesically convex if W is convex. It is geodesically strictly convex along the subset of measures with the same barycenter if W is strictly convex.*

The geodesic convexity of the internal energy is somewhat more involved; in particular, as discussed above, unlike the other three forms, it does not hold for all generalized geodesics. Our proof, under additional restrictions on ν, uses the structure of the transport map G in a crucial way.

We assume that the reference measure ν is absolutely continuous with respect to n-dimensional Lebesgue measure on the n-dimensional submanifold Y (using the convention in Remark 14) and consider only geodesics μ_t whose endpoints μ_0 and μ_1 are absolutely continuous with respect to m-dimensional Lebesgue measure on X with densities with respective densities $\overline{\mu}_0$ and $\overline{\mu}_1$. Letting G be the optimal map in (5), guaranteed to exist by Theorem 24, we will use $x_t = (1-t)x_0 + tG(x_0)$ to denote a point in the support of μ_t. Lemma 26 then yields uniqueness of the geodesic μ_t, and we adopt similar notation to Theorem 20 and Lemma 26: v_i is the Kantorovich potential for transport between μ_i and ν, and $v_t = (1-t)v_0 + tv_i$ plays the same role between μ_t and ν. Letting T_t be the optimal transport map sending μ_t to ν, set $JT_t^y(x_t) := \frac{\sqrt{\det(Df^T Df(y))}}{\det[-D^2 v_t(y) - x_t \cdot D^2 f(y)]}$, and $\overline{\mu}_t^y(x_t) = \overline{\mu}_t(x_t)/(JT_t^y(x_t)\overline{\nu}(y))$

the density with respect to $(m-n)$-dimensional Hausdorff measure of the conditional probability μ_t^y after disintegrating μ_t with respect to ν, using the map T_t; that is, $\mu_t(x_t) := \mu_t^y(x_t) \otimes \nu(y)$.

We recall that Lemma 26 imlies that μ_t^y is the displacement interpolant between μ_0 and μ_1 and that the Brenier map between μ_0^y and μ_t^y is then given by $\nabla\varphi_t^y = (1-t)I + t\nabla\varphi^y$, where $\nabla\varphi^y : T_0^{-1}(y) \to T_1^{-1}(y)$ is the Brenier map between μ_0^y and μ_1^y.

Lemma 28. *Assume that the reference measure ν is absolutely continuous with respect to n-dimensional Hausdorff measure on a smooth n-dimensional submanifold Y and that μ_0 and μ_1 are absolutely continuous with respect to Lebesgue measure on X. The internal energy at the interpolant is given by*

$$\int_X U(\overline{\mu}_t(x))dx =$$
$$\int_Y \frac{1}{\overline{\nu}(y)} \int_{T_0^{-1}(y)} U\Big(\frac{\overline{\mu}_0^y(x_0)\overline{\nu}(y)JT_t^y(\nabla\varphi_t^y(x_0))}{\det(D^2\varphi_t^y(x_0))}\Big) \frac{1}{JT_t^y(\nabla\varphi_t^y(x_0))} \det(D^2\varphi_t^y(x_0))d\mathcal{H}^{m-n}(x_0)dy$$
$$(29)$$

Proof. Similarly to (18), we can write the internal energy as

$$\int_X U(\overline{\mu}_t(x_t))dx_t = \int_Y \frac{1}{\overline{\nu}(y)} \int_{T_t^{-1}(y)\cap X(v_t^c)} U(\overline{\mu}_t^y(x_t)\overline{\nu}(y)JT_t^y(x_t)) \frac{1}{JT_t^y(x_t)} d\mathcal{H}^{m-n}(x_t)dy.$$

The remainder of the proof is then a matter of rewriting the inner integral as an integral over $T_0^{-1}(y)$, which is standard. For fixed y, standard optimal transport theory applied to the parallel affine subspaces $X_1^c(y, Dv_i(y))$ implies that the optimal map $\nabla\varphi^y$ is differentiable $\mathcal{H}^{m-n}$ almost everywhere and the Monge-Ampere equation is satisfied for $\mathcal{H}^{m-n}$ almost every x_0:

$$\overline{\mu}_0^y(x_0) = \overline{\mu}_1^y(\nabla\varphi^y(x_0))\det D^2\varphi^y(x_0)$$

and similarly note that since $\nabla\varphi_t^y = (1-t)I + t\nabla\varphi^y : T_0^{-1}(y) \to T_t^{-1}(y)$ pushes μ_0^y forward to μ_y^t and is optimal, we have

$$\overline{\mu}_0^y(x_0) = \overline{\mu}_t^y(\nabla\varphi_t^y(x_0))\det D^2\varphi_t^y(x_0) = \overline{\mu}_t^y(\nabla\varphi_t^y(x_0))\det((1-t)I_{m-n}+tD^2\varphi^y(x_0)).$$

The result then follows by using the standard optimal transport change of variables formula on $X_1(y, Dv_0(y))$, using the change of variables $x_t = \nabla\varphi_t^y(x_0) = (1-t)x_0 + t\nabla\varphi^y(x_0)$ $\qquad\square$

Remark 29. *It is natural to note that the factor* $\det\Big(D^2\varphi_t^y(x_0)\Big)/JT_t^y(\nabla\varphi_t^y(x_0))$ *appearing in (29) is the determinant of the Jacobian matrix $DG_t(x_0)$ for the optimal mapping G_t in Theorem 24 between μ_0 and μ_t. It may be helpful to note that the block diagonal matrix M_t introduced in the proof of the following result is* not *equal to $DG_t(x_0)$. $DG_t(x_0)$ is in fact block upper triangular, with*

the same diagonal blocks are M_t. The key point in the proof is the observation that $DG_t(x_0)$ and M_t have the same determinant; it is then more convenient to think of $\det\left(D^2\varphi_t^y(x_0)\right)/JT_t^y(\nabla\varphi_t^y(x_0))$ as the determinant of M_t, since this matrix is positive definite and we can apply Minkowski's determinant inequality to it.

Corollary 30. *Under the assumption in the preceding Lemma, the internal energy (28) is geodesically convex assuming $U(0) = 0$ and $r \mapsto r^m U(r^{-m})$ is convex and non-increasing.*

Proof. Note that $\det\left(D^2\varphi_t^y(x_0)\right)/JT_t^y(\nabla\varphi_t^y(x_0))$ is in fact the determinant of the $m \times m$ block diagonal matrix M_t with an upper $(m-n) \times (m-n)$ block $D^2\varphi_t^y(x_0) = (1-t)I_{m-n} + tD^2\varphi^y(x_0)$ and a lower $n \times n$ block $\frac{1}{|\det(Df^T Df(y))|^{1/2n}}[-D^2v_t(y) - x_t \cdot D^2 f(y)]$ whose determinant is $\frac{1}{JT_t^y(\nabla\varphi_t^y(x_0))} = \frac{1}{|\det(Df^T Df(y))|^{1/2}}\det[-D^2v_t(y) - x_t \cdot D^2 f(y)]$. Noting that $x_t = \nabla\varphi_t^y(x_0)$ interpolates linearly between x_0 and $x_1 = \nabla\varphi_t^y(x_0)$, and the potential $v_t = (1-t)v_0 + tv_1$ for optimal transport between μ_t and ν interpolates linearly between v_0 and v_1, M_t is affine in t.

Now, the first block $D^2\varphi_t^y(x_0)$ of M_t is clearly non-negative definite by convexity of φ^y. Note that since v_t is the Kantorovich potential for transport between μ_t and ν, we have for $y = T_t(x_t)$, $D^2 v_t(y) \leq D_{yy}^2 c(x_t, y) = x_t \cdot D^2 f(y)$; therefore, the second block $\frac{1}{(\det(Df^T Df(y)))^{1/2n}}[-D^2 v_t(y) - x_t \cdot D^2 f(y)]$ of M_t is also non-negative definite. It follows that M_t is as well, and Minkowski's determinant inequality implies that $(\det M_t)^{1/m}$ is concave as a function of t; the convexity of the integrand now follows by a standard argument. Indeed, the integrand is the composition of the concave mapping $t \mapsto [\det((1-t)I_{m-n} + tD^2\varphi^y(x_0))/JT_t^y(\nabla\varphi_t^y(x_0))]^{1/m} = (\det[M_t])^{1/m}$ and the convex, non-increasing mapping $r \mapsto U(\frac{\overline{\mu}_0^y(x_0)\overline{\nu}(y)}{r^m})r^m$ and therefore convex. $\qquad\square$

Remark 31. *An important special case occurs when the function f is affine, in which case $c(x, y) = -x \cdot f(y)$ becomes an index cost. As shown in Proposition 16, when $Y \subseteq \mathbb{R}$ is one dimensional, the metric W_ν becomes the layerwise Wasserstein metric from [12], and Corollary 30 generalizes Corollary 4.2 there (which asserts convexity of the Shannon entropy along layerwise-Wasserstein geodesics).*

Remark 32. *In the case where f is affine, so that $Df(y)$ is constant, slightly improved convexity results can be obtained when we restrict our interpolations to the class $\mathcal{P}_{\nu,v}(X)$ of probability measures on X whose Kantorovich potential corresponding to optimal transport to ν is a fixed v (up to addition of a constant). (It is worth noting that in our context, although the potential v does not uniquely determine the measure μ, which would make $\mathcal{P}_{\nu,v}(X)$ a singleton, it is easy to see that $\mathcal{P}_{\nu,v}(X)$ is geodesically convex).*

In this case, since $v_t = v$ is constant, $X_1(y, Dv(y)) = \{x : x \cdot Df(y) = Dv(y)\}$ on which μ_t^y is supported is constant as well. The curve $\nabla\varphi_t^y(x_0) =

$(1-t)x_0 + t\nabla\varphi^y(x_0)$ *then moves along* $X_1(y, Dv(y))$, *so that* $JT_t^y(\nabla\varphi_t^y(x_0))$ *is constant as a function of* t, *and so we can use concavity of* $\det^{m-n}$ *on the set of positive definite* $(m-n) \times (m-n)$ *dimensional matrices to obtain convexity of* (28) *as soon at* $r \mapsto r^{m-n}U(r^{n-m})$ *is convex and nonincreasing, a slight improvement over the classical condition.*

The set $\mathcal{P}_{\nu,v}(X)$ *is potentially relevant in the root comparison applications in [12]. In these applications, the main objective is to interpolate between the shapes, and as part of the analysis they are first rescaled so that they have uniform marginals in one (the vertical) direction; that is, one compares measures* $\tilde{\mu}_0$ *and* $\tilde{\mu}_1$ *on* $[0,1] \times \mathbb{R}^{m-1}$, *whose first marginals are Lebesgue. In this case, if we set* ν *to be Lebesgue measure on the line segment* $[0,1] \times \{0\}$, *then optimal transport between* μ_i *and* ν *is given by the projection map* $x \mapsto x^1$ *and the Kantorovich potential on the* y *side are* $v_i(y) = y^2/2$ *for both* $i = 0, 1$.

Therefore, interpolating between these rescaled versions corresponds to working in $\mathcal{P}_{\nu,v}(X)$, *where* ν *is uniform measure on* $\{0\} \times [0,1]$ *and* $v(y) = y^2/2$, *and the slightly improved convexity results described above apply.*

5.2 Geodesic convexity of the set of nested sources

Assuming the ν concentrates on a smooth n-dimensional submanifold Y, we now turn our attention to the structure of the set of measures $\mu \in \mathcal{P}_\mu^u(X)$ such that optimal transport with quadratic cost between μ and ν satisfies (a strengthening of) the generalized nested condition. We will show that this set is W_ν convex.

We adopt the convention in Remark 14; the cost is $c(x, y) = -x \cdot f(y)$, where $f : Y \to \mathbb{R}^m$ for some $Y \subset \mathbb{R}^n$. Note that this cost naturally extends to all $x \in \mathbb{R}^m$ using the same formula. Our strengthening of the nestedness condition requries some additional notation. For each $y \in Y$, decompose $x \in \mathbb{R}^m$ into $x = x_{||}^y + x_\perp^y$ with $x_\perp^y \cdot Df(y) = 0$ and $x_{||}^y$ orthogonal to the null space of $Df(y)$.

We define the *enhancement of X by* $y \in Y$ as

$$X^y = \{x = x_\perp^y + x_{||}^y : x_\perp^y + \tilde{x}_{||}^y, \tilde{x}_\perp^y + x_{||}^y \in X \text{ for some } \tilde{x}_{||}^y \in [null(Df(y))]^\perp, \tilde{x}_\perp^y \in null(Df(y))\}$$

We say that (c, μ, ν) satisfies the *enhanced generalized nestedness condition* at y if

$$X_1^y(y, Dv(y)) = \partial^c v(y) := \{x \in X^y : -x \cdot f(y) - v(y) \le -x \cdot f(\tilde{y}) - v(\tilde{y}) \; \forall \tilde{y} \in Y\},$$

where v is the Kantorovich potential. We say that (c, μ, ν) satisfies the *enhanced generalized nestedness condition* if it satisfies this condition at almost every $y \in Y$. Since $X \subseteq X^y$ for each y, $X_1(y, Dv(y)) \subseteq X_1^y(y, Dv(y))$, and so the enhanced generalized nestedness condition implies the generalized nestedness condition $X_1(y, Dv(y)) = \partial^c v(y)$ introduced in [16].

Proposition 33. *Let ν be absolutely continuous with respect to n-dimensional Hausdorff measure on the n-dimensional submanifold Y. Then the set of μ such that (c, μ, ν) is enhanced generalized nested at y is geodesically convex with respect to the metric W_ν.*

Proof. Given Kantorovich potentials v_0 and v_1 corresponding to μ_0 and μ_1, the potential $v_t := (1-t)v_0 + tv_1$ corresponds to the interpolant μ_t. Observe that, since $D_y c(x, y) = -x \cdot Df(y)$ is affine in x, the enhanced level sets $X_1^y(y, Dv_t(y)) = (1-t)X_1^y(y, Dv_0(y)) + tX_1^y(y, Dv_1(y))$ is the Minkowski sum of the two endpoint enhanced level sets. Letting $x_t = (1-t)x_0 + tx_1 \in X_1^y(y, Dv_t(y))$, with each $x_i \in X_1^y(y, Dv_i(y))$, we have by assumption that

$$-x_i \cdot f(y) - v_i(y) \leq x_i \cdot f(\tilde{y}) - v_i(\tilde{y})$$

for all $\tilde{y} \in Y$. Taking the weighted average of these two inequalities yields

$$
\begin{aligned}
-(1-t)x_0 \cdot f(y) - (1-t)v_0(y) - tx_1 \cdot f(y) - tv_1(y) &\leq -(1-t)x_0 \cdot f(\tilde{y}) - (1-t)v_i(\tilde{y}) \\
&\quad -tx_1 \cdot f(\tilde{y}) - tv_1(\tilde{y}) \\
-x_t \cdot f(y) - v_t(y) &\leq x_t \cdot f(\tilde{y}) - v_t(\tilde{y})
\end{aligned}
$$

for all $\tilde{y} \in Y$. Thus, $x_t \in \partial^c v_t(y)$ as desired. $\qquad\square$

6 Fixed point characterization of solutions to variational problems

In this section, we consider a fixed measure μ on the bounded domain $X \subseteq \mathbb{R}^m$, absolutely continuous with respect to Lebesgue measure with density $\overline{\mu}$, bounded above and below, and a fixed domain $Y := (\underline{y}, \overline{y}) \subset \mathbb{R}$. We are interested in the minimization problem (22). We recall that this problem belongs to a class arising in certain game theory problems, introduced in a series of papers by Blanchet-Carlier [2–4], in which the fixed measure μ represents a distribution of players and Y a space of strategies; minimizing measures ν in (22) represent equilibrium distributions of strategies. In general, the dimensions of m and n of the spaces of players and strategies represent the number of charateristics used to distinguish between them. The case $m > n$ (including the case $n = 1$ treated here) are of particular interest, since it is often the case that players are highly homogeneous, whereas the strategies available to them are less varied.

When $m = 1$ (so that both players and strategy spaces are one dimensional), Blanchet-Carlier introduced a characterization of minimizers as fixed points of a certain mapping. They then iterated that mapping as a way to compute solutions [3]; however, they did not prove convergence of this scheme.

Here we introduce a similar fixed point characterization for $m \geq 1$. Furthermore, we prove for any m, under certain assumptions, that the mapping

is a contraction with respect to a strategically chosen metric, implying convergence of the iterative scheme; even for $m = 1$, this is a new result. The key to the proof is choosing a suitable metric; our choice is a slight variant of the dual metric $W^*_{\mu,\mathcal{L},c,1}$ introduced in Section 2.3.

Choose $\underline{d} = \min_{x \in \overline{X}, y \in \overline{Y}} \frac{\partial c}{\partial y}(x, y), \overline{d} = \max_{x \in \overline{X}, y \in \overline{Y}} \frac{\partial c}{\partial y}(x, y)$.

Let $k \in L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right)$, the set of L^1 functions on $(\underline{y}, \overline{y})$ taking values in $[\underline{d}, \overline{d}]$, and set $v_k(y) := \int_{\underline{y}}^y k(s) ds$. Note that $L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right)$ is a complete metric space. Let

$$\overline{\nu}_k(y) := \frac{e^{-v_k(y)}}{\int_{\underline{y}}^{\overline{y}} e^{-v_k(s)} ds}. \tag{30}$$

This is a probability density on $(\underline{y}, \overline{y})$; we denote the corresponding measure by $\nu_k(y)$. Note that (30) then defines a mapping, $L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right) \to \mathcal{P}^{ac}((\underline{y}, \overline{y}))$, where $\mathcal{P}^{ac}((\underline{y}, \overline{y}))$ is the set of absolutely continuous probability measures on $(\underline{y}, \overline{y})$:

$$k \mapsto \nu_k \tag{31}$$

We note that, for any $k \in L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right)$, we have

$$\begin{aligned}
\nu_k(\underline{y}, y) &= \int_{\underline{y}}^y \overline{\nu}_k(s) ds \\
&\geq \int_{\underline{y}}^y \frac{\underline{d} e^{-\overline{d}(s - \underline{y})}}{1 - e^{-\underline{d}(\overline{y} - \underline{y})}} ds \\
&= \frac{\underline{d}(1 - e^{-\overline{d}(y - \underline{y})})}{\overline{d}(1 - e^{-\underline{d}(\overline{y} - \underline{y})})}.
\end{aligned} \tag{32}$$

Similarly,

$$\nu_k(y, \overline{y}) \geq \frac{\underline{d}(e^{-\overline{d}(y - \underline{y})} - e^{-\overline{d}(\overline{y} - \underline{y})})}{\overline{d}(1 - e^{-\underline{d}(\overline{y} - \underline{y})})}. \tag{33}$$

We then note that, for a given $\nu \in \mathcal{P}^{ac}((\underline{y}, \overline{y}))$ the mass splitting property (21) selects a unique $k = k(y) \in [\underline{d}, \overline{d}]$ for each $y \in (\underline{y}, \overline{y})$. The resulting function $k(y)$ is then bounded and so clearly in $L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right)$. We can then think of this process as defining a mapping $\mathcal{P}^{ac}((\underline{y}, \overline{y})) \to L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right), \nu \mapsto k(\cdot)$. Composing this mapping with (31) then yields a $F : L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right) \to L^1\left((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\right)$. Explicitly, defining $\tilde{k} := F[k]$, we have

$$\nu_k(\underline{y}, y) := \mu(X_{\geq}(y, \tilde{k}(y))). \tag{34}$$

Lemma 34. *Let ν be a minimizer of* (22) *such that (c, μ, ν) is nested. Then $k(y) = v'(y)$ is a fixed point of F, where v is the corresponding Kantorovich potential.*

Proof. The first order conditions corresponding to the variational problem are

$$v(y) + \log(\overline{\nu}(y)) = C$$

for some constant C, ν almost everywhere (see, for example, Section 7.4 in [19]), where $\overline{\nu}$ is the density of minimizer $\overline{\nu}$ and v the corresponding Kantorovich potential. This implies that that $\nu = \nu_k$.

Since (c, μ, ν) is nested, the $\tilde{k} = F[K]$ defined by (34) coincides with the derivative of the Kantorovich potential for optimal transport between μ and $\nu_k = \nu$, which is exactly k. $\qquad\square$

We will prove that, under certain conditions, F is a contraction. The Banach fixed point theorem will then imply that F has a unique fixed point. Under conditions ensuring nestedness of the minimizer, Lemma 34, will then ensure that that fixed point is the minimizer of (22).

We need the following conditions on X, c and μ to ensure F is a contraction:

- $A := \max |D_x c_y(x, y)| < \infty$,

- $B := \min \overline{\mu}(x) > 0$,

- There is a $C > 0$ such that, for all $y \in Y$, $x \in X$, using $c_y(x, y)$ as a shorthand for $\frac{\partial c}{\partial y}(x, y)$,

$$\mathcal{H}^{m-1}(X_1^c(y, c_y(x, y))) \geq C \min \left\{ \mu(X_{\geq}^c(y, c_y(x, y))), \mu(X_{\leq}^c(y, c_y(x, y))) \right\}.$$
(35)

Let us discuss briefly the third condition above. For any x, y, setting $p = c_y(x, y)$, the level set $X_1^c(y, p)$ is an $m - 1$ dimensional submanifold, dividing the region X into the sub and super level sets $X_{\geq}^c(y, p)$ and $X_{\leq}^c(y, p)$. The condition implies that if $X_1^c(y, p)$ is small, then at least one of $\mu(X_{\geq}^c(y, p))$ or $\mu(X_{\leq}^c(y, p))$ must also be (quantitatively) small.

In addition, we set $H(y) = \max\{ \frac{|e^{(\overline{d} - 2\underline{d})(y - \underline{y})} - 1|}{|1 - e^{-\overline{d}(y - \underline{y})}|}, \frac{|e^{(\overline{d} - 2\underline{d})(\overline{y} - \underline{y})} - e^{(\overline{d} - 2\underline{d})(y - \underline{y})}|}{|(e^{-\overline{d}(y - \underline{y})} - e^{-\overline{d}(\overline{y} - \underline{y})}|} \}$.

Theorem 35. *Assume the three conditions above, and that*

$$\frac{2A\overline{d}^3(1 - e^{-\underline{d}(\overline{y} - \underline{y})})^2}{BC\underline{d}^2[1 - e^{-\overline{d}(\overline{y} - \underline{y})}]^2|\overline{d} - 2\underline{d}|} \int_{\underline{y}}^{\overline{y}} H(y)dy < 1.$$

Then F is a contraction.

Remark 36. *The condition appearing in the statement of the theorem looks complicated; however, all quantities except for B and C can be computed using the cost function alone (B involves the reference marginal μ while C involves both the cost and μ). Since the factor in front of the integral and the function $H(y)$ are bounded, the limit as $\overline{y} - \underline{y} \to 0$ of the left hand side above is 0; therefore, the condition will always hold for sufficiently small intervals. We illustrate these points in a simple example below, where we compute each of these quantites explicly.*

Example 37. *Let $X = [-1,1]^2 \subset \mathbb{R}^2$, $Y = [0,\overline{y}] \subset \mathbb{R}$, and $c(x,y) = -x_1 y$. We take μ to be uniform measure, normalized to have total mass 1, so that the density is $\overline{\mu}(x) = 1/4$; clearly this means $B = 1/4$. We then have $c_y(x,y) = -x_1$ so that $\overline{d} = 1$, $\underline{d} = -1$, and $D_x c_y(x,y) = -(1,0)$, so that $A = 1$. We also have that each $X_1^c(y,p) = \{x : x_1 = -p, \ -1 < x_2 < 1\}$ is a vertical line and so $\mathcal{H}^{m-1}(X_1^c(y, c_y(x,y))) = \mathcal{H}^1(\{x : x_1 = -p, \ -1 < x_2 < 1\}) = 2$. Since $\mu(X_{\geq}^c(y, c_y(x,y))) + \mu(X_{\leq}^c(y, c_y(x,y)) = 1$, the minimum of the two values is at most $1/2$, so we can take $C = 4$.*

It is easy to check that $\frac{|e^{(\overline{d}-2\underline{d})(y-\underline{y})}-1|}{|1-e^{-\overline{d}(y-\underline{y})}|} = \frac{e^{3y}-1}{1-e^{-y}}$ is increasing and so is bounded above by its value $\frac{e^{3\overline{y}}-1}{1-e^{-\overline{y}}}$ at the endpoint $\overline{y}$. Similarly, $\frac{|e^{(\overline{d}-2\underline{d})(\overline{y}-\underline{y})}-e^{(\overline{d}-2\underline{d})(y-\underline{y})}|}{|(e^{-\overline{d}(y-\underline{y})}-e^{-\overline{d}(\overline{y}-\underline{y})}|} = \frac{e^{3\overline{y}}-e^{3y}}{e^{-y}-e^{-\overline{y}}}$ is decreasing provided $\overline{y}$ is small enough, so it is bounded above by its value $\frac{e^{3\overline{y}}-1}{1-e^{-\overline{y}}}$ at $y = 0$. Therefore, $H(y) \leq \frac{e^{3\overline{y}}-1}{1-e^{-\overline{y}}}$ and we have, for sufficiently small $\overline{y}$,

$$\frac{2A\overline{d}^3(1-e^{-\underline{d}(\overline{y}-\underline{y})})^2}{BC\underline{d}^2[1-e^{-\overline{d}(\overline{y}-\underline{y})}]^2|\overline{d}-2\underline{d}|} \int_{\underline{y}}^{\overline{y}} H(y)dy = \frac{2(1-e^{\overline{y}})^2}{\frac{1}{4}4(1-e^{-\overline{y}})^2 3} \int_0^{\overline{y}} H(y)dy$$

$$\leq \frac{2(1-e^{\overline{y}})^2}{3(1-e^{-\overline{y}})^2}\overline{y}\frac{e^{3\overline{y}}-1}{1-e^{-\overline{y}}}$$

This tends to 0 as $\overline{y}$ does, since $\frac{(1-e^{\overline{y}})^2}{(1-e^{-\overline{y}})^2} \to 1$ and $\frac{e^{3\overline{y}}-1}{1-e^{-\overline{y}}} \to 3$, so that for sufficiently small $\overline{y}$, the condition in Theorem 35 holds.

Remark 38 (Relationship to dual metrics)**.** *We show that F is a contraction with respect to the L^1 metric. Noting that the mapping $\nu \mapsto k$, where k is defined by the mass splitting condition $\nu(\underline{y}, y) = \mu(X_{\geq}(y, k(y)))$ is a bijection on the set of non-atomic measures $\nu \in \overline{\mathcal{P}}(Y)$ (and in particular on the set $\mathcal{P}^{ac}(Y)$ of measures which are absolutely continuous with respect to Lebesgue measure $\mathcal{L}$), we can consider the L^1 metric to be a metric on $\mathcal{P}^{ac}(Y)$. If (c, μ, ν) is nested, then $k(y) = v'(y)$ where v is the Kantorovich dual potential. In this case, this metric is exactly the dual metric $W^*_{\mu,\mathcal{L},c,1}$ from Section 2.3.*

Proof. Let $k_0, k_1 \in L^1\Big((\underline{y}, \overline{y}); [\underline{d}, \overline{d}]\Big)$. Then note that, setting $v_i(y) := v_{k_i}(y)$,

$$
\begin{aligned}
|\overline{\nu}_0(y) - \overline{\nu}_1(y)| &= \frac{e^{-v_0(y)}}{\int_{\underline{y}}^{\overline{y}} e^{-v_0(s)} ds} - \frac{e^{-v_1(y)}}{\int_{\underline{y}}^{\overline{y}} e^{-v_1(s)} ds} \\
&= \left| \frac{1}{\int_{\underline{y}}^{\overline{y}} e^{v_0(y) - v_0(s)} ds} - \frac{1}{\int_{\underline{y}}^{\overline{y}} e^{v_1(y) - v_1(s)} ds} \right| \\
&= \frac{|\int_{\underline{y}}^{\overline{y}} [e^{v_1(y) - v_1(s)} - e^{v_0(y) - v_0(s)}] ds|}{\int_{\underline{y}}^{\overline{y}} e^{v_0(y) - v_0(s)} ds \int_{\underline{y}}^{\overline{y}} e^{v_1(y) - v_1(s)} ds} \\
&\leq \frac{\int_{\underline{y}}^{\overline{y}} |e^{v_1(y) - v_1(s)} - e^{v_0(y) - v_0(s)}| ds}{[e^{\underline{d}(y - \underline{y})} \int_{\underline{y}}^{\overline{y}} e^{-\overline{d}(s - \underline{y})} ds]^2} \\
&\leq \frac{\overline{d}^2 \int_{\underline{y}}^{\overline{y}} e^{\overline{d}(y - \overline{y}) - \underline{d}(s - \overline{y})} [|v_0(y) - v_1(y)| + |v_0(s) - v_1(s)|] ds}{[e^{\underline{d}(y - \underline{y})}(1 - e^{-\overline{d}(\overline{y} - \underline{y})})]^2} \\
&\leq \frac{2\overline{d}^2 \|v_0 - v_1\|_{L^\infty} e^{\overline{d}(y - \overline{y})} \int_{\underline{y}}^{\overline{y}} e^{-\underline{d}(s - \overline{y})} ds}{[e^{\underline{d}(y - \underline{y})}(1 - e^{-\overline{d}(\overline{y} - \underline{y})})]^2} \\
&\leq \frac{2\overline{d}^2 \|v_0 - v_1\|_{L^\infty} e^{(\overline{d} - 2\underline{d})(y - \overline{y})}(1 - e^{-\underline{d}(\overline{y} - \underline{y})})}{\underline{d}[1 - e^{-\overline{d}(\overline{y} - \underline{y})}]^2} \\
&\leq \frac{2\overline{d}^2 \|k_0 - k_1\|_{L^1} e^{(\overline{d} - 2\underline{d})(y - \overline{y})}(1 - e^{-\underline{d}(\overline{y} - \underline{y})})}{\underline{d}[1 - e^{-\overline{d}(\overline{y} - \underline{y})}]^2}.
\end{aligned}
\tag{36}
$$

Now, from the mass splitting definition of $\tilde{k}$, assuming without loss of generality that $\tilde{k}_0(y) \geq \tilde{k}_1(y)$, we compute, using the co-area formula

$$
\int_{\underline{y}}^{y} [\overline{\nu}_1(s) - \overline{\nu}_0(s)] ds = \mu\{x : \tilde{k}_0(y) \geq c_y(x, y) \geq \tilde{k}_1(y)\}
\tag{37}
$$

$$
= \int_{\{x : \tilde{k}_0(y) \geq c_y(x,y) \geq \tilde{k}_1(y)\}} \overline{\mu}(x) dx
\tag{38}
$$

$$
= \int_{\tilde{k}_1(y)}^{\tilde{k}_0(y)} \Big(\int_{X_=(y,k)} \frac{1}{|D_x c_y(x, y)|} \overline{\mu}(x) d\mathcal{H}^{m-1}(x) \Big) dk
\tag{39}
$$

$$
\geq |\tilde{k}_0(y) - \tilde{k}_1(y)| \frac{1}{\max |D_x c_y(x, y)|} [\min \overline{\mu}(x)] [\min_{\tilde{k}_1(y) \leq k \leq \tilde{k}_0(y)} \mathcal{H}^{m-1}(X_=(y, k))].
$$

Now, suppose the minimum of $\mathcal{H}^{m-1}(X_=(y, k))$ over $k \in [\tilde{k}_1(y), \tilde{k}_0(y)]$ is attained at some k; by assumption (35), either

$$
\mathcal{H}^{m-1}(X_=(y, k)) \geq C\mu(X_\leq(y, k))
\tag{40}
$$

or

$$
\mathcal{H}^{m-1}(X_=(y, k)) \geq C\mu(X_\geq(y, k)).
\tag{41}
$$

34

We assume the second case (the proof in the first case is similar), in which case, we have

$$\mathcal{H}^{m-1}(X_=(y,k)) \geq C\mu(X_\geq(y,k)) \geq C\mu(X_\geq(y,\tilde{k}_1(y))).$$

Therefore,

$$
\begin{aligned}
|\tilde{k}_1(y) - \tilde{k}_0(y)| &\leq \frac{2A\overline{d}^2||k_0 - k_1||_{L^1}(1 - e^{-\underline{d}(\overline{y}-\underline{y})})}{BC\underline{d}[1 - e^{-\overline{d}(\overline{y}-\underline{y})}]^2\mu[X_\geq(y,\tilde{k}_1(y))]} \int_{\underline{y}}^y e^{(\overline{d}-2\underline{d})(s-\overline{y})}ds \\
&= \frac{2A\overline{d}^2||k_0 - k_1||_{L^1}(1 - e^{-\underline{d}(\overline{y}-\underline{y})})}{BC\underline{d}[1 - e^{-\overline{d}(\overline{y}-\underline{y})}]^2\mu[X_\geq(y,\tilde{k}_1(y))]} \frac{(e^{(\overline{d}-2\underline{d})(y-\overline{y})} - 1)}{\overline{d} - 2\underline{d}} \\
&= \frac{2A\overline{d}^2||k_0 - k_1||_{L^1}(1 - e^{-\underline{d}(\overline{y}-\underline{y})})}{BC\underline{d}[1 - e^{-\overline{d}(\overline{y}-\underline{y})}]^2\nu_{k_1}(\underline{y},y)} \frac{(e^{(\overline{d}-2\underline{d})(y-\overline{y})} - 1)}{\overline{d} - 2\underline{d}} \\
&\leq \frac{2A\overline{d}^3||k_0 - k_1||_{L^1}(1 - e^{-\underline{d}(\overline{y}-\underline{y})})}{BC\underline{d}^2[1 - e^{-\overline{d}(\overline{y}-\underline{y})}]^2} \frac{(e^{(\overline{d}-2\underline{d})(y-\overline{y})} - 1)}{\overline{d} - 2\underline{d}} \frac{(1 - e^{-\underline{d}(\overline{y}-y)})}{(1 - e^{-\overline{d}(y-\underline{y})})}.
\end{aligned}
\tag{42}
$$

where we have used (32) in the last line. If instead of (41) we have (40), a very similar argument holds, in which $\mu[X_\geq(y,\tilde{k}_1(y))]$ and $\nu_{k_1}(\underline{y},y)$ are replaced by $\mu[X_\leq(y,\tilde{k}_0(y))]$ and $\nu_{k_0}(y,\overline{y})$, the domain of integration in (37) is replaced by $[y,\overline{y}]$, and we use (33) instead of (32). In this case, we obtain

$$|\tilde{k}_1(y) - \tilde{k}_0(y)| \leq \frac{2A\overline{d}^3||k_0 - k_1||_{L^1}(1 - e^{-\underline{d}(\overline{y}-\underline{y})^2})}{BC\underline{d}^2[1 - e^{-\overline{d}(\overline{y}-\underline{y})}]^2} \frac{(e^{(\overline{d}-2\underline{d})(\overline{y}-\underline{y})} - e^{(\overline{d}-2\underline{d})(y-\underline{y})})}{\overline{d} - 2\underline{d}} \frac{1}{(e^{-\overline{d}(y-\underline{y})} - e^{-\overline{d}(\overline{y}-\underline{y})})}.$$

Therefore, for every $y \in (\underline{y},\overline{y})$, we have

$$|\tilde{k}_1(y) - \tilde{k}_0(y)| \leq \frac{2A\overline{d}^3||k_0 - k_1||_{L^1}[1 - e^{-\underline{d}(\overline{y}-\underline{y})}]^2}{BC\underline{d}^2[1 - e^{-\overline{d}(\overline{y}-\underline{y})}]^2|\overline{d} - 2\underline{d}|}H(y).$$

Integrating yields:

$$||\tilde{k}_1 - \tilde{k}_0||_{L^1} \leq \left(\frac{2A\overline{d}^3[1 - e^{-\underline{d}(\overline{y}-\underline{y})}]^2}{BC\underline{d}^2|\overline{d} - 2\underline{d}|[1 - e^{-\overline{d}(\overline{y}-\underline{y})}]^2} \int_{\underline{y}}^{\overline{y}} H(y)dy\right)||k_0 - k_1||_{L^1}$$

which is the desired result. $\qquad\square$

As discussed above, together with the Lemma 34, condition (23) ensuring nestedness of the solution, and the Banach Fixed Point Theorem, this yields convergence of a simple iterative algorithm to compute the solution.

Corollary 39. *Under the assumptions in Theorem 35, and as well as assumption (23), for any $k \in L^1\left((\underline{y},\overline{y}); [\underline{d},\overline{d}]\right)$, the sequence $F^n[k]$ converges to a function k_{fixed} whose anti-derivative is the Kantorovich potential corresponding to the unique minimizer ν of (22).*

A   Hierarchical metrics with multiple reference measures

The ν-based Wasserstein metric can be extended to align with several different probability measures in a hierarchical way. Given $\nu_1, \ldots, \nu_k \in \mathcal{P}(X)$, we define iteratively $\Pi^1_{opt} = \Pi_{opt}(\nu_1, \mu_i)$ and $\Pi^j_{opt}(\nu_1, \ldots, \nu_j, \mu_i) \subseteq \mathcal{P}(X^{j+1})$ by

$$\Pi^j_{opt}(\nu_1, ..., \nu_j, \mu_i) = \text{argmin}_{\substack{\gamma_{y_1, \ldots y_{j-1} x_i} \in \Pi^{j-1}_{opt}(\nu_1, \ldots, \nu_{j-1}, \mu_i) \\ \gamma_{y_j} = \nu_j}} \int_{X^{j+1}} |x_i - y_j|^2 d\gamma(y_1, \ldots., y_j, x_i).$$

A natural analogue of W_ν is then defined by:

$$W^2_{\nu_1, \ldots \nu_k}(\mu_0, \mu_1) = \inf_{\gamma_{y_1 \ldots y_k x_i} \in \Pi^k_{opt}(\nu_1, \ldots, \nu_k, \mu_i), i=0,1} \int_{X^{k+2}} |x_0 - x_1|^2 d\gamma(y_1, \ldots., y_k, x_0, x_1). \tag{43}$$

As before, optimal couplings can be recovered as limits of multi-marginal problems, where the weights on the interaction terms reflect the hierarchy of the measures $\nu_1, ..., \nu_k$ in the definition of $W_{\nu_1, \ldots \nu_k}$:

Proposition 40. *Let $\nu_1, \ldots., \nu_k, \mu_0, \mu_1$ be probability measures on X. Consider the multi-marginal optimal transport problem:*

$$\int_{X^{k+2}} c_\varepsilon(x_0, x_1, y_1, y_2, \ldots y_k) d\gamma \tag{44}$$

where $c_\varepsilon = \varepsilon^{n-1}|x_0 - x_1|^2 + \varepsilon^{n-2}(|x_0 - y_{n-1}|^2 + |x_1 - y_{n-1}|^2) + \varepsilon^{n-3}(|x_0 - y_{n-2}|^2 + |x_1 - y_{n-2}|^2) + \ldots. + \varepsilon(|x_0 - y_2|^2 + |x_1 - y_2|^2) + (|x_0 - y_1|^2 + |x_1 - y_1|^2)$. Then any weak limit $\overline{\gamma}$ of solutions γ_ε is optimal in (43)

Proof. The proof is similar to the proof of the second part of Theorem 12. Letting $\gamma \in \Pi^k_{opt}(\nu_1, ..., \nu_k, \mu_i)$ for $i = 0, 1$, optimality of γ_ε in (44) implies

$$\int_{X^{k+2}} c_\varepsilon(x_0, x_1, y_1, y_2, \ldots y_k) d\gamma_\varepsilon \leq \int_{X^{k+2}} c_\varepsilon(x_0, x_1, y_1, y_2, \ldots y_k) d\gamma. \tag{45}$$

Passing to the limit implies that $\overline{\gamma} \in \Pi^1_{opt}(\mu_i, \nu)$. Now, combining the fact that $\gamma \in \Pi^1_{opt}(\mu_i, \nu)$ with (45) implies

$$\int_{X^{k+2}} c_\varepsilon(x_0, x_1, y_1, y_2, \ldots y_k) - (|x_0 - y_1|^2 + |x_1 - y_1|^2) d\overline{\gamma}_\varepsilon$$

$$\leq \int_{X^{k+2}} c_\varepsilon(x_0, x_1, y_1, y_2, \ldots y_k) - (|x_0 - y_1|^2 + |x_1 - y_1|^2) d\gamma.$$

Dividing by ε and taking the limit then implies that $\overline{\gamma} \in \Pi_{opt}(\nu_1, \nu_2, \mu_i)$ for $i = 0, 2$. Continuing inductively in this way yields the desired result.   □

As a consequence, we easily obtain the following result, which is similar to the main result in [6].

Corollary 41. *Let $k = n - 1$, and $\nu_1, ... \mu_{n-1}$ be probability measures concentrated on mutually orthogonal line segments, each absolutely continuous with respect to one dimensional Hausdorff measure. Then solutions γ_ε to the multi-marginal optimal transport problem (44) converge weakly to γ, whose $1, 2$ marginal is $(Id, G)_{\#}\mu_0$, where G is the Knothe-Rosenblatt transport from μ_0 to μ_1.*

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics. Birkhäuser Boston, 2nd edition, 2004.

[2] Adrien Blanchet and Guillaume Carlier. From Nash to Cournot-Nash equilibria via the Monge-Kantorovich problem. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 372(2028):20130398, 11, 2014.

[3] Adrien Blanchet and Guillaume Carlier. Remarks on existence and uniqueness of cournot–nash equilibria in the non-potential case. *Mathematics and Financial Economics*, 8(4):417–433, 2014.

[4] Adrien Blanchet and Guillaume Carlier. Optimal transport and Cournot-Nash equilibria. *Math. Oper. Res.*, 41(1):125–145, 2016.

[5] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[6] G. Carlier, A. Galichon, and F. Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM J. Math. Anal.*, 41(6):2554–2576, 2009/10.

[7] Pierre-André Chiappori, Robert McCann, and Brendan Pass. Transition to nestedness in multi- to one-dimensional optimal transport. *European J. Appl. Math.*, 30(6):1220–1228, 2019.

[8] Pierre-Andre Chiappori, Robert J McCann, and Brendan Pass. Multidimensional matching. Preprint at arXiv:1604.05771.

[9] Pierre-André Chiappori, Robert J. McCann, and Brendan Pass. Multi-to one-dimensional optimal transport. *Comm. Pure Appl. Math.*, 70(12):2405–2444, 2017.

[10] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions.* Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015.

[11] Claus Griessler. *C*-cyclical monotonicity as a sufficient criterion for optimality in the multimarginal Monge-Kantorovich problem. *Proc. Amer. Math. Soc.*, 146(11):4735–4740, 2018.

[12] Young-Heon Kim, Brendan Pass, and David J Schneider. Optimal transport and barycenters for dendritic measures. *Pure and Applied Analysis*, 2(3):581–601, 2020.

[13] Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Math. J.*, 4:39–52, 1957.

[14] R. J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[15] Robert J. McCann. Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.*, 11(3):589–608, 2001.

[16] Robert J. McCann and Brendan Pass. Optimal transportation between unequal dimensions. *Arch. Ration. Mech. Anal.*, 238(3):1475–1520, 2020.

[17] Luca Nenna and Brendan Pass. Variational problems involving unequal dimensional optimal transport. *Journal de Mathématiques Pures et Appliquées*, 139:83–108, 2020.

[18] Murray Rosenblatt. Remarks on a multivariate transformation. *Ann. Math. Statistics*, 23:470–472, 1952.

[19] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Progress in Nonlinear Differential Equations and their Applications, 87. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[20] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, 2003.

[21] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

[22] Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comput. Vis.*, 101(2):254–269, 2013.