Strongly Isomorphic Neural Optimal Transport Across Incomparable Spaces

Athina Sotiropoulou [1]   **David Alvarez-Melis** [2] [3]

Abstract

Optimal Transport (OT) has recently emerged as a powerful framework for learning minimal-displacement maps between distributions. The predominant approach involves a neural parametrization of the Monge formulation of OT, typically assuming the same space for both distributions. However, the setting across "incomparable spaces" (e.g., of different dimensionality), corresponding to the Gromov-Wasserstein distance, remains underexplored, with existing methods often imposing restrictive assumptions on the cost function. In this paper, we present a novel neural formulation of the Gromov-Monge (GM) problem rooted in one of its fundamental properties: invariance to strong isomorphisms. We operationalize this property by decomposing the learnable OT map into two components: (i) an approximate strong isomorphism between the source distribution and an intermediate reference distribution, and (ii) a GM-optimal map between this reference and the target distribution. Our formulation leverages and extends the Monge gap regularizer of Uscidda & Cuturi (2023) to eliminate the need for complex architectural requirements of other neural OT methods, yielding a simple but practical method that enjoys favorable theoretical guarantees. Our preliminary empirical results show that our framework provides a promising approach to learn OT maps across diverse spaces.

1. Introduction

Transforming samples between distributions lies at the core of machine learning, with applications ranging from generative modeling (Creswell et al., 2018; Rezende & Mohamed, 2015; Song et al., 2020) to domain adaptation (Courty et al., 2017) and cell-genomics (Bunne et al., 2023). Optimal transport (OT) provides an elegant, geometrically-driven approach to address this challenge. Given a source measure μ supported on a domain $\mathcal{X}$ and a target measure ν on $\mathcal{Y}$, OT, in its fundamental form, aims at finding a map $T : \mathcal{X} \longrightarrow \mathcal{Y}$, which transports mass from μ to ν through the push-forward operation $T\sharp\mu = \nu$, while minimizing a transportation cost (Santambrogio, 2015).

Most existing work on OT focuses either on defining meaningful distances between distributions (Arjovsky et al., 2017; Alvarez-Melis & Fusi, 2020; Bunne et al., 2019), or on deriving optimal couplings between discrete sets of unpaired samples (Sebbouh et al., 2024; Alvarez-Melis et al., 2019; Fickinger et al., 2021). In the latter case, the derived coupling cannot generalise to out of distributions samples. To this end, one has to derive the optimal transformation map T in the continuous setting. Yet, its computation on high dimensional settings is notoriously challenging.

An emerging area of work that aims to address this issue is that of *neural* Optimal Transport, where the OT map T *itself* is parameterized as a neural network, essentially *learning* the *actual* solution T^* to the OT problem (Makkuva et al., 2020; Korotin et al., 2023; Uscidda & Cuturi, 2023; Rout et al., 2022; Korotin et al., 2020; Rout et al., 2022). Yet, these methods impose a strict requirement: that the source and target measures are supported on the same space, or at the very least on different but 'comparable' spaces (i.e., those across which a meaningful distance can be defined). In particular, they must be of the same dimensionality. This amounts to solving the classic OT formulation of Monge (1781), or the relaxed formulation by Kantorovich (2006), where correspondences are defined as probabilistic couplings. Despite their limitations, these formulations are endowed with the richness of classical optimal transport theory, which is utilized for the design of neural frameworks with robust theoretical guarantees. These (notably) include Brenier's Theorem (Brenier, 1987; Makkuva et al., 2020), cyclical monotonicity (Santambrogio, 2015; Uscidda & Cuturi, 2023) and duality (Santambrogio, 2015; Korotin et al., 2023; Rout et al., 2022; Fan et al., 2023).

However, transforming samples between distributions that live in *incomparable* spaces, e.g., of different dimensionality

[1]Independent Researcher [2]Harvard University [3]Microsoft Research, Cambridge, MA. Correspondence to: Athina Sotiropoulou <athina.sotiropoulou@alumni.ethz.ch>.

or structural form, is essential across a wide range of applications, such as aligning latent representations learned by different models (Alvarez-Melis et al., 2019), or matching samples across different modalities (Demetci et al., 2022; Bunne et al., 2019). The challenge lies in the absence of a meaningful transportation cost across spaces of different dimensionality. To this end, the Gromov-Wasserstein (GW) problem (Mémoli, 2011) has been proposed to generalize OT to incomparable spaces. Instead of a cross inter-domain cost $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$, GW computes individual costs $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}_+$ and $c_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ defined within each space and solves a quadratic problem w.r.t the space of couplings. Restricting the space of couplings to an explicit (deterministic) transport map T, amounts to the "hard-assignment" version of GW, the Gromov-Monge (GM) problem (Mémoli & Needham, 2022).

Unlike the extensive work for comparable spaces, neural frameworks for the GW and GM problems are largely unexplored. An important challenge in this regard is that the main theorems of classic OT do *not* apply to these versions of the problem. In fact, to the best of our knowledge, the only works in this direction are by Nekrashevich et al. (2023) and Klein et al. (2024). Nekrashevich et al. (2023) restrict their framework for the specific case of using inner product inter-domain costs in GW, solving an equivalent min-max-min problem, while Klein et al. (2024) propose a general framework for the entropic GW using neural flow matching. To the best of our knowledge, Neural frameworks for the GM problem, for general inter-domain cost functions, have yet to be explored.

Contributions. In this work, we propose a *Neural Gromov-Monge* framework, that allows for *any* choice of inter-domain costs. Just as Brenier's Theorem is the theoretical powerhouse of classic neural OT, we utilize a fundamental property of the GM (and GW) problem: its *invariance to strong isomorphisms* (Mémoli & Needham, 2022). We show that this property can be utilized to estimate the optimal solution map T^* of the GM problem (the GM-optimal map) with neural networks. Our main contributions are:

- We show that the solution to the general GM problem between two measures can be decomposed into two maps: an isomorphism and a GM-optimal map (Proposition 3.1, Figure 1).
- We show that by parameterizing each map by a neural network, their composition constitutes a universal approximator of *any* transport map between incomparable spaces (Theorem 4.1).
- We propose a neural algorithm for learning the aforementioned neural composition. By extending the Monge-gap regularizer of Uscidda & Cuturi (2023) to the GM case, we showcase that if appropriately minimized, our proposed loss is guaranteed to recover true

GM-optimal maps.
- We empirically demonstrate that our algorithm can recover known GM-optimal maps on synthetic data.

2. Preliminaries and Background

2.1. Metric measure spaces and Strong Isomorphism

We consider a compact metric space $(\mathcal{X}, c_{\mathcal{X}})$ endowed with a continuous and measurable metric $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}_+$. Let $\mu \in \mathcal{P}(\mathcal{X})$ be a Borel probability measure. That is, μ is fully supported on its domain, i.e supp$(\mu) = \mathcal{X}$, and $\mu(\mathcal{X}) = 1$. Then the triplet $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ constitutes a *metric measure space (mm-space)* (Mémoli, 2011). When it is clear from the context, we will denote $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ as simply $\mathcal{X}_\mu$. Following Vayer (2020), we denote the space of *all* mm-spaces, with finite L^p-size, as $\mathcal{M}_p = \{\mathcal{X}_\mu := (\mathcal{X}, c_{\mathcal{X}}, \mu) \mid size_p(\mathcal{X}_\mu) < +\infty\}$, where $size_p(\mathcal{X}_\mu) = \int_{\mathcal{X} \times \mathcal{X}} c_{\mathcal{X}}(x, x')^p \, d\mu \otimes d\mu$, with $\otimes$ denoting the product measure. Let $\mathcal{Y}_\nu := (\mathcal{Y}, c_{\mathcal{Y}}, \nu)$ be a second mm-space endowed with the Borel probability measure $\nu \in \mathcal{P}(\mathcal{Y})$. To this end, we introduce the concept of *strong isomorphism* between two mm-spaces as:

Definition 2.1. [Strong Isomorphism] Two mm-spaces $\mathcal{X}_\mu \triangleq (\mathcal{X}, c_{\mathcal{X}}, \mu) \in \mathcal{M}_p$ and $\mathcal{Y}_\nu \triangleq (\mathcal{Y}, c_{\mathcal{Y}}, \nu) \in \mathcal{M}_p$ are strongly isomorphic, denoted as $\mathcal{X}_\mu \cong^s \mathcal{Y}_\nu$, if there exists a bijective map $\phi : \mathcal{X} \longrightarrow \mathcal{Y}$ s.t:

1. ϕ is an *isometry*, i.e., $\forall x, x' \in \mathcal{X}: c_{\mathcal{X}}(x, x') = c_{\mathcal{Y}}(\phi(x), \phi(x'))$.
2. ϕ pushes μ forward to ν, i.e., $\phi\sharp\mu = \nu$.

Note that the push-forward of a measure through a map ϕ, is the measure $\phi\sharp\mu \in \mathcal{P}(\mathcal{Y})$ satisfying $\phi\sharp\mu(A) := \mu(T^{-1}(A))$ for any measurable set $A \subseteq \mathcal{X}$.

In essence, Definition 2.1 is a definition of *equivalence*, both from a geometric (condition 1. of *exact* distance preservation) and from a statistical perspective (condition 2. of exact measure preservation). Intuitively, the existence of an isometry between the metric spaces $(\mathcal{X}, c_{\mathcal{X}}),(\mathcal{Y}, c_{\mathcal{Y}})$ ensures they have the same spacial configuration. On the other hand, the push-forward condition ensures the corresponding mm-spaces have the same probabilistic configuration, i.e the probability mass in both spaces is distributed in exactly the same manner. As such, any map ϕ that satisfies the conditions in Definition 2.1, is a measure preserving isometry, which we will refer to as an *isomorphism*.

2.2. The Gromov-Monge problem

In this section, we introduce the GM distance between two mm-spaces. Here we consider $\mathcal{X}_\mu$ and $\mathcal{Y}_\nu$ to be two arbitrary spaces, i.e not necessarily isomorphic. In the general setting, the domains $\mathcal{X}$ and $\mathcal{Y}$ are incomparable. Consider the collection of all possible measure preserving transport

maps from $\mathcal{X}_\mu$ to $\mathcal{Y}_\nu$:

$$\mathcal{T}(\mu, \nu) = \{T : \mathcal{X} \longrightarrow \mathcal{Y} \quad | \quad T\sharp\mu = \nu\} \qquad (1)$$

The *distortion* (or more precisely, the p-distortion) induced by *any* such map $T \in \mathcal{T}(\mu, \nu)$ is defined as (Mémoli & Needham, 2022):

$$\text{dis}_p(T)^p = \iint\limits_{\mathcal{X}\times\mathcal{X}} |c_\mathcal{X}(\mathbf{x}, \mathbf{x}') - c_\mathcal{Y}(T(\mathbf{x}), T(\mathbf{x}'))|^p \ d\mu \otimes \mu$$

$$(2)$$

Following Mémoli & Needham (2022) we can re-write eq. (2) in a more concise manner by using the L^p norm of the function spaces, i.e $\text{dis}_p(T)^p := \|c_\mathcal{X} - c_\mathcal{Y}\|_{L^p(\mu_T \otimes \mu_T)}$. Here, $\mu_T = (\text{Id} \times T)\sharp\mu$ is a measure on $\mathcal{X} \times \mathcal{X}$ and Id is the identity map on $\mathcal{X}$. We will resort to this form when it is clear from the context. Essentially, eq. (2) captures the degree to which the geometric structure of $\mathcal{X}_\mu$ is altered during the (exact) transportation of probability mass from μ to ν. A lower distortion implies that T preserves the inter-domain distances between points more accurately, i.e is closer to being an isometry, in the p-norm sense. To this end, the Gromov-Monge problem aims at finding the map which induces the least amount of distortion (Mémoli & Needham, 2022):

$$\text{GM}_p(\mu, \nu) = \inf_{T\in\mathcal{T}(\mu,\nu)} \text{dis}_p(T) \qquad (3)$$

GM-optimal maps. We will refer to the solution of the optimization problem in eq. (3), i.e $T^* \in \arg\min_{T\in\mathcal{T}(\mu,\nu)} \text{dis}_p(T)^{\frac{1}{p}}$, as the *GM-optimal map* between the two spaces. Intuitively, eq. (3) tries to match the probability distributions through an exact mapping, whilst also minimizing structural deformity. Naturally, following Definition 2.1, when $\mathcal{X}_\mu$ and $\mathcal{Y}_\nu$ are strongly isomorphic, we have that $\text{GM}_p(\mu, \nu) = 0$ and T^* will be an isomorphism (Mémoli & Needham, 2022). This means that the GM distance between mm-spaces is *invariant* to strong isomorphisms. In fact, according to Mémoli & Needham (2022a) and Mémoli & Needham (2022b), GM defines a Lawvere metric on $\mathcal{M}_p$.

2.3. The Gromov-Wasserstein Problem

If instead of optimizing over the set of *exact* correspondences (i.e, eq. (1)), we consider *probabilistic* correspondences (i.e, couplings), we get the Gromov-Wasserstein distance as proposed by Mémoli (2011):

$$\text{GW}_p(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \text{dis}_p(\pi) \qquad (4)$$

where the infimum is over the set of couplings between μ and ν, i.e, $\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X} \times$

$\mathcal{Y}) \quad | \quad \pi(A \times Y) = \mu(A) \quad ; \quad \pi(\mathcal{X} \times B) = \nu(B)$ for any measurable $A \subset \mathcal{X}, B \subset \mathcal{Y}\}$ and $\text{dis}_p^p(\pi) = \iint\limits_{(\mathcal{X}\times\mathcal{Y})^2} |c_\mathcal{X}(\mathbf{x}, \mathbf{x}') - c_\mathcal{Y}(\mathbf{y}, \mathbf{y}')|^p \ d\pi \otimes \pi$ represents the distortion induced by a coupling $\pi \in \Pi(\mu, \nu)$.

In essence, the GM problem is the restricted ("hard" assignment) version of the GW problem. The fact that it optimizes over a set of exact point registrations, i.e functions, makes the solution space of GM more suitable for parametrization by neural networks.

Existence of solutions. Similar to the Monge problem for comparable spaces (Monge, 1781), a solution to the GM problem in eq. (3) might not always exist, i.e, the set of transport maps in eq. (1) might be empty. In fact, theoretical guarantees for the existence of GM-optimal maps is an ongoing area of research (Vayer, 2020; Sturm, 2020; Dumont et al., 2024; Salmona et al., 2021) and are restricted to very specific conditions that are non-trivial to guarantee in practice. In contrast, the Gromov-Wasserstein problem between mm-spaces (i.e, eq. (4)) *always* admits a solution (Vayer, 2020; Chowdhury & Mémoli, 2018)[1], albeit not necessarily unique , i.e, the set of couplings $\Pi(\mu, \nu)$ is always non-empty. Nevertheless, when a solution to eq. (3), T^*, *does* exist and the measures μ and ν are non-atomic, the solutions to the two problems coincide (Theorem 2 in Mémoli & Needham (2022a)), i.e, $\pi^* = (\text{Id}, T^*)\sharp\mu$ and $GM(\mu, \nu) = GW(\mu, \nu)$.

3. GM-Optimal Map Decomposition

We will henceforth refer to the mm-spaces $\mathcal{X}_\mu$ and $\mathcal{Y}_\nu$ as the *source* and *target* space respectively and, as in section 2.2, we consider them to be arbitrary (not necessarily isomorphic). Analogously, we refer to μ and ν as the *source* and *target* measures and define their respective supports as compact subsets of Euclidean spaces, i.e $\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} \subseteq \mathbb{R}^q$. We assume the general case, where $p \neq q$. Let $\mathcal{Z}_\rho := (\mathcal{Z}, c_\mathcal{Z}, \rho)$ be a third space, with Borel probability measure $\rho \in \mathcal{P}(\mathcal{Z})$ and $\mathcal{Z} \subseteq \mathbb{R}^p$ compact, which we will refer to as the *reference* space. We consider $\mathcal{Z}_\rho$ to be *strongly isomorphic* to the source space according to Definition 2.1, i.e, $\mathcal{X}_\mu \cong^s \mathcal{Z}_\rho$. Note that we take the source and reference domains to be of equal dimension p. This constitutes a necessary (albeit not sufficient) condition for strong isomorphism.

In this section, we show that the solution of eq. (3) that directly transports μ onto ν, can be decomposed into a sequence of two maps: an isomorphism from μ to the reference measure ρ followed by an optimal transport map from

---

[1]Here we are referring to Theorem 12 in Chowdhury & Mémoli (2018) which considers the GW distance between measure networks (m-nets) but is directly applicable to mm-spaces (which are a specific case of m-nets).
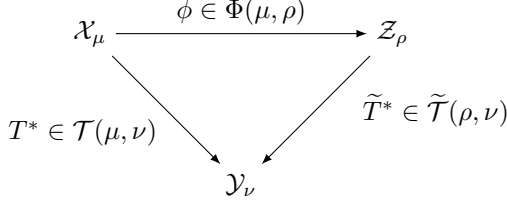
$$\begin{array}{ccc}
\mathcal{X}_\mu & \xrightarrow{\phi \in \Phi(\mu,\rho)} & \mathcal{Z}_\rho \\
& & \\
T^* \in \mathcal{T}(\mu,\nu) & & \widetilde{T}^* \in \widetilde{\mathcal{T}}(\rho,\nu) \\
& \searrow \quad \swarrow & \\
& \mathcal{Y}_\nu &
\end{array}$$

Figure 1. A tripod structure between mm-spaces where $\mathcal{X}_\mu \cong^s \mathcal{Z}_\rho$. $\phi \in \Phi(\mu,\rho)$ represents the collection of isomorphisms, while $\mathcal{T}(\mu,\nu)$ and $\widetilde{\mathcal{T}}(\rho,\nu)$ are the sets of all transport maps between the corresponding spaces.

ρ to ν.

3.1. A tripod structure of mm-spaces

We denote as $\Phi(\mu,\rho)$ the collection of all measure preserving isometries between $\mathcal{X}_\mu$ and $\mathcal{Z}_\rho$. We now focus on the GM distances (and corresponding GM-optimal maps) between the three mm-spaces. Naturally, between the source and reference we have $\mathrm{GM}(\mu,\rho) = 0$, where the corresponding GM-optimal map is an isomorphism $\phi \in \Phi(\mu,\rho)$.

Now let us focus on the following tripod structure:

$$\mathcal{X}_\mu \xrightarrow{\mathcal{T}(\mu,\nu)} \mathcal{Y}_\nu \xleftarrow{\widetilde{\mathcal{T}}(\rho,\nu)} \mathcal{Z}_\rho$$

where $\widetilde{\mathcal{T}}(\rho,\nu)$ is the collection of all transport maps between $\mathcal{Z}_\rho$ and $\mathcal{Y}_\nu$:

$$\widetilde{\mathcal{T}}(\rho,\nu) = \{\widetilde{\mathcal{T}} : \mathcal{Z} \longrightarrow \mathcal{Y} \mid \widetilde{\mathcal{T}}\sharp\rho = \nu\} \qquad (5)$$

and $\mathcal{T}(\mu,\nu)$ is given by eq. (1). Subsequently, the GM problem $\mathrm{GM}_p(\mu,\nu)$ between the source and target spaces is given by eq. (3). The GM distance between $\mathcal{Z}_\rho$ and $\mathcal{Y}_\nu$ will then be defined as:

$$\mathrm{GM}_p(\rho,\nu) = \inf_{\widetilde{T} \in \widetilde{\mathcal{T}}(\rho,\nu)} \mathrm{dis}_p(\widetilde{T}) \qquad (6)$$

where $\mathrm{dis}_p(\widetilde{T})^p := \|c_{\mathcal{Z}} - c_{\mathcal{Y}}\|_{L^p(\rho_{\widetilde{T}} \otimes \rho_{\widetilde{T}})}$, with $\rho_{\widetilde{T}} = (\mathrm{Id} \times \widetilde{T})\sharp\rho \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$. We denote as $\widetilde{T}^* \in \widetilde{\mathcal{T}}(\rho,\nu)$, the solution to eq. (6). It can be easily shown that the GM problems eq. (3) and eq. (6) are equivalent. Formally, we introduce the following Proposition, which is a consequence of the isomorphic invariance of the GM problem, as introduced in section 2.2.

Proposition 3.1. *Let $\mathcal{X}_\mu, \mathcal{Z}_\rho, \mathcal{Y}_\nu \in \mathcal{M}_p$ such that $\mathcal{X}_\mu \cong^s \mathcal{Z}_\rho$. Then for $p \in [1,\infty)$ it holds that $GM_p(\mu,\nu) = GM_p(\rho,\nu)$.*

We provide the proof of Proposition 3.1 in Appendix A.1.

3.2. Optimal map decomposition

Now let us focus on the transport problem from the source to the target space, i.e, $\mathcal{X}_\mu \longrightarrow \mathcal{Y}_\nu$. Consider the structure illustrated in Figure 1. Instead of transporting mass directly from $\mathcal{X}_\mu$ to $\mathcal{Y}_\nu$, we introduce a "detour", by first mapping μ onto ρ through an isomorphism $\phi \in \Phi(\mu,\rho)$ and subsequently mapping ρ onto ν through a transport map $\widetilde{T} \in \widetilde{\mathcal{T}}(\rho,\nu)$, i.e, $\mathcal{X}_\mu \xrightarrow{\phi} \mathcal{Z}_\rho \xrightarrow{\widetilde{T}} \mathcal{Y}_\nu$. This alternate mapping can be expressed, naturally, as the composition $\widetilde{T} \circ \phi : \mathcal{X} \longrightarrow \mathcal{Y}$. It is straightforward to show that since $\mathcal{X}_\mu \cong^s \mathcal{Z}_\rho$, i.e $\phi\sharp\mu = \rho$, we have $(\widetilde{T} \circ \phi)\sharp\mu = \widetilde{T}\sharp(\phi\sharp\mu) = \widetilde{T}\sharp\rho = \nu$. As such, we can define the collection of all such composition maps as:

$$\mathcal{I}(\mu,\nu) = \{T : \mathcal{X} \to \mathcal{Y} \mid T\sharp\mu = \nu,$$
$$T \triangleq \widetilde{T} \circ \phi, \ \phi \in \Phi(\mu,\rho), \ \widetilde{T} \in \widetilde{\mathcal{T}}(\rho,\nu)\} \qquad (7)$$

where $\mathcal{I}(\mu,\nu) \subseteq \mathcal{T}(\mu,\nu)$. Since we restrict the collection of transport maps between the source and target space to only those that can be decomposed as in eq. (7), the proposed GM problem becomes a *constrained* version of eq. (3):

$$\mathrm{CGM}_p(\mu,\nu) = \inf_{T \in \mathcal{I}(\mu,\nu)} \mathrm{dis}_p(T) \qquad (8)$$

where $\mathrm{dis}_p(T)$ is given by eq. (2). Note that since the isomorphism $\phi \in \Phi(\mu,\rho)$ is already a solution (albeit not necessarily unique) to the $\mathrm{GM}(\mu,\rho) = 0$ problem, it is considered *fixed*. As such, we do not need to use a double infinum in eq. (8). In other words, given *any* isomorphism between the source and target space, we only need to optimize over the set $\widetilde{\mathcal{T}}(\rho,\nu)$, i.e finding the GM-optimal map $\widetilde{T}$ what pushes μ onto $\phi\sharp\mu$. To this end, eq. (8) can be re-written as:

$$\mathrm{CGM}_p(\mu,\nu) = \inf_{\widetilde{T}\sharp\rho=\nu} \mathrm{dis}_p(\widetilde{T} \circ \phi) \qquad (9)$$

where $\mathrm{dis}_p(\widetilde{T} \circ \phi) = \|c_{\mathcal{X}} - c_{\mathcal{Y}}\|_{L^p(\mu_{\widetilde{T}\circ\phi} \otimes \mu_{\widetilde{T}\circ\phi})}$ with $\mu_{\widetilde{T}\circ\phi} = (\mathrm{Id}, \widetilde{T} \circ \phi)$. Then, using Proposition 3.1 we show the following result.

Proposition 3.2. *Assume problem eq. (3) admits at least one solution, i.e $\exists \ T^* \in \arg\min_{T \in \mathcal{T}(\mu,\nu)} dis_p(T)$. Then, given a reference space s.t $\mathcal{X}_\mu \cong^s \mathcal{Z}_\rho$ and any isomorphism $\phi \in \Phi(\mu,\rho)$, there exists $\widetilde{T}^* \in \arg\min_{\widetilde{\mathcal{T}}(\rho,\nu)} dis_p(\widetilde{T})$ s.t the composition map $\widetilde{T}^* \circ \phi$ is also a solution to eq. (3). That is, the optimization problems eq. (3) and eq. (9) are equivalent, i.e, $GM_p(\mu,\nu) = CGM_p(\mu,\nu)$.*

We provide the proof of Proposition 3.2 in Appendix A.2.

Proposition 3.2, shows that we can decompose the search for an optimal solution $T^* \in \mathcal{T}(\mu, \nu)$ to $\mathrm{GM}_p(\mu, \nu)$ into the search for an isomorphism $\phi \in \Phi(\mu, \rho)$ and a map $\widetilde{T}^* \in \widetilde{\mathcal{T}}(\rho, \nu)$. It shows that we can effectively break the complex problem of optimally transporting one measure onto another into two structured sub-problems, without losing GM-optimality. Note that, assuming non-unique optimality, the above result does not necessarily mean that *every* GM-optimal map of eq. (3) can be decomposed in this manner but that at least *some* are, which is sufficient for our analysis.

Geometric Intuition. From a geometric perspective, we can interpret Proposition 3.2 as follows: Any exact transportation of probability mass that aims to minimize distortion from one mm-space onto another, can be represented as a sequence of two geometric transformations. An isomorphic transformation, which in Euclidean space includes rotations reflections and translations and a geometric "deformation" which represents the distortion of the initial geometry of the source metric space, e.g shearing. This two-stage process highlights the interplay in GM-optimal transport between maintaining intrinsic geometric properties and adapting to new probabilistic configurations.

4. Isomorphism Invariant Neural Gromov-Monge

In this section, we introduce our neural framework for approximating the GM-optimal composition map $\widetilde{T}^* \circ \phi$ in Proposition 3.2 for $p = 2$. We use neural networks $\phi_\omega : \mathcal{X} \longrightarrow \mathcal{Z}$ and $\widetilde{T}_\theta : \mathcal{Z} \longrightarrow \mathcal{Y}$ to parameterize ϕ and $\widetilde{T}^*$ respectively. For convenience, we will refer to ϕ_ω as the *isomorphism network* and to $\widetilde{T}_\theta$ as the *transport network*. To this end, in section 4.1, we prove the following theoretical result: that the composition of neural networks $\widetilde{T}_\theta \circ \phi_\omega$ can approximate *any* transport map between a source and a target measure. Subsequently, in section 2.2, we present our proposed learning procedure for approximating GM-optimal maps, s.t $\widetilde{T}_\theta \circ \phi_\omega \approx \widetilde{T}^* \circ \phi$.

4.1. Neural Network Compositions as Universal Approximators of Transport Maps

Our analysis is inspired by Theorem 1 in Korotin et al. (2023), which states that neural networks can approximate any stochastic transport map in the L^2 norm.

Transport maps in the L^p space. Note that the set of transport *maps* between mm-spaces can include both continuous *and* discontinuous functions. More formally, let $L^2_\mu(\mathcal{X}, \mathcal{Y})$ be the space of quadratically integrable functions w.r.t μ, i.e $\{f : \mathcal{X} \longrightarrow \mathcal{Y} \mid \|f\|_{L^2(\mu)} < +\infty\}$, which is known to include both continuous and irregular functions. Assuming μ has a finite second moment, it can be shown that, for any

transport map T from $\mathcal{X}_\mu$ to $\mathcal{Y}_\nu$, we have $T \in L^2_\mu(\mathcal{X}, \mathcal{Y})$ (Korotin et al., 2023). As such, given a neural network ϕ_ω which is *specifically modeled to approximate an isometry*, we aim to show that $\widetilde{T}_\theta \circ \phi_\omega$ is dense in the function space $L^2_\mu(\mathcal{X}, \mathcal{Y})$. We base our analysis on the result by Kratsios & Bilokopytov (2020) who show that the composition of a continuous injective map ϕ with a ReLU neural network is dense in the space of continuous functions. In our case, since ϕ is an isomorphism, it inherently satisfies both continuity and injectivity (note that an isomorphism is also bijective). We present our result in the following Theorem.

Theorem 4.1. *Let $T \in \mathcal{T}(\mu, \nu)$ be a transport map between μ and ν, where ν has finite second moment. Then, there exists a feed-forward ReLU neural network $\widetilde{T}_\theta : \mathcal{Z} \longrightarrow \mathcal{Y}$ and a neural network $\phi_\omega : \mathcal{X} \longrightarrow \mathcal{Z}$ with any nonaffine continuous activation function which is continuously differentiable at at least one point, such that, $\forall \epsilon > 0$ we have:*

$$\|T - \widetilde{T}_\theta \circ \phi_\omega\|_{L^2_\mu} \leq \epsilon \tag{10}$$

We provide the full proof of Theorem 4.1 in Appendix A.3.

Intuition behind Theorem 4.1. In Kratsios & Bilokopytov (2020), they consider ϕ to be a (continuous and injective) *feature map* from a (possibly) non-Euclidean to a Euclidean feature space. Their result shows that the presence of such a map, in the initial layer of an architecture, does not compromise its approximation capabilities. Our result can be viewed in a similar fashion. An isomorphism $\phi \in \Phi(\mu, \nu)$ can be viewed as a 'feature map' from the source $\mathcal{X}_\mu$ to the reference space $\mathcal{Z}_\rho$. To this end, it can be interpreted as a transformation that re-configures the source space into a "canonical form", i.e, a standardized representation that retains the original geometric properties. Additionally, Theorem 4.1 shows that if the true map ϕ is not available, we can approximate it by the neural network ϕ_ω and still retain the overall architecture's universal approximation properties.

Note that, approximating $\widetilde{T}^* \circ \phi$ through $\widetilde{T}_\theta \circ \phi_\omega$, essentially entails the individual approximation of two GM-optimal maps $\widetilde{T}^*$ and ϕ. Recall, that an isomorphism $\phi \in \Phi(\mu, \rho)$ is, itself, a solution to a GM problem, i.e $\mathrm{GM}(\mu, \rho)$. Thus, we only need to define a single learning procedure for learning GM-optimal maps, which can be used for both $\widetilde{T}_\theta$ and ϕ_ω, adapted to their respective domains.

4.2. Learning GM-optimal maps

In this section, we start by introducing a general learning framework to approximate *any* GM-optimal map between two arbitrary mm-spaces. Subsequently, based on the aforementioned framework, we introduce our learning procedure for training networks $\widetilde{T}_\theta$ and ϕ_ω. We base our method on the work by Uscidda & Cuturi (2023), who propose a loss for approximating OT maps between comparable spaces, i.e,

Monge maps. Here, we briefly summarize their contribution and how our work is an extension of their framework to the GM case.

The Monge Gap regularizer (Uscidda & Cuturi, 2023). Note that, the Monge problem between a source $\mu \in \mathcal{P}(\Omega)$ and target $\nu \in \mathcal{P}(\Omega)$ space, supported on the same domain Ω, is defined as (Monge, 1781):

$$M_c(\mu, \nu) = \inf_{T\#\mu=\nu} \int_\Omega c(\mathbf{x}, T(\mathbf{x}))d\mu \quad (11)$$

where $c : \Omega \times \Omega \longrightarrow \mathbb{R}_+$ is an intra-domain cost. Given any map T which is used to approximate a solution to eq. (11), Uscidda & Cuturi (2023) propose a loss function which quantifies its deviation from Monge optimality. Their loss consists of two terms: a *fitting loss* and a *regularizer* referred to as the *Monge Gap*. The fitting loss quantifies how well T satisfies the constraint $T\sharp\mu = \nu$. Conversely, the Monge gap is defined as (Uscidda & Cuturi, 2023):

$$\mathcal{M}(\mu, T\sharp\mu) = \int_\Omega c(\mathbf{x}, T(\mathbf{x}))d\mu - M_c(\mu, T\#\mu) \quad (12)$$

In essence, eq. (12) quantifies the deviation of T from being the c-optimal map between μ and $T\sharp\mu$, i.e, the map which induces the least amount of c cost, while performing the push-forward operation $T\sharp\mu$. Naturally, when $T\sharp\mu = \nu$ is satisfied, i.e, when the fitting loss is zero, eq. (12) will quantify the deviation of T from the solution of eq. (11).

The Gromov-Monge Gap regularizer. We can extend the above elegant framework to the Gromov-Monge problem. For a given map $T : \mathcal{X} \longrightarrow \mathcal{Y}$, we define its induced *Gromov-Monge (GM) gap* as follows:

$$\mathcal{GM}^p(\mu, T\sharp\mu) = \text{dis}_p(T)^p - \text{GM}_p^p(\mu, T\sharp\mu) \quad (13)$$

where $\text{GM}_p^p(\mu, T\sharp\mu)$ is the p-GM distance between measures μ and $T\sharp\mu$. Note that eq. (13) is a direct extension of eq. (12) to incomparable spaces. Intuitively, the first term represents the distortion induced by T when moving mass from μ to $T\sharp\mu$. The second term represents the distortion induced by the GM-optimal map between the measures μ and $T\sharp\mu$. Thus, as in the Monge gap case, when $\mathcal{GM}^p(\mu, T\sharp\mu) = 0$ and $T\sharp\mu = \nu$, T is theoretically guaranteed to be the solution T^* to eq. (3). Consequently, we can utilise the GM gap of eq. (13) as a regularizer in a loss function, designed to recover GM-optimal maps.

GM-optimality loss. Let T_θ be any parameterized map. Similar to the Monge-optimality loss proposed in Uscidda & Cuturi (2023), we propose the GM-optimality loss defined as:

$$\mathcal{L}(\theta) := \Delta(T_\theta\sharp\mu, \nu) + \lambda \, \mathcal{GM}^p(\mu, T_\theta\sharp\mu) \quad (14)$$

where $\Delta(T_\theta\sharp\mu, \nu)$ is the fitting loss and λ is a regularization weight to stabilize training. We can use any valid discrepancy between measures on comparable spaces, such as the Sinkhorn divergence (Genevay et al., 2019), i.e, $S_{c,\varepsilon}(\mu, T\sharp\mu)$ or the entropic Wasserstein distance, i.e, $W_{c,\varepsilon}(\mu, T\sharp\mu)$. Naturally, we get $\mathcal{L}(\theta) = 0$ when both terms of eq. (14) are 0, i.e, when both probability mass and inter-domain distances are preserved. As such, for any family of parameterized maps $\{T_\theta\}_\theta$ between two mm-spaces, we can recover the GM-optimal map by solving the optimization problem $T_\theta := \arg\min_\theta \mathcal{L}(\theta)$.

Optimization procedure. Following the tripod structure of Figure 1, the networks ϕ_ω and $\widetilde{T}_\theta$ should be trained such that:

1. ϕ_ω moves mass from μ to ρ in a GM-optimal way.

2. $\widetilde{T}_\theta$ pushes forward the measure transformed by ϕ_ω, i.e, $\rho' := \phi_\omega\sharp\mu$, onto ν in a GM-optimal way.

Based on the GM-optimality loss of eq. (14), condition 1. can be formulated as the following optimization problem:

$$\phi_\omega := \arg\min_\omega \Delta(\phi_\omega\sharp\mu, \rho) + \lambda \, \mathcal{GM}^2(\mu, \phi_\omega\sharp\mu) \quad (15)$$

where $\mathcal{GM}^2(\mu, \phi_\omega\sharp\mu) = \text{dis}_2(\phi_\omega)^2 - \text{GM}_2^2(\mu, \phi_\omega\sharp\mu)$, with $\text{dis}_2(\phi_\omega)^2 := \|c_\mathcal{X} - c_\mathcal{Z}\|_{L^2(\mu_{\phi_\omega} \otimes \mu_{\phi_\omega})}$. Analogously, condition 2. can be formulated as:

$$\widetilde{T}_\theta := \arg\min_\theta \Delta(\widetilde{T}_\theta\sharp\rho', \nu) + \lambda \, \mathcal{GM}^2(\rho', \widetilde{T}_\theta\sharp\rho') \quad (16)$$

where $\mathcal{GM}^2(\rho', \widetilde{T}_\theta\sharp\rho') = \text{dis}_2(\widetilde{T}_\theta)^2 - \text{GM}_2^2(\rho', \widetilde{T}_\theta\sharp\rho')$, with $\text{dis}_2(\widetilde{T}_\theta)^2 := \|c_\mathcal{Z} - c_\mathcal{Y}\|_{L^2(\rho'_{\widetilde{T}_\theta} \otimes \rho'_{\widetilde{T}_\theta})}$. Given the theoretical properties of the GM-optimality loss and GM-gap, achieving zero in the objective functions of eq. (15) and eq. (16) theoretically ensures that their solutions, ϕ_ω and $\widetilde{T}_\theta$, are the GM-optimal maps ϕ and $\widetilde{T}^*$, respectively. Consequently, based on Theorem 4.1 and Proposition 3.1, it will hold that $\widetilde{T}_\theta \circ \phi_\omega \approx \widetilde{T}^* \circ \phi$, thus approximating a GM-optimal map between $\mathcal{X}_\mu$ and $\mathcal{Y}_\nu$. We present the detailed learning procedure in Algorithm 1.

Estimation from samples. In practice, we have access to finite sets of samples, assumed to be drawn iid from the underlying, true, continuous distributions. As such, we consider the following sets consisting of n number of samples: $\{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \overset{\text{iid}}{\sim} \mu$, $\{\mathbf{y}_i\}_{i=1}^n$, $\mathbf{y}_i \overset{\text{iid}}{\sim} \nu$ and $\{\mathbf{z}_i\}_{i=1}^n$, $\mathbf{z}_i \overset{\text{iid}}{\sim} \rho$. For the corresponding push-forward measures we have: $\{\phi_\omega(\mathbf{x}_i)\}_{i=1}^n$, $\phi_\omega(\mathbf{x}_i) \overset{\text{iid}}{\sim} \phi_\omega\sharp\mu \in \mathcal{P}(\mathcal{Z})$ and $\{\widetilde{T}_\theta(\mathbf{z}'_i)\}_{i=1}^n$, $\widetilde{T}_\theta(\mathbf{z}'_i) \overset{\text{iid}}{\sim} \widetilde{T}_\theta\sharp\rho' \in \mathcal{P}(\mathcal{Y})$, where $\mathbf{z}'_i = \phi_\omega(\mathbf{x}_i) \overset{\text{iid}}{\sim} \phi_\omega\sharp\mu = \rho'$. The discrete empirical estimates of each corresponding (continuous) measure will,
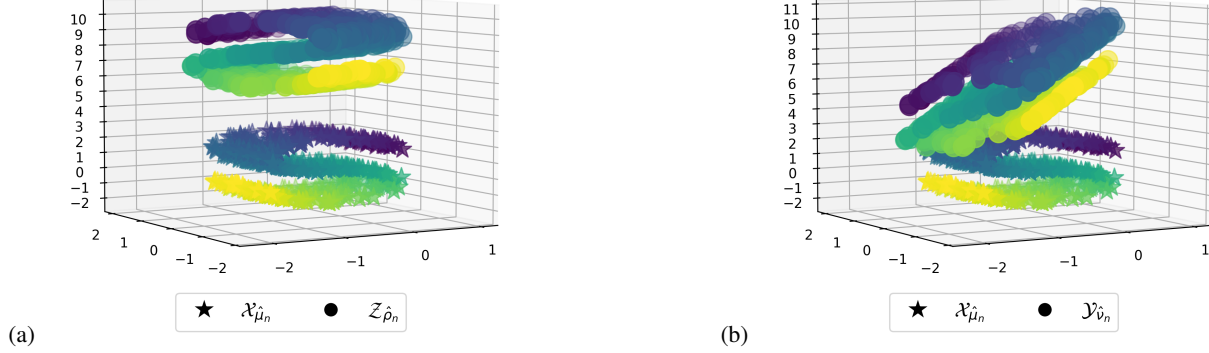
Figure 2. (a) **Isomorphic transformation of the source space.** Depiction of the empirical source and reference spaces $\mathbf{X} := \mathcal{X}_{\hat{\mu}_n}$ and $\mathbf{Z} := \mathcal{Z}_{\hat{\rho}_n}$ respectively. We obtain $\mathbf{Z}$ through the rigid transformation of $\mathbf{X}$, such that $\mathcal{X}_{\hat{\mu}_n} \cong^s \mathcal{Z}_{\hat{\rho}_n}$. To this end, we have $\mathbf{Z} = \phi(\mathbf{X}) := \mathbf{R}\mathbf{X} + \mathbf{t}$, where $\mathbf{R} \in \{\mathbf{R} \in \mathbb{R}^{3 \times 3} \mid \mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = \mathbf{I}\}$ is an orthogonal (rotation) matrix and $\mathbf{t} \in \mathbb{R}^3$ a translation vector. (b) **Non-Isomorphic transformation of the reference space.** Depiction of the empirical source and target spaces $\mathbf{X} := \mathcal{X}_{\hat{\mu}_n}$ and $\mathbf{Y} := \mathcal{Y}_{\hat{\nu}_n}$, respectively. The target point cloud $\mathbf{Y}$ is obtained through a non-rigid transformation of the reference space $\mathbf{Z}$, i.e., $\mathbf{Y} = \widetilde{T}^*(\mathbf{Z}) = \mathbf{A}\mathbf{Z}$, where $\mathbf{A} \in \mathbb{R}^{3 \times 3}$ is a shearing matrix. Unlike ϕ, the transport map $\widetilde{T}^*$ introduces a distortion of the spacial configuration of its input.

therefore, be: $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$, $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{y}_i}$, $\hat{\rho}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{z}_i}$, $\phi_\omega\sharp\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\phi_\omega(\mathbf{x}_i)} = \hat{\rho}'_n$ and $\widetilde{T}_\theta\sharp\hat{\rho}'_n = \frac{1}{n}\sum_{i=1}^n \delta_{\widetilde{T}_\theta(\mathbf{z}'_i)}$. Given the above measure estimates, we can re-write each term in eq. (15) and eq. (16) in a discrete formulation. Specifically, we can define the empirical estimate of the GM-gap $\mathcal{GM}^2(\mu, \phi_\omega\sharp\mu)$ as:

$$\mathcal{GM}^2(\hat{\mu}_n, \phi_\omega\sharp\hat{\mu}_n) = \hat{\text{dis}}_2(\phi_\omega)^2 - \text{GW}_2^2(\hat{\mu}_n, \phi_\omega\sharp\hat{\mu}_n) \tag{17}$$

where,
$\hat{\text{dis}}_2(\phi_\omega)^2 = \frac{1}{n^2}\sum_{i,j}^n \left(c_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) - c_{\mathcal{Z}}(\phi_\omega(\mathbf{x}_i), \phi_\omega(\mathbf{x}_j))\right)^2$.
Analogously, for $\mathcal{GM}^2(\rho', \widetilde{T}_\theta\sharp\rho')$ we have:

$$\mathcal{GM}^2(\hat{\rho}'_n, \widetilde{T}_\theta\sharp\hat{\rho}'_n) = \hat{\text{dis}}_2(\widetilde{T}_\theta)^2 - \text{GW}_2^2(\hat{\rho}'_n, \widetilde{T}_\theta\sharp\hat{\rho}'_n) \tag{18}$$

where,
$\hat{\text{dis}}_2(\widetilde{T}_\theta)^2 = \frac{1}{n^2}\sum_{i,j}^n \left(c_{\mathcal{Z}}(\mathbf{z}_i, \mathbf{z}_j) - c_{\mathcal{Y}}(\widetilde{T}_\theta(\mathbf{z}_i), \widetilde{T}_\theta(\mathbf{z}_j))\right)^2$.

In both cases we use the GW distance with entropic regularization (Peyré et al., 2016; Solomon et al., 2016) to approximate the GM distances $\text{GM}_2^2(\mu, \phi_\omega\sharp\mu)$ and $\text{GM}_2^2(\rho', \widetilde{T}_\theta\sharp\rho')$ respectively. Note that according to Theorem 2 in Mémoli & Needham (2022a), the GW distance is equivalent to the GM only for *non-atomic* measures. Here we make the assumption that given enough samples n, the discrete entropic GW will approximate its continuous non-atomic counterpart and subsequently the corresponding GM distance.

Algorithm 1 GM composition map estimation

1: **Data:** Source μ, target ν and reference ρ measures accessible through empirical estimates $\hat{\mu}_n, \hat{\rho}_n$ and $\hat{\nu}_n$; transport network $\widetilde{T}_\theta$; isomorphism network ϕ_ω; cost functions $c_{\mathcal{X}}, c_{\mathcal{Y}}$ and $c_{\mathcal{Z}}$; regularization weight λ; entropic regularization parameter ε; learning rate η; batch size n, number of iterations K_{outer} and K_{inner};
2: **Output:** Estimated GM-optimal map $T := \widetilde{T}_\theta \circ \phi_\omega$.
3: **for** $k = 1, \ldots, K_{\text{outer}}$ **do**
4:     Sample batches $\hat{\mu}_n, \hat{\rho}_n$
5:     Compute fitting loss $\Delta(\phi_\omega\sharp\hat{\mu}_n, \hat{\rho}_n)$
6:     Compute regularizer $\mathcal{GM}^2(\hat{\mu}_n, \phi_\omega\sharp\hat{\mu}_n)$
7:     $\hat{L}_\phi(\omega) \leftarrow \Delta(\phi_\omega\sharp\hat{\mu}_n, \hat{\rho}_n) + \lambda\,\mathcal{GM}^2(\hat{\mu}_n, \phi_\omega\sharp\hat{\mu}_n)$
8:     Update ω using $\frac{\partial \hat{L}_\phi(\omega)}{\partial \omega}$
9:     **for** $k = 1, \ldots, K_{\text{inner}}$ **do**
10:         Re-sample batch $\hat{\mu}_n$
11:         Set $\hat{\rho}'_n \leftarrow \phi_\omega\sharp\hat{\mu}_n$
12:         Sample batch $\hat{\nu}_n$
13:         Compute fitting loss $\Delta(\widetilde{T}_\theta\sharp\hat{\rho}'_n, \hat{\nu}_n)$
14:         Compute regularizer $\mathcal{GM}^2(\hat{\rho}'_n, \widetilde{T}_\theta\sharp\hat{\rho}'_n)$
15:         $\hat{L}_T(\theta) \leftarrow \Delta(\widetilde{T}_\theta\sharp\hat{\rho}'_n, \hat{\nu}_n) + \lambda\,\mathcal{GM}^2(\hat{\rho}'_n, \widetilde{T}_\theta\sharp\hat{\rho}'_n)$
16:         Update θ using $\frac{\partial \hat{L}_T(\theta)}{\partial \theta}$
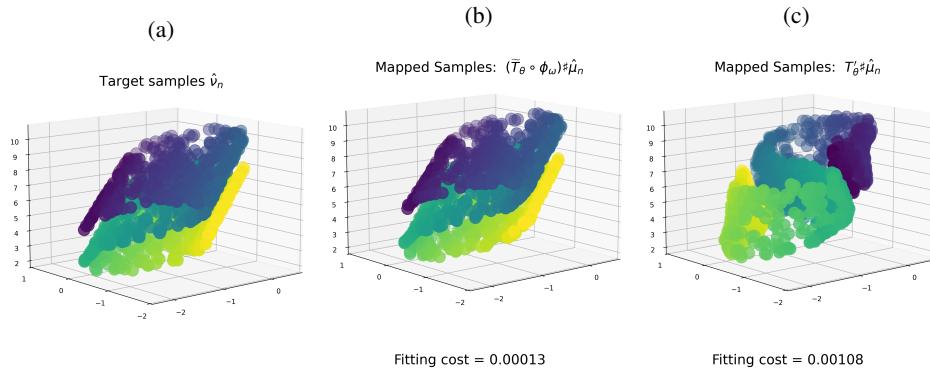17:   **end for**
18: **end for**

Figure 3. (a) **Ground truth target samples.** (b) **Mapped samples using neural composition.** Fitted samples using the learned composition map, i.e, $(\widetilde{T}_\theta \circ \phi_\omega)\sharp\hat{\mu}_n$. We first train ϕ_ω for 5,000 iterations as a direct map between the source and reference samples. Then, using the pre-trained ϕ as initialization, we train the composition map $\widetilde{T}_\theta \circ \phi_\omega$, following Algorithm 1 for $K_{outer} = 5$ and $K_{inner} = 2,000$ iterations. (c) **Mapped samples using direct parameterization.** Fitted samples using a direct map from the source to the target space $T'_\theta\sharp\hat{\mu}_n$ trained for 5,000 iterations. Unlike the direct map, the composition map clearly resembles the target distribution. We use the squared Euclidean Sinkhorn divergence, i.e, $S_{l_2^2,\varepsilon}$, with entropic regularization strength $\varepsilon = 0.1$ as the quantitative evaluation metric (i.e, evaluation fitting cost). We provide all details of our experimental framework in Appendix B.

5. Experiments

In this section, we evaluate the ability of our method to recover GM-optimal maps on synthetic data. We implement our framework using the OTT-JAX[2] package (Cuturi et al., 2022).

Experimental tripod structure. We consider a controlled setting, where the GM-optimal maps ϕ and $\widetilde{T}^*$ are known. We generate samples $\{\mathbf{x}_i\}_{i=1}^n$ in $\mathbb{R}^3$, following the associated empirical measure $\hat{\mu}_n$ as described in section 4.2. As such, we define the *empirical* source mm-space $\mathcal{X}_{\hat{\mu}_n}$ as the 3D point cloud $\mathbf{X} := \mathcal{X}_{\hat{\mu}_n} \in \mathbb{R}^{n\times3}$. Following the tripod structure of Figure 1, we apply a known isomorphism, i.e a *rigid* transformation ϕ to $\mathcal{X}_{\hat{\mu}_n}$, to obtain the (empirical) reference mm-space, i.e, $\mathbf{Z} = \phi(\mathbf{X}) := \mathcal{Z}_{\hat{\rho}_n} \in \mathbb{R}^{n\times3}$ (Figure 2 (a)). Subsequently, we generate the empirical target space by applying a *non-rigid* transformation, $\widetilde{T}^*$, to the reference point-cloud $\mathbf{Z}$, i.e $\mathbf{Y} = \widetilde{T}^*(\mathbf{Z}) := \mathcal{Y}_{\hat{\nu}_n} \in \mathbb{R}^{n\times3}$ (Figure 2 (b)).

Results. Following Algorithm 1, we train neural networks ϕ_ω and $\widetilde{T}_\theta$ on the empirical sample sets. Since the target point cloud $\mathbf{Y}$ is the result of the GM-optimal transformation $\widetilde{T}^* \circ \phi$, if $\widetilde{T}_\theta \circ \phi_\omega$ approximates GM-optimality, we expect the mapped points to approximate $\mathbf{Y}$, i.e $\widetilde{T}_\theta \circ \phi_\omega(\mathbf{X}) \approx \mathbf{Y}$. We present the mapped samples of the learned composition in Figure 3 (b), and compare them to the ground truth target samples (Figure 3 (a)). Indeed, the learned map $\widetilde{T}_\theta \circ \phi_\omega$ is able to fully match the ground truth target point cloud. To validate the importance of learning a composition instead of a direct approximation, we present the equivalent results when learning a direct mapping $T'_\theta : \mathcal{X}_{\hat{\mu}_n} \longrightarrow \mathcal{Y}_{\hat{\nu}_n}$,

from the source to the target point cloud (Figure 3 (c)). In this case, we train $T'_\theta : \mathcal{X}_{\hat{\mu}_n}$ on a single loop using the loss $\hat{L}_{T'}(\theta) \leftarrow \Delta(T'_\theta\sharp\hat{\mu}_n, \hat{\nu}_n) + \lambda\,\mathcal{GM}^2(\hat{\mu}_n, T'_\theta\sharp\hat{\mu}_n)$. Unlike the composition map, the direct mapping is unable to accurately match the geometry of the target distribution.

6. Discussion

In this work, we introduce a theoretically grounded framework to approximate optimal transport maps between incomparable spaces, based on the geometric property of invariance to strong isomorphism. We show that the GM-optimal map between a source and target distribution can be decomposed into an isomorphism and a subsequent GM-optimal map, which can be adequately approximated by two separate neural networks. Intuitively, this implies the hypothesis that any geometrically informed transformation between two distributions is a sequence of a rigid and a non-rigid transformation. This suggests that by parameterizing (and training) each transformation step independently, we more effectively approximate the true optimal transportation map between incomparable distributions. Empirical results on toy data reinforce this hypothesis. We believe this could open a new avenue to address the fundamental problem of geometrically informed transformations of samples between distributions.

References

Alvarez-Melis, D. and Fusi, N. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

Alvarez-Melis, D., Jegelka, S., and Jaakkola, T. S. Towards

optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1870–1879. PMLR, 2019.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Brenier, Y. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. Learning generative models across incomparable spaces. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 851–861. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/bunne19a.html.

Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023.

Chowdhury, S. and Mémoli, F. The Gromov-Wasserstein distance between networks and stable network invariants. *arXiv e-prints*, art. arXiv:1808.04337, August 2018. doi: 10.48550/arXiv.1808.04337.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. Scot: single-cell multi-omics alignment with optimal transport. *Journal of computational biology*, 29 (1):3–18, 2022.

Dumont, T., Lacombe, T., and Vialard, F.-X. On the existence of monge maps for the gromov–wasserstein problem. *Foundations of Computational Mathematics*, pp. 1–48, 2024.

Fan, J., Liu, S., Ma, S., Zhou, H., and Chen, Y. Neural monge map estimation and its applications. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Fickinger, A., Cohen, S., Russell, S., and Amos, B. Cross-domain imitation learning via optimal transport. *arXiv preprint arXiv:2110.03684*, 2021.

Folland, G. B. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences, 2019.

Kantorovich, L. V. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Klein, D., Uscidda, T., Theis, F., and Cuturi, M. Entropic (gromov) wasserstein flow matching with genot, 2024.

Korotin, A., Egiazarian, V., Asadulaev, A., Safin, A., and Burnaev, E. Wasserstein-2 generative networks, 2020.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Kratsios, A. and Bilokopytov, I. Non-euclidean universal approximation. *Advances in Neural Information Processing Systems*, 33:10635–10646, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Makkuva, A. V., Taghvaei, A., Oh, S., and Lee, J. D. Optimal transport mapping via input convex neural networks, 2020.

Mémoli, F. and Needham, T. Distance distributions and inverse problems for metric measure spaces. *Studies in Applied Mathematics*, 149(4):943–1001, 2022.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Mémoli, F. Gromov-wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. URL http://dblp.uni-trier.de/db/journals/focm/focm11.html#Memoli11.

Mémoli, F. and Needham, T. Comparison results for gromov-wasserstein and gromov-monge distances, 2022a.

Mémoli, F. and Needham, T. Distance distributions and inverse problems for metric measure spaces, 2022b. URL https://arxiv.org/abs/1810.09646.

Nekrashevich, M., Korotin, A., and Burnaev, E. Neural gromov-wasserstein optimal transport. *arXiv preprint arXiv:2303.05978*, 2023.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rout, L., Korotin, A., and Burnaev, E. Generative modeling with optimal transport maps, 2022.

Salmona, A., Delon, J., and Desolneux, A. Gromov-wasserstein distances between gaussian distributions. *arXiv preprint arXiv:2104.07970*, 2021.

Santambrogio, F. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015. ISBN 9783319208282. URL https://books.google.ch/books?id=UOHHCgAAQBAJ.

Sebbouh, O., Cuturi, M., and Peyré, G. Structured transforms across spaces with cost-regularized optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 586–594. PMLR, 2024.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sturm, K.-T. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces, 2020.

Uscidda, T. and Cuturi, M. The monge gap: A regularizer to learn all transport maps, 2023.

Vayer, T. A contribution to optimal transport on incomparable spaces. *arXiv preprint arXiv:2011.04447*, 2020.

A. Proofs

A.1. Proof of Proposition 3.1

Since $\mathcal{X}_\mu \cong^s \mathcal{Z}_\rho$, $\exists\, \phi \in \Phi(\mu, \rho)$, which is a bijection. Thus, $\exists\, \phi^{-1} \in \Phi(\rho, \mu)$, s.t $GM(\rho, \mu) = 0$. Since GM defines a Lawvere metric on $\mathcal{M}_p$ (see Mémoli & Needham (2022b) and Theorem 1 in Mémoli & Needham (2022a)), following the triangle inequality we have $\mathrm{GM}(\mu, \nu) \leq \mathrm{GM}(\mu, \rho) + \mathrm{GM}(\rho, \nu)$ and $\mathrm{GM}(\rho, \nu) \leq \mathrm{GM}(\rho, \mu) + \mathrm{GM}(\mu, \nu)$. Combining the above inequalities and since $\mathrm{GM}(\mu, \rho) = 0$, we get $\mathrm{GM}(\mu, \nu) = \mathrm{GM}(\rho, \nu)$.

A.2. Proof of Proposition 3.2

Constraining the optimal transport plan to be within $\mathcal{I}(\mu, \nu)$, will result in a *constraint* GM-problem:

$$
\begin{aligned}
\mathrm{CGM}_p(\mu, \nu) = \inf_{T \in \mathcal{I}(\mu, \nu)} \left(\iint_{\mathcal{X} \times \mathcal{X}} |c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') - c_{\mathcal{Y}}(T(\mathbf{x}), T(\mathbf{x}'))|^p \, d\mu(\mathbf{x}) \, d\mu(\mathbf{x}') \right)^{1/p} = \\
\inf_{(\widetilde{T} \circ \phi) \# \mu = \nu} \left(\iint_{\mathcal{X} \times \mathcal{X}} |c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') - c_{\mathcal{Y}}((\widetilde{T} \circ \phi)(\mathbf{x}), (\widetilde{T} \circ \phi)(\mathbf{x}'))|^p \, d\mu(\mathbf{x}) \, d\mu(\mathbf{x}') \right)^{1/p}
\end{aligned}
\tag{19}
$$

Based on Definition 2.1, any isomorphism $\phi \in \Phi(\mu, \rho)$ is an invertible bijective map. Thus, for every pair $(\mathbf{z}, \mathbf{z}') \in \mathcal{Z} \times \mathcal{Z}$ there exists a unique pair $(\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X}$ such that $\mathbf{z} = \phi(\mathbf{x})$ and $\mathbf{z}' = \phi(\mathbf{x}')$. Additionally, $\phi^{-1} : \mathcal{Z} \longrightarrow \mathcal{X}$ is also a bijective isometry. Consequently, for every $\mathbf{z} \in \mathcal{Z}$ there is a unique $\mathbf{x} \in \mathcal{X}$ s.t $\mathbf{x} = \phi^{-1}(\mathbf{z})$. Also, since $\phi \# \mu = \rho$, it holds for every $B \subseteq \mathcal{Z}$ that $\rho(B) = \mu(\phi^{-1}(B))$. Considering all the above, we are permitted to perform a change of variables over the integral in eq. (19). Thus, we have:

$$
\begin{aligned}
\mathrm{CGM}_p(\mu, \nu) = \inf_{\widetilde{T} \# \rho = \nu} \left(\iint_{\mathcal{X} \times \mathcal{X}} |c_{\mathcal{Z}}(\phi(\mathbf{x}), \phi(\mathbf{x}')) - c_{\mathcal{Y}}((\widetilde{T} \circ \phi)(\mathbf{x}), (\widetilde{T} \circ \phi)(\mathbf{x}'))|^p \, d\mu(\mathbf{x}) \, d\mu(\mathbf{x}') \right)^{1/p} = \\
\inf_{\widetilde{T} \# \rho = \nu} \left(\iint_{\mathcal{Z} \times \mathcal{Z}} |c_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') - c_{\mathcal{Y}}(\widetilde{T}(\mathbf{z}), \widetilde{T}(\mathbf{z}'))|^p \, d\mu(\phi^{-1}(\mathbf{z})) \, d\mu(\phi^{-1}(\mathbf{z}')) \right)^{1/p} = \\
\inf_{\widetilde{T} \# \rho = \nu} \left(\iint_{\mathcal{Z} \times \mathcal{Z}} |c_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') - c_{\mathcal{Y}}(\widetilde{T}(\mathbf{z}), \widetilde{T}(\mathbf{z}'))|^p \, d\rho(\mathbf{z}) \, d\rho(\mathbf{z}') \right)^{1/p} = \\
\mathrm{GM}_p(\rho, \nu)
\end{aligned}
\tag{20}
$$

Then, following Proposition 3.1 we get $\mathrm{CGM}_p(\mu, \nu) = \mathrm{GM}_p(\rho, \nu) = \mathrm{GM}_p(\mu, \nu)$.

A.3. Proof of Theorem 4.1

Since $\mathcal{X}_\mu \cong^s \mathcal{Z}_\rho$, there exists a measure preserving isometry $\phi \in \Phi(\mu, \rho)$ between $\mathcal{X}$ and $\mathcal{Z}$. Since ϕ is an isometry it will be continuous and injective. According to Kratsios & Bilokopytov (2020), for any given continuous and injective function $\phi : \mathcal{X} \longrightarrow \mathcal{Z}$, the collection of functions $f \circ \phi \in C_c(\mathcal{X}, \mathcal{Y})$, where $f : \mathcal{Z} \longrightarrow \mathcal{Y}$ is a deep-forward ReLU neural network, will be dense in $C_c(\mathcal{X}, \mathcal{Y})$. Following Proposition 7.9 in Folland (1999), $C_c(\mathcal{X}, \mathcal{Y})$ is dense in $L^2_\mu(\mathcal{X}, \mathcal{Y})$. Consequently, the set of functions $f \circ \phi$ is also dense in $L^2_\mu(\mathcal{X}, \mathcal{Y})$. Following the Proof of Theorem 1 in Korotin et al. (2023), it is straightforward to show that since ν has a finite second moment, for any transport $T \in \mathcal{T}(\mu, \nu)$ map between μ and ν we have $T \in L^2_\mu(\mathcal{X}, \mathcal{Y})$.

Thus, there exists a ReLU neural network f_θ s.t for any $\epsilon > 0$ we have:

$$
\|T - f_\theta \circ \phi\|_{L^2_\mu} \leq \epsilon/2
\tag{21}
$$

It holds that $f_\theta \circ \phi, f_\theta \circ \phi_\omega \in C_c(\mathcal{X}, \mathcal{Y})$ and $\phi, \phi_\omega \in C_c(\mathcal{X}, \mathcal{Z})$. Continuous functions with compact support are also p-integrable w.r.t any finite measure in their domain. Thus it holds: $f_\theta \circ \phi, f_\theta \circ \phi_\omega \in L^2_\mu(\mathcal{X}, \mathcal{Y})$ and $\phi, \phi_\omega \in L^2_\mu(\mathcal{X}, \mathcal{Z})$.

Therefore, the norms $\|T - f_\theta \circ \phi_\omega\|_{L^2_\mu}$, $\|f_\theta \circ \phi - f_\theta \circ \phi_\omega\|_{L^2_\mu}$ and $\|\phi - \phi_\omega\|_{L^2_\mu}$ are well-defined and for any T and ϕ_ω we have:

$$
\begin{aligned}
\|T - f_\theta \circ \phi_\omega\|_{L^2_\mu} &= \|T - f_\theta \circ \phi + f_\theta \circ \phi - f_\theta \circ \phi_\omega\|_{L^2_\mu} \\
&\leq \|T - f_\theta \circ \phi\|_{L^2_\mu} + \|f_\theta \circ \phi - f_\theta \circ \phi_\omega\|_{L^2_\mu}
\end{aligned}
\tag{22}
$$

where the inequality follows from Minkowski's inequality.

We will now focus on deriving a bound for the term $\|f_\theta \circ \phi - f_\theta \circ \phi_\omega\|_{L^2_\mu}$ in inequality 22.

Every feed-forward ReLU neural network is Lipschitz continuous w.r.t the L_2 norm. Thus, $\exists$ a constant $L > 0$ s.t $\forall z, z' \in \mathcal{Z}$:

$$
\|f_\theta(z) - f_\theta(z')\|_2 \leq L\|z - z'\|_2
\tag{23}
$$

We have $\phi_\omega : \mathcal{X} \longrightarrow \mathcal{Z}$ and $\phi : \mathcal{X} \longrightarrow \mathcal{Z}$, being two distinct mappings with the same domain and codomain.

Thus, $\forall x \in \mathcal{X}$ we have $\phi(x), \phi_\omega(x) \in \mathcal{Z}$. Note that it doesn't necessarily hold that $\phi(x) = \phi_\omega(x)$. As such, we can assign $z = \phi(x)$ and $z' = \phi_\omega(x)$ and based on inequlity 23 we get:

$$
\|f_\theta \circ \phi(x) - f_\theta \circ \phi_\omega(x)\|_2^2 \leq L^2 \|\phi(x) - \phi_\omega(x)\|_2^2
\tag{24}
$$

Integrating both sides of inequality 24 over $\mathcal{X}$ w.r.t μ we get: (note that integration is permitted since we have already established that all functions are p-integrable w.r.t. μ)

$$
\int_{x \in \mathcal{X}} \|f_\theta \circ \phi_\omega(x) - f_\theta \circ \phi(x)\|_2^2 \, d\mu(x) \leq L^2 \int_{x \in \mathcal{X}} \|\phi(x) - \phi_\omega(x)\|_2^2 \, d\mu(x)
\tag{25}
$$

which can be written as:

$$
\|f_\theta \circ \phi_\omega - f_\theta \circ \phi\|_{L^2_\mu} \leq L^2 \|\phi_\omega - \phi\|_{L^2_\mu}
\tag{26}
$$

Note that the above inequality, as well as inequality 22, holds for any neural network $\phi_\omega : \mathcal{X} \longrightarrow \mathcal{Z}$ with the above-mentioned activation functions. In the Proof of Theorem 1 in Korotin et al. (2023), it is shown that neural networks of the aforementioned form will be dense in $L^2_\mu(\mathcal{X}, \mathcal{Z})$. As such, for convenience, we can choose a bound $\epsilon/2L^2$ s.t for any ϵ, there exists a neural network ϕ_ω s.t:

$$
\|\phi_\omega - \phi\|_{L^2_\mu} \leq \epsilon/2L^2
\tag{27}
$$

Combining inequalities 26, 27 and 22 we get:

$$
\|T - f_\theta \circ \phi_\omega\| \leq \epsilon
\tag{28}
$$

which holds for any $\epsilon > 0$, thus concluding the proof.

B. Experimental details

Neural networks. We use vanilla MLPs with hidden sizes [128, 64, 64] and a ReLU activation function for all perametrizations, i.e for all $\phi_\omega, \widetilde{T}_\theta, T'_\theta$. Specifically for the isomorphism network ϕ_ω we add a residual connection from its first to its last layer. Inspired by the initialization scheme in Uscidda & Cuturi (2023), this approach encourages the network to learn an affine transformation between the source and reference samples. We initialize all networks using the orthogonal initialization, i.e which uses uniformly distributed orthogonal matrices. We train ϕ_θ with a learning rate of $\eta = 10^{-3}$ and all

other networks with $\eta = 10^{-4}$. In all cases, we train our neural networks using the ADAM optimizer (Kingma & Ba, 2017), with a batch size of $n = 1024$.

Loss functions. For the training of ϕ_ω, both during its pre-training phase and during the training of the composition map, we use the Sinkhorn divergence with the squared Euclidean as its fitting loss, i.e $\mathcal{S}_{l_2^2} = \Delta(\phi_\omega \sharp \hat{\mu}_n, \hat{\rho}_n)$, with an etropic regularization parameter of $\varepsilon = 0.01$. When training networks $\widetilde{T}_\theta$ and T'_θ, we use the entropic Wasserstein distance (Cuturi, 2013) with the Euclidean distance as the cross-domain cost. For all fitting losses, we scale the cross-domain cost matrix using mean scaling. For all regularizers $\mathcal{GM}^2(\hat{\mu}_n, \phi_\omega \sharp \hat{\mu}_n)$, $\mathcal{GM}^2(\hat{\rho}_n, \widetilde{T}_\theta \sharp \hat{\rho}_n)$, $\mathcal{GM}^2(\hat{\mu}_n, T'_\theta \sharp \hat{\mu}_n)$ we use the quadratic entropic Gromov-Wasserstein distance (Peyré et al., 2016). We use the Euclidean distance for all inter-domain costs $c_\mathcal{X} = c_\mathcal{Y} = c_\mathcal{Z} := \|\cdot\|_2$ and scale them using max scaling. We use a regularization strength of $\lambda_{GM} = 1$ across all losses. For all fitting losses and regularizers, we use an entropic regularization parameter of $\varepsilon = 0.001$.