Multi-Marginal Optimal Transport Defines a Generalized Metric

Liang Mi
Arizona State University
liangmi@asu.edu

José Bento
Boston College
jose.bento@bc.edu

***Abstract*—We prove that the multi-marginal optimal transport (MMOT) problem defines a generalized metric. In addition, we prove that the distance induced by MMOT satisfies a generalized triangle inequality that, to leading order, cannot be improved.**

I. Introduction

The *Optimal Transport* (OT) problem dates back to 1781, when Monge [1] raised the problem of finding a way to transport one distribution of points (formally a probability distribution) into another one at minimal cost. OT theory was greatly developed in the past century, especially assisted by Kantorovich [2] in 1941 and Brenier [3] in 1991, and, in part thanks to contemporary fast OT solvers, e.g. [4], OT has found applications in machine learning, e.g. [5], signal processing, e.g. [6], and information theory, e.g. [7], [8], [9], [10].

Let $(\Omega^i, \mathcal{F}^i, \boldsymbol{p}^i)$ and $(\Omega^j, \mathcal{F}^j, \boldsymbol{p}^j)$ be two probability spaces. Given a cost function $d^{i,j} : \Omega^i \times \Omega^j \to \mathbb{R}^{\geq 0}$, and $\ell \geq 1$, the *Optimal Transport* (OT) problem [2] seeks to find

$$\inf_{\boldsymbol{p}^{i,j}} \left(\int_{\Omega^i \times \Omega^j} (d(\omega^i, \omega^j))^\ell \mathrm{d}\boldsymbol{p}^{i,j}(\omega^i, \omega^j) \right)^{\frac{1}{\ell}}, \quad (1)$$

where the infimum is taken over all measures $\boldsymbol{p}^{i,j}$ on the product space that satisfies $\int_{A \times \Omega^j} \mathrm{d}\boldsymbol{p}^{i,j}(\omega^i, \omega^j) = \boldsymbol{p}^i(A)$ for all $A \in \mathcal{F}^i$, and $\int_{\Omega^i \times A} \mathrm{d}\boldsymbol{p}^{i,j}(\omega^i, \omega^j) = \boldsymbol{p}^j(A)$ for all $A \in \mathcal{F}^j$.

Problem (1) is typically studied under the assumptions that Ω^i and Ω^j are Polish spaces, and that d is a metric. In this case, the minimum cost induced by (1) is called the *Wasserstein distance* (WD), and it is a metric on the space of probability measures. The WD has gained increasing popularity in the past two decades thanks to its superiority over other metrics, and divergences, in many applications, e.g. shape interpolation [11], generative modeling [5], [12], and domain adaptation [13].

More recently, a generalization of OT to multiple marginal measures has gained attention. Given probability spaces $(\Omega^i, \mathcal{F}^i, \boldsymbol{p}^i)$, $i = 1, \dots, n$, a cost function $d : \Omega^1 \times \dots \times \Omega^n \mapsto \mathbb{R}^{\geq 0}$, and $\ell \geq 1$, the Multi-Marginal Optimal Transport (MMOT) problem seeks

$$\inf_{\boldsymbol{p}} \left(\int_{\Omega^1 \times \dots \times \Omega^n} (d(\omega^1, \dots, \omega^n))^\ell \mathrm{d}\boldsymbol{p}(\omega^1, \dots, \omega^n) \right)^{\frac{1}{\ell}}, \quad (2)$$

where the infimum is taken over all measures $\boldsymbol{p}$ on the product space whose ith marginal satisfies, for all $A \in \mathcal{F}^i$,

$$\int_{\Omega^1 \times \dots \times \Omega^{i-1} \times A \times \Omega^{i+1} \times \dots \times \Omega^n} \mathrm{d}\boldsymbol{p}(\omega^1, \dots, \omega^n) = \boldsymbol{p}^i(A). \quad (3)$$

Much of the discussion on MMOT has focused on its existence, the uniqueness and structure of Kantorovich solutions, practical algorithms, and the choice of the cost function, see [14], [15], [16]. Also worth mentioning are [17] which surveys the MMOT problem in the early days, [18] which discusses the pairwise MMOT problem that we focus on, and [19] which computes MMOT for image translation.

There is, however, a lack of discussion about the (generalized) metric properties of MMOT. At the same time, it is the metric property of the WD that makes it useful in so many applications, and thus, understanding when (2) has metric-like properties is very important, not only from a theoretical, but also practical, point of view.

In this paper, we show that a variant of (2) defines a generalized metric, or n-metric [20]. In addition, we show that the generalized triangle inequality that it satisfies cannot be improved (up to leading order of n). To the best of our knowledge, this paper is the first attempt to prove generalized metric properties for MMOT. In the rest of the paper, we provide definitions and notation in Section II, several lemmas in Section III that we will insert for later proofs. we show our main results in Section IV and the detailed proofs supporting them in Section 1 and Section 2, respectively. Finally, we conclude with futher work in Section VII

II. Definitions and notation

For every integer i, we define $[i] \triangleq \{1, \dots, i\}$. To simplify our exposition, we illustrate our theorems for probability spaces where the sample space is finite, the event set σ-algebra is the power set of the sample space, and the probability measure can be fully described using a probability mass function. Our results can be extended to more general settings. We refer to probability mass functions using bold letters, e.g. $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$, etc.

We consider n probability spaces, the ith space being described by a sample space $\Omega^i = \{\Omega^i_1, \dots, \Omega^i_{m^i}\}$, an event space 2^{Ω^i}, and a probability mass function $\boldsymbol{p}^i$, or $\boldsymbol{q}^i$, or $\boldsymbol{r}^i$, etc. Variable m^i specifies the number of atoms of Ω^i. We use $\boldsymbol{p}^i_s$ to denote the probability of the atom Ω^i_s. This notation assumes that the atoms can be given some order, but our results can be extended beyond this assumption. Without loss of generality, we assume that $\Omega^i_s = \Omega^i_t$ if and only if $s = t$.

Given n probability spaces, we use $\boldsymbol{p}^{i_1,\dots,i_k}$ to denote a probability mass function for the probability space with sample space $\Omega^{i_1} \times \dots \times \Omega^{i_k}$ and event space $2^{\Omega^{i_1} \times \dots \times \Omega^{i_k}}$. In particular,

$\boldsymbol{p}^{i_1,\dots,i_k}_{s_1,\dots,s_k}$ is the probability at the atom $(\Omega^{i_1}_{s_1},\dots,\Omega^{i_k}_{s_k})$. Given indices $j_1,\dots,j_r \in [n]$, we also use $\boldsymbol{p}^{i_1,\dots,i_k|j_1,\dots,j_r}$ to denote a probability mass function for the probability space with sample space $\Omega^{i_1}\times\cdots\times\Omega^{i_k}$, and event space $2^{\Omega^{i_1}\times\cdots\times\Omega^{i_k}}$. In particular, for every $(t_1,\dots,t_r)\in[m^{j_1}]\times\cdots\times[m^{j_r}]$, the symbol $\boldsymbol{p}^{i_1,\dots,i_k|j_1,\dots,j_r}_{s_1,\dots,s_k|t_1,\dots,t_r}$ is a probability at atom $(\Omega^{i_1}_{s_1},\dots,\Omega^{i_k}_{s_k})$. This notation is useful to describe conditional probabilities.

Given two probability mass functions $\boldsymbol{p}^i$, $\boldsymbol{p}^j$, for sample spaces Ω^i, Ω^j respectively, we use $\boldsymbol{p}^{i,j}$ to denote a probability mass function for the sample space $\Omega^i\times\Omega^j$, such that

$$(\boldsymbol{p}^{i,j})_{s,t} \triangleq \boldsymbol{p}^{i,j}_{s,t}$$

is the probability of atom (Ω^i_s,Ω^j_t). This notation extends to more than two probability mass functions. Note that $\boldsymbol{p}^{i,j}\neq\boldsymbol{p}^{j,i}$, in terms of the way we index the atoms of both distributions.

Given $k\in[n]$, together with an arbitrary probability mass function $\boldsymbol{q}^k$ for sample space Ω^k and $n-1$ arbitrary probability mass functions $\{\boldsymbol{q}^{i|k}\}_{i\in[n]\setminus\{k\}}$ for sample spaces $\{\Omega^i\}_{i\in[n]\setminus\{k\}}$, we define $\mathcal{G}$ as the map that defines the following probability mass function $\boldsymbol{p}$ on sample space $\Omega^1\times\cdots\times\Omega^n$

$$\begin{aligned}\boldsymbol{p} &= \mathcal{G}\left(\boldsymbol{q}^k,\{\boldsymbol{q}^{i|k}\}_{i\in[n]\setminus\{k\}}\right)\\ &= \left(\prod_{i\in[k-1]}\boldsymbol{q}^{i|k}\right)\boldsymbol{q}^k\left(\prod_{i-k\in[n-k]}\boldsymbol{q}^{i|k}\right). \qquad (4)\end{aligned}$$

To be more specific,

$$\boldsymbol{p}_{s_1,\dots,s_n} = \boldsymbol{q}^k_{s_k}\prod_{i\in[n]\setminus\{k\}}\boldsymbol{q}^{i|k}_{s_i|s_k}. \qquad (5)$$

We use $d^{i,j}_{s,t}$ to denote a distance between Ω^i_s and Ω^j_t. Given sample spaces $\Omega^1,\dots,\Omega^n$, we say that d *is a metric* over these spaces when, for any $i,j,k\in[n]$, and any $s\in[m^i], r\in[m^j], t\in[m^k]$, we have that

i) $d^{i,j}_{r,s}\geq 0$;
ii) $d^{i,j}_{r,s}=d^{j,i}_{s,r}$;
iii) $d^{i,j}_{s,r}=0$ if and only if $\Omega^i_s=\Omega^j_r$;
iv) $d^{i,j}_{s,r}\leq d^{i,k}_{s,t}+d^{k,j}_{t,r}$.

Given two multidimensional arrays of real numbers, A, B, with the same dimensions, and an integer number ℓ, we define

$$\langle A,B\rangle_\ell \triangleq \sum_{s_1,\dots,s_k}(A_{s_1,\dots,s_k})^\ell B_{s_1,\dots,s_k},$$

where $(A_{s_1,\dots,s_k})^\ell$ is the ℓth power of $A_{s_1,\dots,s_k}$. In this paper, unless specified otherwise, any summation is over the full possible range of values for indices in the summation.

When a formula depends on a list of symbols indexed consecutively, we will use : to abbreviate lists. So, e.g., we will write $s_1,\dots,s_k$ as $s_{1:k}$, we will write $\Omega^1,\dots,\Omega^k$ as $\Omega^{1:k}$, and we will write $A_{s_1,\dots,s_k}$ as $A_{s_{1:k}}$. Note that $A_{s_1:s_k}$ has a different meaning than $A_{s_{1:k}}$. Assuming that $s_k>s_1$, the former represents $A_{s_1},A_{s_1+1},A_{s_1+2},\dots,A_{s_k}$.

III. Useful lemmas

Lemma 1. *Let $\boldsymbol{p}$ be defined as in* (4) *and* (5). *Let $\boldsymbol{p}^i$ and $\boldsymbol{p}^{i,k}, i\neq k$, be the marginal probability mass functions for the sample spaces Ω^i and $\Omega^i\times\Omega^k$, respectively, induced by $\boldsymbol{p}$. Let $\boldsymbol{q}^{i,k}=\boldsymbol{q}^{i|k}\boldsymbol{q}^k$, $i\neq k$, and let $\boldsymbol{q}^i$, $i\neq k$, be the marginal probability mass function for the sample space Ω^i induced by $\boldsymbol{q}^{i,k}$. We have that $\boldsymbol{p}^i=\boldsymbol{q}^i\ \forall i$, and $\boldsymbol{p}^{i,k}=\boldsymbol{q}^{i,k}\ \forall i\neq k$.*

Proof. We can think of $\boldsymbol{p}$ as describing n discrete random variables. It follows from (4) that these are independent conditioned on the kth random variable. The result follows. ☐

Lemma 2. *Let d be a metric over $\Omega^{1:n}$, and $\boldsymbol{p}$ be a probability mass function for the sample space $\Omega^1\times\cdots\times\Omega^n$. Let $\boldsymbol{p}^{i,j}$ be the marginal probability mass function for the sample space $\Omega^i\times\Omega^j$ induced by $\boldsymbol{p}$. Define $w_{i,j}=\left(\left\langle d^{i,j},\boldsymbol{p}^{i,j}\right\rangle_\ell\right)^{\frac{1}{\ell}}$. For any $i,j,k\in[n]$ and $1\leq\ell\leq\infty$ we have that $w_{i,j}\leq w_{i,k}+w_{k,j}$.*

Proof. Let $\boldsymbol{p}^{i,j,k}$ be the marginal of $\boldsymbol{p}$ for the sample space $\Omega^i\times\Omega^j\times\Omega^k$. We can write $w_{i,j}=\left(\left\langle d^{i,j},\boldsymbol{p}^{i,j}\right\rangle_\ell\right)^{\frac{1}{\ell}}=\left(\sum_{s,t,r}(d^{i,j}_{s,t})^\ell\boldsymbol{p}^{i,j,k}_{s,t,r}\right)^{\frac{1}{\ell}}\leq\left(\sum_{s,t,r}(d^{i,k}_{s,r}+d^{k,j}_{r,t})^\ell\boldsymbol{p}^{i,j,k}_{s,t,r}\right)^{\frac{1}{\ell}}$. We can now use Minkowski's inequality on a L_ℓ space with measure $\boldsymbol{p}^{i,j,k}$ to bound the last term by $\left(\sum_{s,t,r}(d^{i,k}_{s,r})^\ell\boldsymbol{p}^{i,j,k}_{s,t,r}\right)^{\frac{1}{\ell}}+\left(\sum_{s,t,r}(d^{k,j}_{r,t})^\ell\boldsymbol{p}^{i,j,k}_{s,t,r}\right)^{\frac{1}{\ell}}=w_{i,k}+w_{k,j}$. ☐

The following lemma concerns the map $\mathcal{H}^n$ that takes a tuple $(i,j), 1\leq i<j\leq n$, into a list of either 2, 3 or 4 triples.

$$(i,j)\rightarrow\mathcal{H}^n(i,j)=\mathcal{H}^n_1(i,j)\oplus\mathcal{H}^n_2(i,j)$$

where $\oplus$ denotes a list join operation with *no duplicate removal* – two tuples (resp. triples) are assumed duplicates iff all of their components agree – and

$$\mathcal{H}^n_1(i,j)=\begin{cases}\{(i,n+1,h(i))\} & \text{, if } j=n\wedge i=1\\ \{(i,j,h(i)),(j,n+1,h(i))\} & \text{, if otherwise}\end{cases},$$

and

$$\mathcal{H}^n_2(i,j)=\begin{cases}\{(j,n+1,h(j))\} & \text{, if } i=j-1\\ \{(i,j,h(j)),(i,n+1,h(j))\} & \text{, if } i<j-1\end{cases}.$$

$h(\cdot)$ is also a function of n but for simplicity we omit it in the notation. $h(\cdot)$ is defined as

$$h(i)=1+((i-2)\bmod n). \qquad (6)$$

Lemma 3. *Let $(a,b,c)\in\mathcal{H}^n(i,j)$ for $1\leq i<j\leq n$. Then, $1\leq a<b\leq n+1$, $1\leq c\leq n$, and $c\notin\{a,b\}$. Furthermore, the set*

$$\bigoplus_{1\leq i<j\leq n}\mathcal{H}^n(i,j) \qquad (7)$$

has no duplicates.

Proof. The fact that $1\leq a<b\leq n+1$ and that $1\leq c\leq n$ is immediate. To see that $c\notin\{a,b\}$, we just need to notice that $h(i)\notin\{i,n+1\}$ for $i\in[n]$. The fact that $h(i)\neq n+1$ follows the range of h being $[n]$. If we had $h(i)=i$, then we would have $(i-2)\bmod n=i-1$, which is not possible.

To see that the set (7) does not have duplicates, we just need to see that, starting from two different tuples, the different expressions that define the triples that go into (7) can never be equal. This is a tedious exercise, which we now illustrate for a few cases.

Consider two tuples $(i,j) \neq (i',j')$, $i<j$, and $i'<j'$. Recall that we consider tuples, or triples, equal iff all of their components agree. If $j=j'<n$, we can see e.g. that $(i,j,h(i)) \neq (i',j',h(i'))$, and also that $(i,j,h(i)) \neq (i',n+1,h(i'))$, because $i \neq i'$. If $j<n$ and $i'=j'-1$, we can see e.g. that $(j,n+1,h(i)) \neq (j',n+1,h(j'))$, since equality would require $j'=j$, and also $i=j'$, which in turn would imply $i=j$, which is false. As a final example, if $j<n$, and $i'<j'-1$ we can see e.g. that $(j,n+1,h(i)) \neq (i',n+1,h(j'))$, since equality requires $j=i'$ and $i=j'$, which implies $i'>j'$, which is false. □

For example, if $n=3$, then the possible tuples $(1,2)$, and $(1,3)$, and $(2,3)$, get mapped respectively to $(1,2,3)$, $(2,4,3)$, $(2,4,1)$, and $(1,4,3)$, $(1,3,2)$, $(1,4,2)$, and $(2,3,1)$, $(3,4,1)$, $(3,4,2)$, all of which are different and satisfy the claims in Lemma 3.

The next lemma concerns a different map $\mathcal{H'}^n$ from a triple $(i,j,r), 1 \leq i<j \leq n, r \in [n-2]$ to either 2, 3, or 4 triples:

$$(i,j,r) \rightarrow \mathcal{H'}^n(i,j,r) = \mathcal{H'}^n_1(i,j,r) \oplus \mathcal{H'}^n_2(i,j,r),$$

where

$$\mathcal{H'}^n_1(i,j,r) = \begin{cases} \{(i,r,h'(i,r)))\} & ,\text{if } j = h'(i,r) \\ \{(i,j,h'(i,r),(j,r,h'(i,r))\} & ,\text{if } j \neq h'(i,r) \end{cases},$$

$$\mathcal{H'}^n_2(i,j,r) = \begin{cases} \{(j,r,h'(j,r)))\} & ,\text{if } i = h'(j,r) \\ \{(i,j,h'(j,r)),(i,r,h'(j,r))\} & ,\text{if } i \neq h'(j,r) \end{cases}.$$

We assume that the first two components of each output triple are ordered. For example, $(i,r,h'(j,r)) \equiv (\min\{i,r\},\max\{i,r\},h'(j,r))$. We also assume that the last component of each output triple is taken modulo n in the range $[n]$. For example, if $h'(i,r)=0$, then $(i,r,h'(i,r)) \equiv (i,r,n)$; if $h'(i,r)=n+1$ then $(i,r,h'(i,r)) \equiv (i,r,1)$.

Now, we define function h', also a function of n, as

$$h'(i,r) = \begin{cases} 1+((i+r-1) \bmod n) & ,\text{if } i<n \\ 1+(r \bmod n-1) & ,\text{if } i=n \end{cases} \quad (8)$$

Lemma 4. *Let $(a,b,c) \in \mathcal{H}^n(i,j,r)$, $1 \leq i<j \leq n$, $r \in [n-1]$. Then, $1 \leq a \leq b \leq n+1$, $1 \leq c \leq n$, and $c \notin \{a,b\}$. Furthermore,*

$$\bigoplus_{1 \leq i<j \leq n} \mathcal{H'}^n(i,j) \quad (9)$$

has at most 5 copies of each triple.

Proof. The fact that $1 \leq a \leq b \leq n+1$ and that $1 \leq c \leq n$ is immediate. The fact that $c \neq \{a,b\}$ amounts to checking that $h'(i,r) \notin \{i,r\}$ for all $i \in [n]$, and $r \in [n-1]$. This can be checked directly from (8). E.g. $h'(i,r)=i$ would imply either $(i=n) \wedge (r \bmod n-1 = i-1)$, or $(i<n) \wedge ((i+r-1) \bmod n = i-1)$, both of which are impossible. The rest of the proof amounts to checking that the different expressions that define the triples that go into the set (9) via mapping two triples $(i,j,r) \neq (i',j',r')$, can coincide at most 5 times. Just like for Lemma 3, this is a tedious but simple exercise that we omit for space reasons. □

Remark 1. *Unlike in Lemma 3, we might have $a=b$ in an triple (a,b,c) output by $\mathcal{H'}^n$.*

For example, if $n=4$, all 5 triples $(1,2,3)$, $(1,3,2)$, $(2,3,2)$, $(2,3,3)$, and $(2,3,4)$ map to $(2,3,1)$. Also, both $(1,2,1)$ and $(1,4,1)$ map to $(1,1,2)$ whose first two components equal.

IV. Main results

Given n probability mass functions $\boldsymbol{p^*}^{1:n}$, the ith function being associated with a sample space Ω^i, we define a distance $\mathcal{W}$ between subsets of these mass functions as follows. First, without loss of generality, we assume that $\boldsymbol{p^*}^i_s > 0$ for all $i \in [n]$, and $s \in [m^i]$. Let $\boldsymbol{p}$ be a probability mass function for the sample space $\Omega^1 \times \cdots \times \Omega^n$. Let $\boldsymbol{p}^i$ be the marginal probability of $\boldsymbol{p}$ on Ω^i, and $\boldsymbol{p}^{i,j}$ be the marginal probability of $\boldsymbol{p}$ on $\Omega^i \times \Omega^j$. Let d be a metric over $\Omega^{1:n}$. Let $i_1,\ldots,i_k \in [n]$.

Definition 1.

$$\mathcal{W}^{i_{1:k}} = \min_{\boldsymbol{p}:\boldsymbol{p}^{i_s} = \boldsymbol{p^*}^{i_s} \forall s \in [k]} \sum_{1 \leq s<t \leq k} \left(\langle d^{i_s,i_t}, \boldsymbol{p}^{i_s,i_t} \rangle_\ell\right)^{\frac{1}{\ell}}. \quad (10)$$

Notice in particular that, if $n=2$, then this definition reduces to the classical Wasserstein distance. Our definition is a special case of the Kantorovich formulation for the general multi-marginal optimal transport problem discussed in [17]. Whereas most results regarding the multi-marginal optimal transport problem have focused on questions of existence of Monge solutions, as well as the uniqueness and structure of Kantorovich solutions, we focus on the generalized metric properties of the distances that these problems define.

Theorem 1. *$\mathcal{W}$ defines an n-metric. Namely, for any spaces $\Omega^{1:n}$, mass functions $\boldsymbol{p^*}^{1:n}$, and metric d over $\Omega^{1:n}$, we have*

i) $\mathcal{W}^{1,\ldots,n} \geq 0$;
ii) $\mathcal{W}^{1,\ldots,n} = \mathcal{W}^{\sigma(1,\ldots,n)}$, *for any permutation map σ;*
iii) $\mathcal{W}^{1,\ldots,n} = 0 \iff \boldsymbol{p^*}^i = \boldsymbol{p^*}^j$, *and $\Omega^i = \Omega^j$, $\forall i,j \in [n]$;*
iv) $C(n)\mathcal{W}^{1,\ldots,n-1} \leq \sum_{r=1}^{n-1} \mathcal{W}^{1,\ldots,r-1,r+1,\ldots,n}$;

where $C(n) \geq 1$.

Remark 2. *When we write $\boldsymbol{p^*}^i = \boldsymbol{p^*}^j$ and $\Omega^i = \Omega^j$, we mean that $m^i = m^j$, and there exists a bijection $b^{i,j}(\cdot)$ from $[m^i]$ to $[m^j]$ such that $\boldsymbol{p^*}^i_s = \boldsymbol{p^*}^j_{b^{i,j}(s)}$ and $\Omega^i_s = \Omega^j_{b^{i,j}(s)}$, $\forall\, s \in [m^i]$.*

Theorem 2. *In Theorem 1, the constant $C(n)$ can be made larger than $(n-2)/5$ for $n>7$, and there exists sample spaces $\Omega^{1:n}$, mass functions $\boldsymbol{p^*}^{1:n}$, and a metric d over $\Omega^{1:n}$ such that iv) holds with $C(n)=n-2$.*

V. Proof of Theorem 1

We prove the four metric properties in order. It is trivial to prove the first three properties given the definition of our distance function for the transport problem. Then, we provide a detailed proof for the triangle inequality.

A. Non-Negativity

Proof. The non-negativity of $d^{i,j}$ and $\boldsymbol{p}^{i,j}$, implies that $\langle d^{i,j}, \boldsymbol{p}^{i,j}\rangle_\ell \geq 0$, and hence that $\mathcal{W} \geq 0$. □

B. Symmetry

Proof. Consider a generic permutation map σ, and a map σ' obtained by flipping the order of $1 \leq i < j \leq n$. Let $\boldsymbol{p}^*$ be a minimizer of (10) for $\mathcal{W}^{\sigma(1,\dots,n)}$. Define ${\boldsymbol{p}^*}'_{1,\dots,n} = \boldsymbol{p}^*_{1,\dots,i-1,j,i+1,\dots,j-1,i,j+1,\dots,n}$. ${\boldsymbol{p}^*}'_{1,\dots,n}$ satisfies the constraints of (10) for $\mathcal{W}^{\sigma'(1,\dots,n)}$. Furthermore, for any ordered tuple $(a,b) \neq (i,j)$, we have $\boldsymbol{p}^{*a,b} = \boldsymbol{p}^{*\prime a,b}$ and hence $\langle d^{a,b}, \boldsymbol{p}^{*a,b}\rangle_\ell = \langle d^{a,b}, \boldsymbol{p}^{*\prime a,b}\rangle_\ell$. We know that $d^{i,j} = d^{j,i}$, and that $\boldsymbol{p}^{*i,j} = \boldsymbol{p}^{*\prime j,i}$. Therefore, $\langle d^{i,j}, \boldsymbol{p}^{*i,j}\rangle_\ell = \langle d^{j,i}, \boldsymbol{p}^{*\prime j,i}\rangle_\ell$. Hence, $\mathcal{W}^{\sigma'(1,\dots,n)} \leq \mathcal{W}^{\sigma(1,\dots,n)}$. Since σ is generic, by a similar argument, we have that $\mathcal{W}^{\sigma'(1,\dots,n)} \geq \mathcal{W}^{\sigma(1,\dots,n)}$. Since any permutation can be constructed from a series of swaps, the symmetry property follows. □

C. Identity

Proof. We prove each direction of the equivalence separately.

"$\Longleftarrow$": If for each $i,j \in [n]$ we have $\Omega^i = \Omega^j$, then $m^i = m^j$, and there exists a bijection $b^{i,j}(\cdot)$ from $[m^i]$ to $[m^j]$ such that $\Omega^i_s = \Omega^j_{b^{i,j}(s)}$ for all s. If furthermore $\boldsymbol{p}^{*i} = \boldsymbol{p}^{*j}$, we can define a $\boldsymbol{p}$ such that its univariate marginal $\boldsymbol{p}^i$ satisfies $\boldsymbol{p}^i = \boldsymbol{p}^{*i}$, and such that its bivariate marginal $\boldsymbol{p}^{i,j}_{s,t}$ satisfies $\boldsymbol{p}^{i,j}_{s,t} = \boldsymbol{p}^{*i}_s$, if $t = b^{i,j}(s)$, and zero otherwise. Such a $\boldsymbol{p}$ achieves an objective value of 0 in (10), the smallest value possible by the first metric property (already proved). Therefore, $\mathcal{W}^{1,\dots,n} = 0$.

"$\Longrightarrow$": Now let $\boldsymbol{p}^*$ be a minimizer of (10) for $\mathcal{W}^{1,\dots,n}$. If $\mathcal{W}^{1,\dots,n} = 0$ then $\langle d^{i,j}, \boldsymbol{p}^{*i,j}\rangle_\ell = 0$ for all i,j. Let us consider a specific pair i,j, and, without loss of generality, let us assume that $m^i \leq m^j$. Since, by assumption, we have that $\boldsymbol{p}^{*i}_s > 0$ for all $s \in [m^i]$, and $\boldsymbol{p}^{*j}_s > 0$ for all $s \in [m^j]$, there exists an injection $b^{i,j}(\cdot)$ from $[m^i]$ to $[m^j]$ such that $\boldsymbol{p}^{*i,j}_{s,b^{i,j}(s)} > 0$ for all $s \in [m^i]$. Therefore, $\langle d^{i,j}, \boldsymbol{p}^{*i,j}\rangle_\ell = 0$ implies that $d^{i,j}_{s,b^{i,j}(s)} = 0$ for all $s \in [m^i]$. Therefore, since d is a metric, it must be that $\Omega^i_s = \Omega^j_{b^{i,j}(s)}$ for all $s \in [m^i]$. Now lets us suppose that there exists an $r \in [m^j]$ that is not in the range of $b^{i,j}$. Since, by assumption, all of the elements of the sample spaces are different, it must be that $d^{i,j}_{s,r} > 0$ for all $s \in [m^i]$. Therefore, since $\langle d^{i,j}, \boldsymbol{p}^{*i,j}\rangle_\ell = 0$, it must be that $\boldsymbol{p}^{*i,j}_{s,r} = 0$ for all $s \in [m^i]$. This contradicts the fact that $\sum_{s\in[m^i]} \boldsymbol{p}^{*i,j}_{s,r} = \boldsymbol{p}^{*j}_r > 0$ (the last inequality being true by assumption). Therefore, $m^i = m^j$, and the existence of $b^{i,j}$ proves that $\Omega^i = \Omega^j$. At the same time, since $d^{i,j}_{s,t} > 0$ for all $t \neq b^{i,j}(s)$, it must be that $\boldsymbol{p}^{*i,j}_{s,t} = 0$ for all $t \neq b^{i,j}(s)$. Therefore, $\boldsymbol{p}^{*i}_s = \boldsymbol{p}^{*j}_{b^{i,j}(s)}$ for all s, i.e. $\boldsymbol{p}^{*i} = \boldsymbol{p}^{*j}$. □

D. Generalized Triangle Inequality

Proof. Let $\boldsymbol{p}^*$ be a minimizer for (the optimization problem associated with) $\mathcal{W}^{1,\dots,n}$, and let $\boldsymbol{p}^{*i,j}$ be the marginal induced by $\boldsymbol{p}^*$ for the sample space $\Omega^i \times \Omega^j$. We can write that

$$\mathcal{W}^{1,\dots,n-1} = \sum_{1\leq i<j\leq n-1} \left(\left\langle d^{i,j}, \boldsymbol{p}^{*i,j}\right\rangle_\ell\right)^{\frac{1}{\ell}}. \tag{11}$$

For $r \in [n-1]$, let $\boldsymbol{p}^{(*r)}$ be a minimizer for $\mathcal{W}^{1,\dots,r-1,r+1,\dots,n}$. For $i,j \in [n]\backslash\{r\}$, let ${\boldsymbol{p}^{(*r)}}^{i,j}$ be the marginal of $\boldsymbol{p}^{(*r)}$ for the sample space $\Omega^i \times \Omega^j$. Recall that since $\boldsymbol{p}^{(*r)}$ satisfies the constraints in (10), its marginal for the sample space Ω^i is $\boldsymbol{p}^{*i}$, which is given in advance.

Let $h(\cdot)$ be the map from $[n-1]$ to $[n-1]$ defined as (6) but with n replaced by $n-1$.

Define the following mass function for $\Omega^1 \times \dots \times \Omega^n$,

$$\boldsymbol{p} = \mathcal{G}\left(\boldsymbol{p}^{*n}, \{{\boldsymbol{p}^{(*h(i))}}^{i|n}\}_{i\in[n-1]}\right), \tag{12}$$

where ${\boldsymbol{p}^{(*h(i))}}^{i|n}$ is defined as the mass function that satisfies ${\boldsymbol{p}^{(*h(i))}}^{i|n}\boldsymbol{p}^{*n} = {\boldsymbol{p}^{(*h(i))}}^{i,n}$. Notice that since $h(i) \notin \{i,n\}$, the probability ${\boldsymbol{p}^{(*h(i))}}^{i,n}$ exists for all $i \in [n-1]$.

Let $\boldsymbol{p}^{1,\dots,n-1}$ be the marginal of $\boldsymbol{p}$ for sample space $\Omega^1 \times \dots \times \Omega^{n-1}$, and $\boldsymbol{p}^{i,j}$ be the marginal of $\boldsymbol{p}$ for $\Omega^i \times \Omega^j$.

By Lemma 1, we know that the ith univariate marginal of $\boldsymbol{p}$ is $\boldsymbol{p}^{*i}$ (given) and hence $\boldsymbol{p}^{1,\dots,n-1}$ satisfies the constraints associated with $\mathcal{W}^{1,\dots,n-1}$. Therefore, we can write that

$$\sum_{1\leq i<j\leq n-1} \left(\left\langle d^{i,j}, \boldsymbol{p}^{*i,j}\right\rangle_\ell\right)^{\frac{1}{\ell}} \leq \sum_{1\leq i<j\leq n-1} \left(\left\langle d^{i,j}, \boldsymbol{p}^{i,j}\right\rangle_\ell\right)^{\frac{1}{\ell}}. \tag{13}$$

By Lemma 2, inequality (14) holds; because d is symmetric, (15) holds; by the definition of $\boldsymbol{p}$, (16) follows. Therefore,

$$\left(\langle d^{i,j}, \boldsymbol{p}^{i,j}\rangle_\ell\right)^{\frac{1}{\ell}} \leq \left(\langle d^{i,n}, \boldsymbol{p}^{i,n}\rangle_\ell\right)^{\frac{1}{\ell}} + \left(\langle d^{n,j}, \boldsymbol{p}^{n,j}\rangle_\ell\right)^{\frac{1}{\ell}} \tag{14}$$

$$= \left(\langle d^{i,n}, \boldsymbol{p}^{i,n}\rangle_\ell\right)^{\frac{1}{\ell}} + \left(\langle d^{j,n}, \boldsymbol{p}^{j,n}\rangle_\ell\right)^{\frac{1}{\ell}} \tag{15}$$

$$= \left(\left\langle d^{i,n}, {\boldsymbol{p}^{(*h(i))}}^{i,n}\right\rangle\right)^{\frac{1}{\ell}} + \left(\left\langle d^{j,n}, {\boldsymbol{p}^{(*h(j))}}^{j,n}\right\rangle_\ell\right)^{\frac{1}{\ell}}. \tag{16}$$

Let $w_{(i,j)}$ denote each term on the r.h.s. of (11), and $w_{(i,j,r)}$ denote $\left(\left\langle d^{i,j}, {\boldsymbol{p}^{(*r)}}^{i,j}\right\rangle_\ell\right)^{\frac{1}{\ell}}$. Combining (13) - (16), we have

$$\sum_{1\leq i<j\leq n-1} w_{(i,j)} \leq \sum_{1\leq i<j\leq n-1} w_{(i,n,h(i))} + w_{(j,n,h(j))}. \tag{17}$$

Finally, we write the r.h.s of iv) in Theorem 1 as in (18) and show that (18) upper-bounds the r.h.s of (17).

$$\sum_{r=1}^{n-1} \sum_{i,j\in[n]\backslash\{r\}, i<j} w_{(i,j,r)}. \tag{18}$$

First, by Lemma 2 and the symmetry of d, we have

$$w_{(i,n,h(i))} \leq w_{(i,j,h(i))} + w_{(j,n,h(i))}, \tag{19}$$

$$w_{(j,n,h(j))} \leq w_{(i,j,h(j))} + w_{(i,n,h(j))}, \tag{20}$$

as long as for each triple (a,b,c) in the above expressions, $c \notin \{a,b\}$. We will use these inequalities to upper bound some of the terms on the r.h.s. of (17), which can be further upper bounded by (18). In particular, we will apply inequalities (19) and (20) such that the terms $w_{a,b,c}$ that we get after their use have triples (a,b,c) that match the triples obtained via the map $\mathcal{H}^{n-1}$ defined in Section III. To be concrete, for example, if $\mathcal{H}^{n-1}$ maps (i,j) to $\{(i,n,h(i)),(j,n,h(j))\}$, then we do not apply (19) and (20), and we leave $w_{(i,n,h(i))} +$

$w_{(j,n,h(j))}$ as is on the r.h.s. of (17). If, for example, $\mathcal{H}^{n-1}$ maps (i,j) to $\{(i,n,h(i)),(i,j,h(j)),(i,n,h(j))\}$, then we leave the first term in $w_{(i,n,h(i))}+w_{(j,n,h(j))}$ in the r.h.s. of (17) untouched, but we upper bound the second term using (20) to get $w_{(i,n,h(i))}+w_{(i,j,h(j))}+w_{(j,n,h(j))}$.

After proceeding in this fashion, and by Lemma 3, we know that all of the terms $w_{(a,b,c)}$ that we obtain have triples (a,b,c) with $c \neq \{a,b\}$, with $c \in [n-1]$, and $1 \leq a < b \leq n$. Therefore, these terms appear in (18). Also by Lemma 3, we know that we do not get each triple more than once. Therefore, the upper bound that we just constructed with the help of $\mathcal{H}^{n-1}$ for the r.h.s of (17) can be upper bounded by (18). □

VI. Proof of Theorem 2

A. Lower bound on $C(n)$

The lower bound $C(n) > (n-2)/5$ for $n > 7$, can be proved by revisiting the the proof of the triangle inequality iv) in Theorem 1, and using Lemma 4 instead of Lemma 3. In particular, we will show that, $(n-2)\mathcal{W}^{1,\dots,n-1}$ can be upper bounded by $5\sum_{r=1}^{n-1}\mathcal{W}^{1,\dots,r-1,r+1,\dots,n}$.

We modify (12) and define the probability mass functions

$$\boldsymbol{p}^{(r)} = \mathcal{G}\left(\boldsymbol{p}^{*r}, \{\boldsymbol{p}^{(*h'(i,r))^{i|r}}\}_{i\in[n-1]}\right), \text{for } r \in [n-2],$$

on $\Omega^1 \times \cdots \times \Omega^n$. The mass functions inside $\mathcal{G}$ are defined as in the proof of Theorem 1, expect that we now use $h'(i,r)$ instead of $h(i)$. The map $h'(\cdot,\cdot)$ is defined as in (8) but with n in its definition replaced by $n-1$, and such that its output values are taken modulo $n-1$ but in the range $[n-1]$. In other words, if $n-1=5$ and $h'=6$, then its output value is actually 1, and if $h'=0$, then its output value is actually 5. Note that $h'(i,r) \notin \{i,r\}$, for all $1 \leq i \leq n-1$, and $r \in [n-2]$, therefore $\boldsymbol{p}^{(*h'(i,r))^{i,r}}$ and $\boldsymbol{p}^{(*h'(i,r))^{i|r}}$ exist.

In the next step, we first rewrite (13) as follows,

$$\sum_{r=1}^{n-2}\sum_{1\leq i<j\leq n-1}\left(\left\langle d^{i,j}, \boldsymbol{p}^{*i,j}\right\rangle_{\ell}\right)^{\frac{1}{\ell}} \leq \sum_{r=1}^{n-2}\sum_{1\leq i<j\leq n-1}\left(\left\langle d^{i,j}, \boldsymbol{p}^{(r)^{i,j}}\right\rangle_{\ell}\right)^{\frac{1}{\ell}}, \quad (21)$$

where $\boldsymbol{p}^{(r)^{i,j}}$ is the marginal for $\Omega^i \times \Omega^j$ induced by $\boldsymbol{p}^{(r)}$. From (21), and using Lemma 2, we rewrite (17) as follows,

$$\sum_{r=1}^{n-2}\sum_{1\leq i<j\leq n-1} w_{(i,j,r)} \leq \sum_{r=1}^{n-2}\sum_{1\leq i<j\leq n-1} v_{(i,r,h'(i,r))} + v_{(r,j,h'(j,r))}, \quad (22)$$

where we are using the following notation: (a) we are implicitly assuming that the first two components of each triple on the r.h.s. of (22) are ordered, i.e. if e.g. $r < i$ then $(r,i,h'(i,r))$ should be red as $(i,r,h'(i,r))$; (b) each $w_{(i,j,r)}$ represents one $\left(\left\langle d^{i,j}, \boldsymbol{p}^{*i,j}\right\rangle_{\ell}\right)^{\frac{1}{\ell}}$ on the l.h.s. of (21); and (c) each $v_{(i,j,r)}$ represents $\left(\left\langle d^{i,j}, \boldsymbol{p}^{(*r)^{i,j}}\right\rangle_{\ell}\right)^{\frac{1}{\ell}}$ if $i \neq j$, and is zero if $i = j$. Recall that since $h'(i,r) \notin \{i,r\}$ and $h'(j,r) \notin \{j,r\}$, when $i \neq j$, both $\boldsymbol{p}^{(*h'(i,r))^{i,r}}$ and $\boldsymbol{p}^{(*h'(j,r))^{r,j}}$ exist.

Finally, we write $5\sum_{r=1}^{n-1}\mathcal{W}^{1,\dots,r-1,r+1,\dots,n}$ as in (23) and show that (23) upper-bounds the r.h.s. of (22).

$$5\sum_{r=1}^{n-1}\sum_{i,j\in[n]\setminus\{r\},i<j} w_{(i,j,r)}. \quad (23)$$

First, by Lemma 2 and the symmetry of d, we have

$$v_{(i,r,h'(i,r))} \leq v_{(i,j,h'(i,r))} + v_{(j,r,h'(i,r))}, \quad (24)$$

$$v_{(j,r,h'(j,r))} \leq v_{(i,j,h'(j,r))} + v_{(i,r,h'(j,r))}, \quad (25)$$

as long as for each triple (a,b,c) in the above expressions, $c \notin \{a,b\}$. We will use inequalities (24) and (25) to upper bound some of the terms on the r.h.s. of (22), which we will then show can be upper bounded by (23). In particular, we will apply inequalities (24) and (25) such that the terms $v_{a,b,c}$ that we get after their use have triples (a,b,c) that match the triples obtained via the map $\mathcal{H}'^{n-1}$ defined in Section III. To be concrete, for example, if $\mathcal{H}'^{n-1}$ maps (i,j,r) to $\{(i,r,h'(i,r)),(r,j,h'(j,r))\}$, then we do not apply (24) and (25), and we leave $v_{(i,r,h'(i,r))} + v_{(r,j,h'(j,r))}$ as is on the r.h.s. of (22). If, for example, $\mathcal{H}^{n-1}$ maps (i,j,r) to $\{(i,r,h'(i,r)),(i,j,h'(j,r)),(i,r,h'(j,r))\}$, then we leave the first term in $v_{(i,r,h'(i,r))} + v_{(r,j,h'(j,r))}$ in the r.h.s. of (22) untouched, but we upper bound the second term using (25) to get $v_{(i,r,h'(i,r))} + v_{(i,j,h'(j,r))} + v_{(i,r,h'(j,r))}$.

After proceeding in this fashion, and by Lemma 4, we know that all of the terms $v_{(a,b,c)}$ that we obtain have triples (a,b,c) with $c \neq \{a,b\}$, $c \in [n-1]$, and $1 \leq a < b \leq n$. Therefore, these terms are either zero (if $a=b$) or appear in (23). Also because of Lemma 4, each triple will not appear for more than 5 times. Therefore, the upper bound we build with the help of $\mathcal{H}^{n-1}$ for the r.h.s of (22) can be upper bounded by (23).

B. Upper bound on $C(n)$

Consider the following setup. Let $m^i = m$ for all $i \in [n]$, and $\Omega^i_s \in \mathbb{R}$ for all $i \in [n]$, $s \in [m]$. Define d such that $d^{i,j}_{s,t}$ is $|\Omega^i_s - \Omega^j_t|$, if $s = t$, and infinity otherwise. Let $\boldsymbol{p}^{*i}_{\ s} = \frac{1}{m}$ for all $i \in [n]$, $s \in [m]$.

Any optimal solution to the MMOT problem must satisfy $\boldsymbol{p}^{*i,j}_{\ s,t} = \frac{1}{m}\delta_{s,t}$, and thus $\left(\left\langle d^{i,j}, \boldsymbol{p}^{*i,j}\right\rangle_{\ell}\right)^{\frac{1}{\ell}} = \frac{1}{m^{\ell}}\|\Omega^i - \Omega^j\|_{\ell}$, where we interpret Ω^i has a vector in $\mathbb{R}^m$, and $\|\cdot\|_{\ell}$ is the vector ℓ-norm. Therefore, ignoring the factor $\frac{1}{m^{\ell}}$, we only need to prove that iv) in Theorem 1 holds with $C(n) = n-2$ when $\mathcal{W}^{1:n-1}$ is defined as $\sum_{1\leq i<j\leq n-1}\|\Omega^i - \Omega^j\|_{\ell}$. This in turn is a standard result, whose proof (in a more general form) can be found e.g. in Example 2.4 in [21].

VII. Future work

It is an open problem to find more general conditions on d that result in (2) being an n-metric. Our results hold for (2) with $d(w^{1:n}) = \left(\sum_{i<j}(d^{i,j}(w^i,w^j))^{\ell}\right)^{\frac{1}{\ell}}$. It is trivial to prove

that if $d^{i,j}$ is a metric $\forall i, j$, then d is an n-metric. Could it be true that as long as d is an n-metric that (2) will be too? Will the bounds on $C(n)$ hold for more general definitions of d? We leave these questions to future work.

References

[1] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[2] L. V. Kantorovich, "On the translocation of masses," in *Dokl. Akad. Nauk SSSR*, vol. 37, 1942, pp. 199–201.

[3] Y. Brenier, "Polar factorization and monotone rearrangement of vector-valued functions," *Communications on pure and applied mathematics*, vol. 44, no. 4, pp. 375–417, 1991.

[4] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *NeurIPS*, 2013.

[5] M. Arjovsky *et al.*, "Wasserstein generative adversarial networks," in *ICML*, 2017.

[6] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE signal processing magazine*, vol. 34, no. 4, pp. 43–59, 2017.

[7] R. M. Gray, "Source coding and simulation," *IEEE Information Theory Society Newsletter*, vol. 58, no. 4, p. 1, 2008.

[8] W. Kreitmeier, "Optimal vector quantization in terms of wasserstein distance," *Journal of Multivariate Analysis*, vol. 102, no. 8, pp. 1225–1239, 2011.

[9] M. Raginsky, I. Sason *et al.*, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends® in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

[10] S.-i. Amari, R. Karakida, and M. Oizumi, "Information geometry connecting wasserstein distance and kullback–leibler divergence via the entropy-relaxed transportation problem," *Information Geometry*, vol. 1, no. 1, pp. 13–37, 2018.

[11] J. Solomon *et al.*, "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Trans. Graph.*, vol. 34, no. 4, p. 66, 2015.

[12] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[13] B. B. Damodaran *et al.*, "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation," *arXiv preprint arXiv:1803.10081*, 2018.

[14] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[15] A. Gerolin, A. Kausamo, and T. Rajala, "Duality theory for multi-marginal optimal transport with repulsive costs in metric spaces," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 25, p. 62, 2019.

[16] A. Moameni and B. Pass, "Solutions to multi-marginal optimal transport problems concentrated on several graphs," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 23, no. 2, pp. 551–567, 2017.

[17] B. Pass, "Multi-marginal optimal transport: theory and applications," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 49, no. 6, pp. 1771–1790, 2015.

[18] C. T. Li and V. Anantharam, "Pairwise multi-marginal optimal transport and embedding for earth mover's distance," *arXiv preprint arXiv:1908.01388*, 2019.

[19] J. Cao, L. Mo, Y. Zhang, K. Jia, C. Shen, and M. Tan, "Multi-marginal wasserstein gan," in *Advances in Neural Information Processing Systems*, 2019, pp. 1774–1784.

[20] M. Khamsi, "Generalized metric spaces: a survey," *Journal of Fixed Point Theory and Applications*, vol. 17, no. 3, pp. 455–475, 2015.

[21] G. Kiss, J.-L. Marichal, and B. Teheux, "A generalization of the concept of distance based on the simplex inequality," *Beiträge zur Algebra und Geometrie/Contributions to Algebra and Geometry*, vol. 59, no. 2, pp. 247–266, 2018.