Linear optimal partial transport embedding

Yikun Bai[1], Ivan Medri[1], Rocio Diaz Martin[2], Rana Muhammad Shahroz Khan[1], and Soheil Kolouri[1]

[1]Department of Computer Science, Vanderbilt University
[2]Department of Mathematics, Vanderbilt University
[1]yikun.bai@vanderbilt.edu
[1]ivan.v.medri@vanderbilt.edu
[2]rocio.p.diaz.martin@vanderbilt.edu
[1]rana.muhammad.shahroz.khan@vanderbilt.edu
[1]soheil.kolouri@vanderbilt.edu

Abstract

Optimal transport (OT) has gained popularity due to its various applications in fields such as machine learning, statistics, and signal processing. However, the balanced mass requirement limits its performance in practical problems. To address these limitations, variants of the OT problem, including unbalanced OT, Optimal partial transport (OPT), and Hellinger Kantorovich (HK), have been proposed. In this paper, we propose the Linear optimal partial transport (LOPT) embedding, which extends the (local) linearization technique on OT and HK to the OPT problem. The proposed embedding allows for faster computation of OPT distance between pairs of positive measures. Besides our theoretical contributions, we demonstrate the LOPT embedding technique in point-cloud interpolation and PCA analysis. Our code is available at `https://github.com/Baio0/LinearOPT`.

1 Introduction

The Optimal Transport (OT) problem has found numerous applications in machine learning (ML), computer vision, and graphics. The probability metrics and dissimilarity measures emerging from the OT theory, e.g., Wasserstein distances and their variations, are used in diverse applications, including training generative models [3, 24, 33], domain adaptation [16, 15], bayesian inference [28], regression [26], clustering [42], learning from graphs [29] and point sets [35, 36], to name a few. These metrics define a powerful geometry for comparing probability measures with numerous desirable properties, for

instance, parameterized geodesics [2], barycenters [18], and a weak Riemannian structure [40].

In large-scale machine learning applications, optimal transport approaches face two main challenges. First, the OT problem is computationally expensive. This has motivated many approximations that lead to significant speedups [17, 13, 39]. Second, while OT is designed for comparing probability measures, many ML problems require comparing non-negative measures with varying total amounts of mass. This has led to the recent advances in unbalanced optimal transport [14, 12, 32] and optimal partial transport [8, 20, 21]. Such unbalanced/partial optimal transport formulations have been recently used to improve minibatch optimal transport [37] and for point-cloud registration [4].

Comparing K (probability) measures requires the pairwise calculation of transport-based distances, which, despite the significant recent computational speed-ups, remains to be relatively expensive. To address this problem, [41] proposed the Linear Optimal Transport (LOT) framework, which linearizes the 2-Wasserstein distance utilizing its weak Riemannian structure. In short, the probability measures are embedded into the tangent space at a fixed reference measure (e.g., the measures' Wasserstein barycenter) through a logarithmic map. The Euclidean distances between the embedded measures then approximate the 2-Wasserstein distance between the probability measures. The LOT framework is computationally attractive as it only requires the computation of one optimal transport problem per input measure, reducing the otherwise quadratic cost to linear. Moreover, the framework provides theoretical guarantees on convexifying certain sets of probability measures [34, 1], which is critical in supervised and unsupervised learning from sets of probability measures. The LOT embedding has recently found diverse applications, from comparing collider events in physics [9] and comparing medical images [5, 30] to permutation invariant pooling for comparing graphs [29] and point sets [35].

Many applications in ML involve comparing non-negative measures (often empirical measures) with varying total amounts of mass, e.g., domain adaptation [19]. Moreover, OT distances (or dissimilarity measures) are often not robust against outliers and noise, resulting in potentially high transportation costs for outliers. Many recent publications have focused on variants of the OT problem that allow for comparing non-negative measures with unequal mass. For instance, the optimal partial transport (OPT) problem [8, 20, 21], Kantorovich–Rubinstein norm [25, 31], and the Hellinger–Kantorovich distance [11, 32]. These methods fall under the broad category of "unbalanced optimal transport" [12, 32]. The existing solvers for "unbalanced optimal transport" problems are generally as expensive or more expensive than the OT solvers. Hence, computation time remains a main bottleneck of these approaches.

To reduce the computational burden for comparing unbalanced measures, [10] proposed a clever extension for the LOT framework to unbalanced nonnegative measures by linearizing the Hellinger-Kantorovich, denoted as Linearized Hellinger-Kantorovich (LHK), distance, with many desirable theoretical properties. However, an unintuitive caveat about HK and LHK formulation is that the geodesic for the transported portion

of the mass does not resemble the OT geodesic. In particular, the transported mass does not maintain a constant mass as it is transported (please see Figure 1). In contrast, OPT behaves exactly like OT for the transported mass with the trade-off of losing the Riemannian structure of HK.

Contributions: In this paper, inspired by OT geodesics, we provide an OPT interpolation technique using its dynamic formulation and explain how to compute it for empirical distributions using barycentric projections. We use this interpolation to embed the space of measures in a euclidean space using optimal partial transport concerning a reference measure. This allows us to extend the LOT framework to LOPT, a linearized version of OPT. Thus, we reduce the computational burden of OPT while maintaining the decoupling properties between noise (created and destroyed mass) and signal (transported mass) of OPT. We propose a LOPT discrepancy measure and a LOPT interpolating curve and contrast them with their OPT counterparts. Finally, we demonstrate applications of the new framework in point cloud interpolation and PCA analysis, showing that the new technique is more robust to noise.

Organization: In section 2, we review Optimal Transport Theory and the Linear Optimal Transport framework to set the basis and intuitions on which we build our new techniques. In Section 3 we review Optimal Partial Transport Theory and present an explicit solution to its Dynamic formulation that we use to introduce the Linear Optimal Partial Transport framework (LOPT). We define LOPT Embedding, LOPT Discrepancy, LOPT interpolation and give explicit ways to work with empirical data. In Section 4 we show applications of the LOPT framework to approximate OPT distances, to interpolate between point cloud datasets, and to preprocess data for PCA analysis. In the appendix, we provide proofs for all the results, new or old, for which we could not find a proof in the literature.

2 Background: OT and LOT

2.1 Static Formulation of Optimal Transport

Let $\mathcal{P}(\Omega)$ be the set of Borel probability measures defined in a convex compact subset Ω of $\mathbb{R}^d$, and consider $\mu^0, \mu^j \in \mathcal{P}(\Omega)$. The Optimal Transport (OT) problem between μ^0 and μ^j is that of finding the cheapest way to transport *all* the mass distributed according to the *reference* measure μ^0 onto a new distribution of mass determined by the *target* measure μ^j. Mathematically, it was stated by Kantorovich as the minimization problem

$$OT(\mu^0, \mu^j) := \inf_{\gamma \in \Gamma(\mu^0, \mu^j)} C(\gamma; \mu^0, \mu^j) \tag{1}$$

$$\text{for} \quad C(\gamma; \mu^0, \mu^j) := \int_{\Omega^2} \|x^0 - x^j\|^2 d\gamma(x^0, x^j), \tag{2}$$

where $\Gamma(\mu^0, \mu^j)$ is the set of all joint probability measures in Ω^2 with marginals μ^0 and μ^j. A measure $\gamma \in \Gamma(\mu^0, \mu^j)$ is called a *transportation plan*, and given measurable

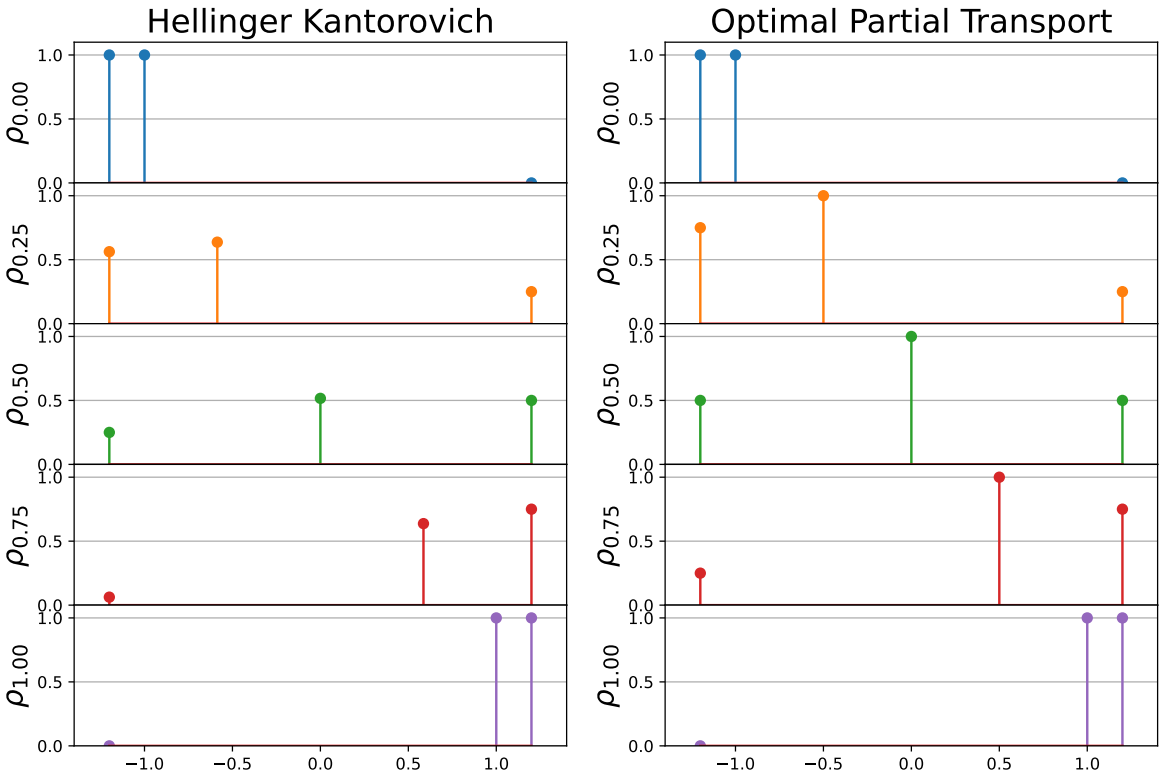


Figure 1: The depiction of the HK and OPT geodesics between two measures, at times $t \in \{0, 0.25, 0.5, 0.75, 1\}$. The top row (Blue) represents two initial deltas of mass one located at positions -1.2 and -1. The bottom row (Purple) shows two final deltas of mass one located at 1 and 1.2. At intermediate time steps $t = 0.25, 0.5, 0.75$, the transported part (middle delta moving from -1 to 1) changes mass for HK while its mass remains constant for OPT. Outer masses (located at -1.2 for initial time $t = 0$, and at 1.2 for final time $t = 1$) are being destroyed and created, so mass changes are expected. Notably, mass is created/destroyed with a linear rate for OPT and a nonlinear rate for HK. See Appendix H.4 for further analysis.

sets $A, B \in \Omega$, the coupling $\gamma(A \times B)$ describes how much mass originally in the set A is transported into the set B. The squared of the Euclidean distance[1] $\|x^0 - x^j\|^2$ is interpreted as the cost of transporting a unit mass located at x^0 to x^j. Therefore, $C(\gamma; \mu^0, \mu^j)$ represents the total cost of moving μ^0 to μ^j according to γ. Finally, we will denote the set of all plans that achieve the infimum in (1), which is non-empty [40], as $\Gamma^*(\mu^0, \mu^j)$.

Under certain conditions (e.g. when μ^0 has continuous density), an optimal plan γ can be induced by a rule/map T that takes all the mass at each position x to a unique point $T(x)$. If that is the case, we say that γ does not split mass and that it is **induced by a map T**. In fact, it is concentrated on the graph of T in the sense that for all measurable sets $A, B \subset \Omega$, $\gamma(A \times B) = \mu^0(\{x \in A : T(x) \in B\})$, and we will write it as the *pushforward* $\gamma = (\mathrm{id} \times T)_{\#}\mu^0$. Hence, (1) reads as

$$OT(\mu^0, \mu^j) = \int_\Omega \|x - T(x)\|^2 d\mu^0(x) \tag{3}$$

The function $T : \Omega \to \Omega$ is called a *Monge map*, and when μ^0 is absolutely continuous it is unique [7].

Finally, the square root of the optimal value $OT(\cdot, \cdot)$ is exactly the so-called **Wasserstein distance**, W_2, in $\mathcal{P}(\Omega)$ [40, Th.7.3], and we will call it also as **OT squared distance**. In addition, with this distance, $\mathcal{P}(\Omega)$ is not only a metric space but also a Riemannian manifold [40]. In particular, the tangent space of any $\mu \in \mathcal{P}(\Omega)$ is $\mathcal{T}_\mu = L^2(\Omega; \mathbb{R}^d, \mu) = \{u : \Omega \to \mathbb{R}^d : \|u\|_\mu^2 < \infty\}$, where

$$\|u\|_\mu^2 := \int_\Omega \|u(x)\|^2 d\mu(x). \tag{4}$$

2.2 Dynamic Formulation of Optimal Transport

To understand the framework of Linear Optimal Transport (LOT) we will use the dynamic formulation of the OT problem. Optimal plans and maps can be viewed as a static way of matching two distributions. They tell us where each mass in the initial distribution should end, but they do not tell the full story of how the system evolves from initial to final configurations.

In the dynamic formulation, we consider $\rho \in \mathcal{P}([0, 1] \times \Omega)$ a curve of measures parametrized in time that describes the distribution of mass $\rho_t := \rho(t, \cdot) \in \mathcal{P}(\Omega)$ at each instant $0 \leq t \leq 1$. We will require the curve to be sufficiently smooth, to have boundary conditions $\rho_0 = \mu^0$, $\rho_1 = \mu^j$, and to satisfy the conservation of mass law. Then, it is well known that there exists a velocity vector field $v_t := v(t, \cdot)$ such that ρ_t satisfies the continuity equation[2] with boundary conditions

$$\partial_t \rho + \nabla \cdot \rho v = 0, \qquad \rho_0 = \mu^0, \quad \rho_1 = \mu^j. \tag{5}$$

[1] More general cost functions might be used, but they are beyond the scope of this article.

[2] The continuity equation is satisfied weakly or in the sense of distributions. See [40, 38].

The length[3] of the curve can be stated as $\int_{[0,1]\times\Omega} \|v\|^2 d\rho := \int_0^1 \|v_t\|^2_{\rho_t} dt$, for $\|\cdot\|_{\rho_t}$ as in (4), and $OT(\mu^0, \mu^j)$ coincides with the length of the shortest curve between μ^0 and μ^j [6]. Hence, the dynamical formulation of the OT problem (1) reads as

$$OT(\mu^0, \mu^j) = \inf_{(\rho,v)\in\mathcal{CE}(\mu^0,\mu^j)} \int_{[0,1]\times\Omega} \|v\|^2 d\rho, \tag{6}$$

where $\mathcal{CE}(\mu^0, \mu^j)$ is the set of pairs (ρ, v), where $\rho \in \mathcal{P}([0,1]\times\Omega)$, and $v : [0,1]\times\Omega \to \mathbb{R}^d$, satisfying (5).

Under the assumption of existence of an optimal Monge map T, an optimal solution for (6) can be given explicitly and is pretty intuitive. If a particle starts at position x and finishes at position $T(x)$, then for $0 < t < 1$ it will be at the point

$$T_t(x) := (1-t)x + tT(x). \tag{7}$$

Then, varying both the time t and $x \in \Omega$, the mapping (7) can be interpreted as a flow whose time velocity[4] is

$$v_t(x) = T(x_0) - x_0, \qquad \text{for } x = T_t(x_0). \tag{8}$$

To obtain the curve of probability measures ρ_t, one can evolve μ^0 through the flow T_t using the formula $\rho_t(A) = \mu_0(T_t^{-1}(A))$ for any measurable set A. That is, ρ_t is the *push-forward* of μ^0 by T_t

$$\rho_t = (T_t)_{\#}\mu^0, \qquad 0 \le t \le 1. \tag{9}$$

The pair (ρ, v) defined by (9) and (8) satisfies the continuity equation (5) and solves (6). Moreover, the curve ρ_t is a *constant speed geodesic* in $\mathcal{P}(\Omega)$ between μ^0 and μ^j [22], i.e., it satisfies that for all $0 \le s \le t \le 1$

$$\sqrt{OT(\rho_s, \rho_t)} = (t-s)\sqrt{OT(\rho_0, \rho_1)}. \tag{10}$$

A confirmation of this comes from comparing the OT cost (3) with (8) obtaining

$$OT(\mu^0, \mu^j) = \int_\Omega \|v_0(x)\|^2 d\mu^0(x) \tag{11}$$

which tells us that we only need the speed at the initial time to compute the total length of the curve. Moreover, $OT(\mu^0, \mu^j)$ coincides with the squared norm of the tangent vector v_0 in the tangent space $\mathcal{T}_{\mu^0}$ of $\mathcal{P}(\Omega)$ at μ^0.

[3]Precisely, the length of the curve ρ, with respect to the Wasserstein distance, should be $\int_0^1 \|v_t\|_{\rho_t} dt$, but this will make no difference in the solutions of (6) since they are constant speed geodesics.

[4]For each $(t, x) \in (0,1)\times\Omega$, the vector $v_t(x)$ is well defined as T_t is invertible. See [38, Lemma 5.29].

2.3 Linear Optimal Transport Embedding

Inspired by the induced Riemannian geometry of the OT squared distance, [41] proposed the so-called **Linear Optimal Transportation** (LOT) framework. Given two target measures μ^i, μ^j, the main idea relies on considering a reference measure μ^0 and embed these target measures into the tangent space $\mathcal{T}_{\mu^0}$. This is done by identifying each measure μ^j with the curve (9) minimizing $OT(\mu^0, \mu^j)$ and computing its velocity (tangent vector) at $t = 0$ using (8).

Formally, let us fix a continuous probability reference measure μ^0. Then, the **LOT embedding** [34] is defined as

$$\mu^j \mapsto u^j := T^j - \mathrm{id} \qquad \forall\ \mu^j \in \mathcal{P}(\Omega) \tag{12}$$

where T^j is the optimal Monge map between μ^0 and μ^j. Notice that by (3), (4) and (11) we have

$$\|u^j\|^2_{\mu^0} = OT(\mu^0, \mu^j). \tag{13}$$

After this embedding, one can use the distance in $\mathcal{T}_{\mu^0}$ between the projected measures to define a new distance in $\mathcal{P}(\Omega)$ that can be used to approximate $OT(\mu^i, \mu^j)$. The **LOT squared distance** is defined as

$$LOT_{\mu^0}(\mu^i, \mu^j) := \|u^i - u^j\|^2_{\mu^0}. \tag{14}$$

2.4 LOT in the Discrete Setting

For discrete probability measures μ^0, μ^j of the form

$$\mu^0 = \sum_{n=1}^{N_0} p_n^0 \delta_{x_n^0}, \qquad \mu^j = \sum_{n=1}^{N_j} p_n^j \delta_{x_n^j}, \tag{15}$$

a Monge map T^j for $OT(\mu^0, \mu^j)$ may not exist. Following [41], in this setting, the target measure μ^j can be replaced by a new measure $\hat{\mu}^j$ for which an optimal transport Monge map exists. For that, given an optimal plan $\gamma^j \in \Gamma^*(\mu^0, \mu^j)$, it can be viewed as a $N_0 \times N_i$ matrix whose value at position (n, m) represents how much mass from x_n^0 should be taken to x_m^j. Then, we define the **OT barycentric projection**[5] of μ^j **with respect to** μ^0 as

$$\hat{\mu}^j := \sum_{n=1}^{N_0} p_n^0 \delta_{\hat{x}_n^j}, \text{ where } \quad \hat{x}_n^j := \frac{1}{p_n^0} \sum_{m=1}^{N_j} \gamma_{n,m}^j x_m^j. \tag{16}$$

The new measure $\hat{\mu}^j$ is regarded as an N_0-point representation of the target measure μ^j. The following lemma guarantees the existence of a Monge map between μ^0 and $\hat{\mu}^j$.

Lemma 2.1. *Let μ^0 and μ^j be two discrete probability measures as in* (15)*, and consider an OT barycentric projection $\hat{\mu}^j$ of μ^j with respect to μ^0 as in* (16)*. Then, the map $x_n^0 \mapsto \hat{x}_n^j$ given by* (16) *solves the OT problem $OT(\mu^0, \hat{\mu}^j)$.*

[5] We refer to [2] for the rigorous definition.

It is easy to show that if the optimal transport plan γ^j is induced by a Monge map, then $\hat{\mu}^j = \mu^j$. As a consequence, the OT barycentric projection is an actual projection in the sense that it is idempotent.

Similar to the continuous case (12), given a discrete reference measure μ^0, we can define the **LOT embedding** for a discrete measure μ^j as the rule

$$\mu^j \mapsto u^j := [(\hat{x}_1^j - x_1^0), \dots, (\hat{x}_{N_0}^j - x_{N_0}^0)].^6 \tag{17}$$

The range $\mathcal{T}_{\mu_0}$ of this application is identified with $\mathbb{R}^{d\times N_0}$ with the norm $\|u\|_{\mu^0} := \sum_{n=1}^{N_0} \|u(n)\|^2 p_n^0$, where $u(n) \in \mathbb{R}^d$ denotes the nth entry of u. We call $(\mathbb{R}^{d\times N_0}, \|\cdot\|_{\mu^0})$ the embedding space.

By the discussion above, if the optimal plan γ^j for problem $\mathrm{OT}(\mu^0, \mu^j)$ is induced by a Monge map, then the discrete embedding is consistent with (13) in the sense that

$$\|u^j\|_{\mu^0}^2 = OT(\mu^0, \hat{\mu}^j) = OT(\mu^0, \mu^j). \tag{18}$$

Hence, as in section 2.3, we can use the distance between embedded measures in $(\mathbb{R}^{d\times N_0}, \|\cdot\|_{\mu_0})$ to define a *discrepancy* in the space of discrete probabilities that can be used to approximate $OT(\mu^i, \mu^j)$. The **LOT discrepancy**[7] is defined as

$$LOT_{\mu^0}(\mu^i, \mu^j) := \|u^i - u^j\|_{\mu^0}^2. \tag{19}$$

We call it a *discrepancy* because it is not a squared metric between discrete measures. It does not necessarily satisfy that $LOT(\mu^i, \mu^j) \neq 0$ for every distinct μ^i, μ^j. Nevertheless, $\|u^i - u^j\|_{\mu^0}$ is a metric in the embedding space.

2.5 OT and LOT Geodesics in Discrete Settings

Let μ^i, μ^j be discrete probability measures as in (15) (with 'i' in place of 0). If an optimal Monge map T for $OT(\mu^i, \mu^j)$ exists, a constant speed geodesic ρ_t between μ^i and μ^j, for the OT squared distance, can be found by mimicking (9). Explicitly, with T_t as in (7),

$$\rho_t = (T_t)_{\#}\mu^i = \sum_{n=1}^{N_i} p_n^i \delta_{(1-t)x_n^i + tT(x_n^i)}. \tag{20}$$

In practice, one replaces μ^j by its OT barycentric projection with respect to μ^i (and so, the existence of an optimal Monge map is guaranteed by Lemma 2.1).

Now, given a discrete reference μ^0, the LOT discrepancy provides a new structure to the space of discrete probability densities. Therefore, we can provide a substitute

[6] In [41], the embedded vector is defined as the the element-wise multiplication $u^j \bigodot \sqrt{p^0}$, In addition, $\hat{\mu}^j, u^j$ are determined by γ^j, but we ignore the subscript 'γ^j' for convenience.

[7] In [41], LOT is defined by the infimum over all possible optimal pairs (γ^i, γ^j). We do not distinguish these two formulations for convenience in this paper. Additionally, (19) is determined by the choice of (γ^i, γ^j).

for the OT geodesic (20) between μ^i and μ^j. Assume we have the embeddings $\mu^i \mapsto u^i$, $\mu^j \mapsto u^j$ as in (17). The geodesic between u^i and u^j in the LOT embedding space $\mathbb{R}^{d\times N_0}$ has the simple form $u_t = (1-t)u^i + tu^j$. This correlates with the curve $\hat{\rho}_t$ in $\mathcal{P}(\Omega)$ induced by the map $\hat{T} : \hat{x}_n^i \mapsto \hat{x}_n^j$ [8] as

$$\hat{\rho}_t := (\hat{T}_t)_{\#}\hat{\mu}^i = \sum_{n=1}^{N_0} p_n^0 \delta_{x_n^0 + u_t(n)}. \tag{21}$$

By abuse of notation, we call this curve the **LOT geodesic** between μ^i and μ^j. Nevertheless, it is a *geodesic between their barycentric projections* since it satisfies the following.

Proposition 2.2. *Let $\hat{\rho}_t$ be defined as* (21)*, and $\hat{\rho}_0 = \hat{\mu}^i, \hat{\rho}_1 = \hat{\mu}^j$, then for all $0 \le s \le t \le 1$*

$$\sqrt{LOT_{\mu^0}(\hat{\rho}_s, \hat{\rho}_t)} = (t-s)\sqrt{LOT_{\mu^0}(\hat{\rho}_0, \hat{\rho}_1)}. \tag{22}$$

3 Linear Optimal Partial Transport Embedding

3.1 Static Formulation of Optimal Partial Transport

In addition to mass transportation, the OPT problem allows mass destruction at the source and mass creation at the target. Let $\mathcal{M}_+(\Omega)$ denote the set of all positive finite Borel measures defined on Ω. For $\lambda \ge 0$ the OPT problem between $\mu^0, \mu^j \in \mathcal{M}_+(\Omega)$ can be formulated as

$$OPT_\lambda(\mu^0, \mu^j) := \inf_{\gamma \in \Gamma_{\le}(\mu^0, \mu^j)} C(\gamma; \mu^0, \mu^j, \lambda) \tag{23}$$

$$\text{for} \quad C(\gamma; \mu^0, \mu^j, \lambda) := \int_{\Omega^2} \|x^0 - x^j\|^2 d\gamma(x^0, x^j) + \lambda(|\mu^0 - \gamma_0| + |\mu^j - \gamma_1|) \tag{24}$$

where $|\mu^0 - \gamma_0|$ is the total mass of $\mu^0 - \gamma_0$ (resp. $|\mu^j - \gamma_1|$), and $\Gamma_{\le}(\mu^0, \mu^j)$ denotes the set of all measures in Ω^2 with marginals γ_0 and γ_1 satisfying $\gamma_0 \le \mu^0$ (i.e., $\gamma_0(E) \le \mu^0(E)$ for all measurable set E), and $\gamma_1 \le \mu^j$. Here, the mass destruction and creation penalty is linear, parametrized by λ. The set of minimizers $\Gamma^*_{\le}(\mu^0, \mu^j)$ of (23) is non-empty [20]. One can further restrict $\Gamma_{\le}(\mu^0, \mu^j)$ to the set of partial transport plans γ such that $\|x^0 - x^j\|^2 < 2\lambda$ for all $(x^0, x^j) \in \text{supp}(\gamma)$ [4, Lemma 3.2]. This means that if the usual transportation cost is greater than 2λ, it is better to create/destroy mass.

3.2 Dynamic Formulation of Optimal Partial Transport

Adding a forcing term ζ to the continuity equation (6), one can take into account curves that allow creation and destruction of mass. That is, those who break the conservation

[8] This map can be understood as the one that transports $\hat{\mu}^i$ onto $\hat{\mu}^j$ pivoting on the reference: $\hat{\mu}^i \mapsto \mu^0 \mapsto \hat{\mu}^j$.

of mass law. Thus, it is natural that the minimization problem (23) can be rewritten [12, Th. 5.2] into a dynamic formulation as

$$OPT_\lambda(\mu^0, \mu^j) = \inf_{(\rho,v,\zeta)\in\mathcal{FCE}(\mu^0,\mu^j)} \int_{[0,1]\times\Omega} \|v\|^2 d\rho + \lambda|\zeta| \tag{25}$$

where $\mathcal{FCE}(\mu^0, \mu^j)$ is the set of tuples (ρ, v, ζ) such that $\rho \in \mathcal{M}_+([0,1]\times\Omega)$, $\zeta \in \mathcal{M}([0,1]\times\Omega)$ (where $\mathcal{M}$ stands for signed measures) and $v : [0,1]\times\Omega \to \mathbb{R}^d$, satisfying

$$\partial_t\rho + \nabla\cdot\rho v = \zeta, \qquad \rho_0 = \mu^0, \quad \rho_1 = \mu^j. \tag{26}$$

As in the case of OT, under certain conditions on the minimizers γ of (23), one curve ρ_t that minimizes the dynamic formulation (25) is quite intuitive. We show in the next proposition that it consists of three parts γ_t, $(1-t)\nu_0$ and $t\nu^j$ (see (27), (28), and (29) below). The first is a curve that only transports mass, and the second and third destroy and create mass at constant rates $|\nu_0|$, $|\nu^j|$, respectively.

Proposition 3.1. *Let $\gamma^* \in \Gamma^*_\leq(\mu^0, \mu^j)$ be of the form $\gamma^* = (\mathrm{id}\times T)_\#\gamma^*_0$ for $T : \Omega \to \Omega$ a (measurable) map. Let*

$$\nu_0 := \mu^0 - \gamma^*_0, \quad \nu^j := \mu^j - \gamma^*_1, \tag{27}$$

$$T_t(x) := (1-t)x + tT(x), \quad \gamma_t := (T_t)_\#\gamma^*_0. \tag{28}$$

Then, an optimal solution (ρ, v, ζ) for (25) is given by

$$\rho_t := \gamma_t + (1-t)\nu_0 + t\nu^j, \tag{29}$$

$$v_t(x) := T(x_0) - x_0, \qquad \text{if } x = T_t(x_0). \tag{30}$$

$$\zeta_t := \nu^j - \nu_0 \tag{31}$$

Moreover, plugging in (ρ, v, ζ) into (25), it holds that

$$OPT_\lambda(\mu^0, \mu^j) = \|v_0\|^2_{\gamma^*_0, 2\lambda} + \lambda(|\nu_0| + |\nu^j|), \tag{32}$$

where $v_0(x) = T(x) - x$ (i.e., v_t at time $t = 0$), and

$$\|v\|^2_{\mu,2\lambda} := \int_\Omega \min(\|v\|^2, 2\lambda) d\mu, \qquad \text{for } v : \Omega \to \mathbb{R}^d.$$

In analogy to the OT squared distance, we also call the optimal partial cost (32) as the **OPT squared distance**.

3.3 Linear Optimal Partial Transport Embedding

Definition 3.2. Let μ^0, $\mu^j \in \mathcal{M}_+(\Omega)$ such that $OPT_\lambda(\mu^0, \mu^j)$ is solved by a plan induced by a map. The **LOPT embedding** of μ^j with respect to μ^0 is defined as

$$\mu^j \mapsto (u^j, \bar{\mu}^j, \nu^j) := (v_0, \gamma_0, \nu^j) \tag{33}$$

where v_0, γ_0, ν^j are defined as in Proposition 3.1.

Let us compare the LOPT (33) and LOT (12) embeddings. The first component v_0 represents the tangent of the curve that transports mass from the reference to the target. This is exactly the same as the LOT embedding. In contrast to LOT, the second component γ_0 is necessary since we need to specify what part of the reference is being transported. The third component ν^j can be thought of as the tangent vector of the part that creates mass. There is no need to save the destroyed mass because it can be inferred from the other quantities.

Now, let $\mu^0 \wedge \mu^j$ be the *minimum measure*[9] between μ^0 and μ^j. By the above definition, $\mu^0 \mapsto (u^0, \bar{\mu}^0, \nu^0) = (0, \mu^0, 0)$. Therefore, (32) can be rewritten

$$OPT_\lambda(\mu^0, \mu^j) = \|u^0 - u^j\|^2_{\bar{\mu}^0 \wedge \bar{\mu}^j, 2\lambda} + \lambda(|\bar{\mu}^0 - \bar{\mu}^j| + |\nu_0| + |\nu^j|) \tag{34}$$

This motivates the definition of the **LOPT discrepancy**.[10]

Definition 3.3. Consider a reference $\mu^0 \in \mathcal{M}_+(\Omega)$ and target measures $\mu^i, \mu^j \in \mathcal{M}_+(\Omega)$ such that $OPT_\lambda(\mu^0, \mu^i)$ and $OPT_\lambda(\mu^0, \mu^j)$ can be solved by plans induced by mappings as in the hypothesis of Proposition 3.1. Let $(u^i, \bar{\mu}^i, \nu^i)$ and $(u^j, \bar{\mu}^j, \nu^j)$ be the LOPT embeddings of μ^i and μ^j with respect to μ^0. The **LOPT discrepancy** between μ^i and μ^j with respect to μ^0 is defined as

$$LOPT_{\mu^0,\lambda}(\mu^i, \mu^j) := \|u^i - u^j\|^2_{\bar{\mu}^i \wedge \bar{\mu}^j, 2\lambda} + \lambda(|\bar{\mu}^i - \bar{\mu}^j| + |\nu^i| + |\nu^j|) \tag{35}$$

Similar to the LOT framework, by equation (34), LOPT can recover OPT when $\mu^i = \mu^0$. That is,

$$LOPT_{\mu^0,\lambda}(\mu^0, \mu^j) = OPT_\lambda(\mu^0, \mu^j).$$

3.4 LOPT in the Discrete Setting

If μ^0, μ^j are N_0, N_j-size discrete non-negative measures as in (15) (but not necessarily with total mass 1), the OPT problem (23) can be written as

$$\min_{\gamma \in \Gamma_\leq(\mu^0, \mu^j)} \sum_{n,m} \|x^0_n - x^j_m\|^2 \gamma_{n,m} + \lambda(|p^0| + |p^j| - 2|\gamma|)$$

where the set $\Gamma_\leq(\mu^0, \mu^j)$ can be viewed as the subset of $N_0 \times N_j$ matrices with non-negative entries

$$\Gamma_\leq(\mu^0, \mu^j) := \{\gamma \in \mathbb{R}_+^{N_0 \times N_j} : \gamma 1_{N_j} \leq p^0, \gamma^T 1_{N_0} \leq p^j\},$$

where 1_{N_0} denotes the $N_0 \times 1$ vector whose entries are 1 (resp. 1_{N_j}), $p^0 = [p^0_1, \ldots, p^0_{N_0}]$ is the vector of weights of μ^0 (resp. p^j), $\gamma 1_{N_j} \leq p^0$ means that component-wise holds

[9] Formally, $\mu^0 \wedge \mu^j(B) := \inf\{\mu^0(B_1) + \mu^j(B_2)\}$ for every Borel set B, where the infimum is taken over all partitions of B., i.e. $B = B_1 \cup B_2$, $B_1 \cap B_2 = \emptyset$, given by Borel sets B_1, B_2.

[10] $LOPT_\lambda$ is not a rigorous metric.

the '$\leq$' (resp. $\gamma^T 1_{N_0} \leq p^j$, where γ^T is the transpose of γ), and $|p^0| = \sum_{n=1}^{N_0} |p_n^0|$ is the total mass of μ^0 (resp. $|p^j|, |\gamma|$). The marginals are $\gamma_0 := \gamma 1_{N_j}$, and $\gamma_1 := \gamma^T 1_{N_0}$.

Similar to OT, when an optimal plan γ^j for $OPT_\lambda(\mu^0, \hat{\mu}^j)$ is not induced by a map, we can replace the target measure μ^j by an **OPT barycentric projection** $\hat{\mu}^j$ for which a map exists. Therefore, allowing us to apply the LOPT embedding (see (33) and (40) below).

Definition 3.4. Let μ^0 and μ^j be positive discrete measures, and $\gamma^j \in \Gamma^*_{\leq}(\mu^0, \mu^j)$. The **OPT barycentric projection**[11] of μ^j **with respect to** μ^0 is defined as

$$\hat{\mu}^j := \sum_{n=1}^{N_0} \hat{p}_n^j \delta_{\hat{x}_n^j}, \qquad \text{where} \tag{36}$$

$$\hat{p}_n^j := \sum_{m=1}^{N_j} \gamma_{n,m}^j, \qquad 1 \leq n \leq N_0, \tag{37}$$

$$\hat{x}_n^j := \begin{cases} \frac{1}{\hat{p}_n^j} \sum_{m=1}^{N_j} \gamma_{n,m}^j x_m^j & \text{if } \hat{p}_n^j > 0 \\ x_n^0 & \text{if } \hat{p}_n^j = 0. \end{cases} \tag{38}$$

Theorem 3.5. *In the same setting of Definition 3.4, the map $x_n^0 \mapsto \hat{x}_n^j$ given by* (38) *solves the problem $OPT_\lambda(\mu^0, \hat{\mu}^j)$, in the sense that induces the partial optimal plan $\hat{\gamma}^j = \operatorname{diag}(\hat{p}_1^j, \ldots, \hat{p}_{N_0}^j)$.*

It is worth noting that when we take a barycentric projection of a measure, some information is lost. Specifically, the information about the part of μ^j that is not transported from the reference μ^0. This has some minor consequences.

First, unlike (18), the optimal partial transport cost $OPT_\lambda(\mu^0, \mu^j)$ changes when we replace μ^j by $\hat{\mu}^j$. Nevertheless, the following relation holds.

Theorem 3.6. *In the same setting of Definition 3.4, if γ^j is induced by a map, then*

$$OPT_\lambda(\mu^0, \mu^j) = OPT_\lambda(\mu^0, \hat{\mu}^j) + \lambda(|\mu^j| - |\hat{\mu}^j|) \tag{39}$$

The second consequence[12] is that the LOPT embedding of $\hat{\mu}^j$ will always have a null third component. That is,

$$\hat{\mu}^j \mapsto ([\hat{x}_1^j - x_1^0, \ldots, \hat{x}_{N_0}^j - x_{N_0}^0], \sum_{n=1}^{N_0} \hat{p}_n^j \delta_{x_n^0}, 0). \tag{40}$$

Therefore, we represent this embedding as $\hat{\mu}^j \mapsto (u^j, \hat{p}^j)$, for $u^j = [\hat{x}_1^j - x_1^0, \ldots, \hat{x}_{N_0}^j - x_{N_0}^0]$ and $\hat{p}^j = [\hat{p}_1^j, \ldots, \hat{p}_{N_0}^j]$. The last consequence is given in the next result.

[11] Notice that in (16) we had $p_n^0 = \sum_{m=1}^{N_j} \gamma_{n,m}^j$. This leads to introducing $\hat{p}_n^j$ as in (37). That is, $\hat{p}_n^j$ plays the role of p_n^0 in the OPT framework. However, here $\hat{p}_n^j$ depends on γ^j (on its first marginal γ_0^j) and not only on μ^0, and so we add a superscript 'j'.

[12] This is indeed an advantage since it allows the range of the embedding to always have the same dimension $N_0 \times (d+1)$.

Proposition 3.7. *If μ^0, μ^i, μ^j are discrete and satisfy the conditions of Definition 3.3, then*

$$LOPT_{\mu^0,\lambda}(\mu^i, \mu^j) = LOPT_{\mu^0,\lambda}(\hat{\mu}^i, \hat{\mu}^j) + \lambda C_{i,j} \tag{41}$$

where $C_{i,j} = |\mu^i| - |\hat{\mu}^i| + |\mu^j| - |\hat{\mu}^j|$.

As a byproduct we can define the **LOPT discrepancy** for **any** pair of discrete measures μ^i, μ^j as the right-hand side of (41). In practice, unless to approximate $OPT_\lambda(\mu^i, \mu^j)$, we set $C_{i,j} = 0$ in (41). That is,

$$LOPT_{\mu^0,\lambda}(\mu^i, \mu^j) := LOPT_{\mu^0,\lambda}(\hat{\mu}^i, \hat{\mu}^j). \tag{42}$$

3.5 OPT and LOPT Interpolation

Inspired by OT and LOT geodesics as defined in section 2.5, but lacking the Riemannian structure provided by the OT squared norm, we propose an OPT interpolation curve and its LOPT approximation.

For the OPT interpolation between two measures μ^i, μ^j for which exists $\gamma \in \Gamma^*_{\leq}(\mu^i, \mu^j)$ of the form $\gamma = (\mathrm{id} \times T)_\# \gamma_0$, a natural candidate is the solution ρ_t of the dynamic formulation of $OPT_\lambda(\mu^i, \mu^j)$. The exact expression is given by Proposition 3.1. When working with general discrete measures μ^i, μ^j (as in (15), with 'i' in place of 0) such γ is not guaranteed to exist. Then, we replace the latter with its OPT barycentric projection with respect to μ^i. And by Theorem 3.5 the map $T : x^i_n \mapsto \hat{x}^j_n$ solves $OPT_\lambda(\mu^i, \hat{\mu}^j)$ and the **OPT interpolating curve** is[13]

$$t \mapsto \sum_{n=1}^{N_i} \hat{p}^j_n \delta_{(1-t)x^i_n + tT(x^i_n)} + (1-t) \sum_{n=1}^{N_i} (p^i_n - \hat{p}^j_n) \delta_{x^i_n}.$$

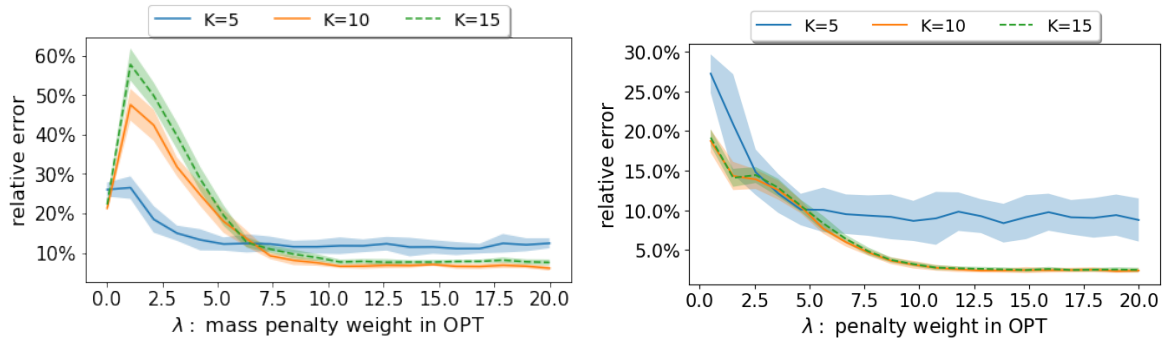
When working with a multitude of measures, it is convenient to consider a reference μ^0 and embed the measures in $\mathbb{R}^{(d+1)\times N_0}$ using LOPT. Hence, doing computations in a simpler space. Below we provide the LOPT interpolation.

Definition 3.8. Given discrete measures μ^0, μ^i, μ^j, with μ^0 as the reference, let $(u^i, \hat{p}^i), (u^j, \hat{p}^i)$ be the LOPT embeddings of μ^i, μ^j. Let $\hat{p}^{ij} := \hat{p}^i \wedge \hat{p}^j$, and $u_t := (1-t)u^i + tu^j$. We define the **LOPT interpolating curve** between μ^i and μ^j by

$$t \mapsto \sum_{k \in D_T} \hat{p}^{ij}_k \delta_{x^0_k + u_t(k)} + (1-t) \sum_{k \in D_D} (\hat{p}^i_k - \hat{p}^{ij}_k) \delta_{x^0_k + u^i_k} + t \sum_{k \in D_C} (\hat{p}^j_k - \hat{p}^{ij}_k) \delta_{x^0_k + u^j_k}$$

where $D_T = \{k : \hat{p}^{ij}_k > 0\}$, $D_D = \{k : \hat{p}^i_k > \hat{p}^{ij}_k)\}$ and $D_C = \{k : \hat{p}^{ij}_k < \hat{p}^j_k)\}$ are respectively the sets where we transport, destroy and create mass.

[13] $\hat{p}^j_n$ are the coefficients of $\hat{\mu}^j$ with respect to μ^i analogous to (36).



(a) Mean of the error

(b) Medean of the error

Figure 2: Graphs of the mean and median relative errors between OPT_λ and $LOPT_{\lambda,\mu_0}$ as a function of the parameter λ.

4 Applications

Approximation of OPT Distance: Similar to LOT [41], and Linear Hellinger Kantorovich (LHK) [10], we test the approximation performance of OPT using LOPT. Given K empirical measures $\{\mu^i\}_{i=1}^K$, for each pair (μ^i, μ^j), we compute $OPT_\lambda(\mu^i, \mu^j)$ and $LOPT_{\mu^0,\lambda}(\mu^i, \mu^j)$ and the mean or median of all pairs (μ^i, μ^j) of relative error defined as

$$\frac{|OPT_\lambda(\mu^i, \mu^j) - LOPT_{\mu^0,\lambda}(\mu^i, \mu^j)|}{OPT_\lambda(\mu^i, \mu^j)}.$$

Similar to LOT and LHK, the choice of μ^0 is critical for the accurate approximation of OPT. If μ^0 is far away from $\{\mu^i\}_{i=1}^K$, the linearization is a poor approximation because the mass in μ^i and μ^0 would only be destroyed or created. In practice, one candidate for μ^0 is the barycenter of the set of measures $\{\mu^i\}$. The OPT can be converted into OT problem [8], and one can use OT barycenter [18] to find μ^0.

For our experiments, we created K point sets of size $N = 500$ for K different Gaussian distributions in $\mathbb{R}^2$. In particular, $\mu^i \sim \mathcal{N}(m^i, I)$, where m^i is randomly selected such that $\|m^i\| = \sqrt{3}$ for $i = 1, ..., K$. For the reference, we picked an N point representation of $\mu^0 \sim \mathcal{N}(\overline{m}, I)$ with $\overline{m} = \sum m^i/K$. We repeated each experiment 10 times. To exhibit the effect of the parameter λ in the approximation, the relative errors are shown in Figure 2. For the histogram of the relative errors for each value of λ and each number of measures K, we refer to Figure 6 in the Appendix H. For large λ, most mass is transported and $OT(\mu^i, \mu^j) \approx OPT_\lambda(\mu^i, \mu^j)$, the performance of LOPT is close to that of LOT, and the relative error is small.

In Figure 3 we report wall clock times of OPT vs LOPT for $\lambda = 5$. We use linear programming [27] to solve each OPT problem with a cost of $\mathcal{O}(N^3\log(N))$ each. Thus, computing the OPT distance pair-wisely for $\{\mu^i\}_{i=1}^K$ requires $\mathcal{O}(K^2N^3\log(N))$. In contrast, to compute $LOPT$, we only need to solve K optimal partial transport problems for the embeddings (see (33) or (40)). Computing LOPT discrepancies after

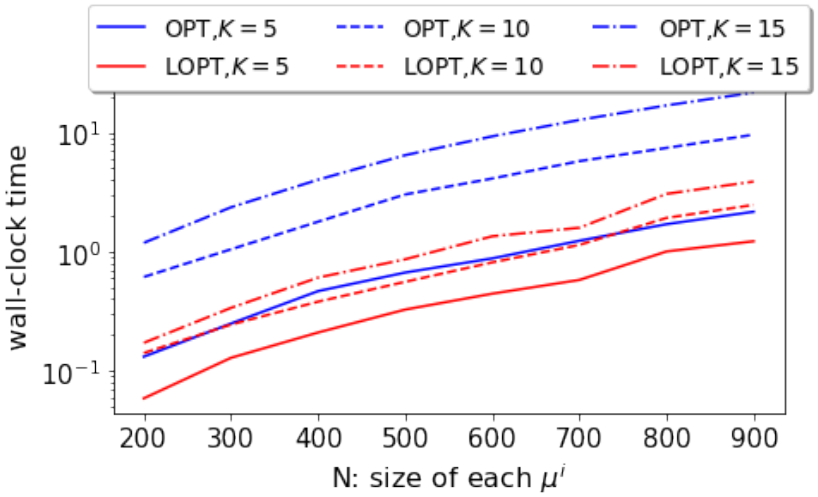


Figure 3: Wall-clock time between OPT and LOPT. The LP solver in PythonOT [23] is applied to each individual OPT problem, with $100N$ maximum number of iterations.

the embeddings is linear. Thus, the total computational cost is $\mathcal{O}(KN^3\log(N) + K^2N)$. The experiment was conducted on a Linux computer with AMD EPYC 7702P CPU with 64 cores and 256GB DDR4 RAM.

Point Cloud Interpolation: We test OT geodesic, LOT geodesic, OPT interpolation, and LOPT interpolation on the point cloud MNIST dataset. We compute different transport curves between point sets of the digits 0 and 9. Each digit is a weighted point set $\{x_n^j, p_n^j\}_{n=1}^{N_j}$, $j = 1, 2$, that we consider as a discrete measure of the form $\mu^j = \sum_{n=1}^{N_j} p_n^j \delta_{x_n^j} + 1/N_j \sum_{m=1}^{\eta N_j} \delta_{y_m^j}$, where the first sum corresponds to the clean data normalized to have total mass 1, and the second sum is constructed with samples from a uniform distribution acting as noise with total mass η. For HK, OPT, LHK and LOPT, we use the distributions μ^j without re-normalization, while for OT and LOT, we re-normalize them. The reference in LOT, LHK and LOPT is taken as the OT barycenter of a sample of the digits 0, 1, and 9 not including the ones used for interpolation, and normalized to have unit total mass. We test for $\eta = 0, 0.5, 0.75$ (see Figure 8 in the Appendix H). The results for $\eta = 0.5$ are shown in Figure 4. We can see that OT and LOT do not eliminate noise points. HK, OPT still retains much of the noise because interpolation is essentially between μ^1 and $\hat{\mu}^2$ (with respect to μ^1). So μ^1 acts as a reference that still has a lot of noise. In LHK, LOPT, by selecting the same reference as LOT we see that the noise significantly decreases. In the HK and LHK cases, we notice not only how the masses vary, but also how their relative positions change obtaining a very different configuration at the end of the interpolation. OPT and LOPT instead returns a more natural interpolation because of the mass preservation of the transported portion and the decoupling between transport and destruction/creation of mass.

PCA analysis: We compare the results of performing PCA on the embedding

Figure 4: We demonstrate the OT geodesic, OPT interpolation, LOT geodesic and LOPT interpolation in MNIST dataset. In LOT geodesic and LOPT interpolation, we use the same reference measure. The percentage of noise η is set to 0.5. In OPT and LOPT interpolation, we set $\lambda = 20$; in HK and LHK, we set the scaling to be 2.5.

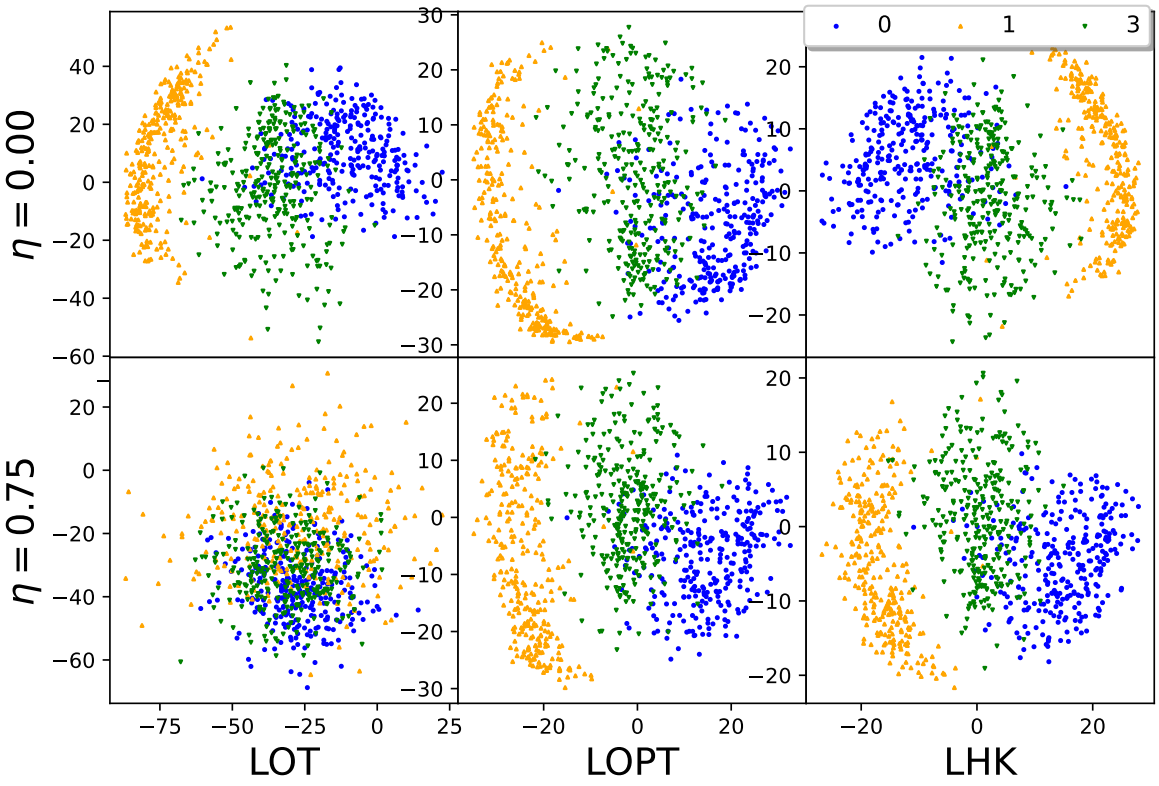


Figure 5: We plot the first two principal components of each u^j based on LOT and LOPT. For LOPT, we set $\lambda = 20.0$, and for LHK, we set the scaling to be 2.5.

space of LOT, LHK and LOPT for point cloud MNIST. We take 900 digits from the dataset corresponding to digits $0, 1$ and 3 in equal proportions. Each element is a point set $\{x_n^j\}_{n=1}^{N_j}$ that we consider as a discrete measure with added noise. The reference, μ^0, is set to the OT barycenter of 30 samples from the clean data. For LOT we re-normalize each μ^j to have a total mass of 1, while we do not re-normalize for LOPT. Let $S_\eta := \{\mu^j : \text{noise level} = \eta\}_{j=1}^{900}$. We embed S_η using LOT, LHK and LOPT and apply PCA on the embedded vectors $\{u^j\}$. In Figure 5 we show the first two principal components of the set of embedded vectors based on LOT, LHK and LOPT for noise levels $\eta = 0, 0.75$. It can be seen that when there is no noise, the PCA dimension reduction technique works well for all three embedding methods. When $\eta = 0.75$, the method fails for LOT embedding, but the dimension-reduced data is still separable for LOPT and LHK. For the running time, LOT, LOPT requires 60-80 seconds and LHK requires about 300-350 seconds. The experiments are conducted on a Linux computer with AMD EPYC 7702P CPU with 64 cores and 256GB DDR4 RAM.

We refer the reader to Appendix H for further details and analysis.

5 Summary

We proposed a Linear Optimal Partial Transport (LOPT) technique that allows us to embed distributions with different masses into a fixed dimensional space in which several calculations are significantly simplified. We show how to implement this for real data distributions allowing us to reduce the computational cost in applications that would benefit from the use of optimal (partial) transport. We finally provide comparisons with previous techniques and show some concrete applications. In particular, we show that LOPT is more robust and computationally efficient in the presence of noise than previous methods. For future work, we will continue to investigate the comparison of LHK and LOPT, and the potential applications of LOPT in other machine learning and data science tasks, such as Barycenter problems, graph embedding, task similarity measurement in transfer learning, and so on.

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contracts No. HR00112190132 and No. HR00112190135. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force, DARPA, or other funding agencies.

References

[1] Akram Aldroubi, Shiying Li, and Gustavo K Rohde. Partitioning signal classes using transport transforms for data analysis and machine learning. *Sampling Theory, Signal Processing, and Data Analysis*, 19(1):1–25, 2021.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[4] Yikun Bai, Bernard Schmitzer, Mathew Thorpe, and Soheil Kolouri. Sliced optimal partial transport. *arXiv preprint arXiv:2212.08049*, 2022.

[5] Saurav Basu, Soheil Kolouri, and Gustavo K Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014.

[6] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

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

[8] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

[9] Tianji Cai, Junyi Cheng, Nathaniel Craig, and Katy Craig. Linearized optimal transport for collider events. *Physical Review D*, 102(11):116019, 2020.

[10] Tianji Cai, Junyi Cheng, Bernhard Schmitzer, and Matthew Thorpe. The linearized hellinger–kantorovich distance. *SIAM Journal on Imaging Sciences*, 15(1):45–83, 2022.

[11] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[12] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[13] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

[14] Lenaic Chizat, Bernhard Schmitzer, Gabriel Peyre, and Francois-Xavier Vialard. An interpolating distance between optimal transport and fisher-rao. *arXiv:1506.06430 [math]*, 2015.

[15] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

[16] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[17] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[18] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[19] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.

[20] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

[21] Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

[22] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. European Mathematical Society, 2021.

[23] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[24] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Gan and vae from an optimal transport point of view. *arXiv preprint arXiv:1706.01807*, 2017.

[25] Kevin Guittet. *Extended Kantorovich norms: a tool for optimization*. PhD thesis, INRIA, 2002.

[26] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein regularization for sparse multi-task regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1407–1416. PMLR, 2019.

[27] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

[28] Sanggyun Kim, Rui Ma, Diego Mesa, and Todd P Coleman. Efficient bayesian inference methods via convex optimization and optimal transport. In *2013 IEEE International Symposium on Information Theory*, pages 2259–2263. IEEE, 2013.

[29] Soheil Kolouri, Navid Naderializadeh, Gustavo K Rohde, and Heiko Hoffmann. Wasserstein embedding for graph learning. In *International Conference on Learning Representations*, 2020.

[30] Shinjini Kundu, Soheil Kolouri, Kirk I Erickson, Arthur F Kramer, Edward McAuley, and Gustavo K Rohde. Discovery and visualization of structural biomarkers from mri using transport-based morphometry. *NeuroImage*, 167:256–275, 2018.

[31] Jan Lellmann, Dirk A Lorenz, Carola Schonlieb, and Tuomo Valkonen. Imaging with kantorovich–rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[32] Matthias Liero, Alexander Mielke, and Giuseppe Savare. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[33] Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4832–4841, 2019.

[34] Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: Provable wasserstein classification for certain rigid transformations and

perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023.

[35] Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-wasserstein embedding. *Advances in Neural Information Processing Systems*, 34:3389–3400, 2021.

[36] Khai Nguyen, Dang Nguyen, and Nhat Ho. Self-attention amortized distributional projection optimization for sliced wasserstein point-cloud reconstruction. *arXiv preprint arXiv:2301.04791*, 2023.

[37] Khai Nguyen, Dang Nguyen, Tung Pham, Nhat Ho, et al. Improving mini-batch optimal transport via partial transportation. In *International Conference on Machine Learning*, pages 16656–16690. PMLR, 2022.

[38] F. Santambrogio. *Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs and Modeling*. Birkhäuser, 2015.

[39] Meyer Scetbon and marco cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[40] Cedric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Society, 2003.

[41] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

[42] Jianbo Ye, Panruo Wu, James Z Wang, and Jia Li. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.

Appendix

We refer to the main text for references.

A Notation

- $(\mathbb{R}^d, \|\cdot\|)$, d-dimensional Euclidean space endowed with the standard Euclidean norm $\|x\| = \sqrt{\sum_{k=1}^{d} |x(k)^2|}$. $d(\cdot,\cdot)$ is the associated distance, i.e., $d(x,y) = \|x-y\|$.
- $x \cdot y$ canonical inner product in $\mathbb{R}^d$.
- 1_N, column vector of size $N \times 1$ with all entries equal to 1.
- $\mathrm{diag}(p_1, \ldots, p_N)$, diagonal matrix.
- w^T, transpose of a matrix w.
- $\mathbb{R}_+$, non-negative real numbers.
- Ω, convex and compact subset of $\mathbb{R}^d$. $\Omega^2 = \Omega \times \Omega$.
- $\mathcal{P}(\Omega)$, set of probability Borel measures defined in Ω.
- $\mathcal{M}_+(\Omega)$, set of positive finite Borel measures defined in Ω.
- $\mathcal{M}$, set of signed measures.
- $\pi_0 : \Omega^2 \to \Omega$, $\pi_0(x^0, x^1) := x^0$; $\pi_1 : \Omega^2 \to \Omega$, $\pi_1(x^0, x^1) := x^1$, standard projections.
- $F_{\#}\mu$, push-forward of the measure μ by the function. $F_{\#}\mu(B) = \mu(F^{-1}(B))$ for all measurable set B, where $F^{-1}(B) = \{x : F(x) \in B\}$.
- δ_x, Dirac measure concentrated on x.
- $\mu = \sum_{n=1}^{N} p_n \delta_{x_n}$, discrete measure ($x_n \in \Omega$, $p_n \in \mathbb{R}_+$). The coefficients p_n are called the weights of μ.
- $\mathrm{supp}(\mu)$, support of the measure μ.
- $\mu^i \wedge \mu^j$ minimum measure between μ^i and μ^j; $p^i \wedge p^j$ vector having at each n entry the minimum value p_n^i and p_n^j.
- μ^0 reference measure. μ^i, μ^j target measures. In the OT framework, they are in $\mathcal{P}(\Omega)$. In the OPT framework, they are in $\mathcal{M}_+(\Omega)$.
- $\Gamma(\mu^0, \mu^j) = \{\gamma \in \mathcal{P}(\Omega^2) : \pi_{0\#}\gamma = \mu^0, \pi_{1\#}\gamma = \mu^j\}$, set of Kantorovich transport plans.

- $C(\gamma; \mu^0, \mu^j) = \int_{\Omega^2} \|x^0 - x^j\|^2 d\gamma(x^0, x^j)$, Kantorovich cost given by the transportation plan γ between μ^0 and μ^j.
- $\Gamma^*(\mu^0, \mu^j)$, set of optimal Kantorovich transport plans.
- T, optimal transport Monge map.
- id, identity map $\mathrm{id}(x) = x$.
- $T_t(x) = (1-t)x + tT(x)$ for $T : \Omega \to \Omega$. Viewed as a function of (t, x), it is a flow.
- In section 2: $\rho \in \mathcal{P}([0,1] \times \Omega)$ curve of measures. At each $0 \leq t \leq$, $\rho_t \in \mathcal{P}(\Omega)$. In section 3: analogous, replacing $\mathcal{P}$ by $\mathcal{M}_+$.
- $v : [0,1] \times \Omega \to \mathbb{R}^d$. at each time $v_t : \Omega \to \mathbb{R}^d$ is a vector field. v_0, initial velocity.
- $\nabla \cdot V$, divergence of the vector field V with respect to the spatial variable x.
- $\partial_t \rho + \nabla \cdot \rho v = 0$, continuity equation.
- $\mathcal{CE}(\mu^0, \mu^j)$, set of solutions of the continuity equation with boundary conditions μ^0 and μ^j.
- $\partial_t \rho + \nabla \cdot \rho v = \zeta$, continuity equation with forcing term ζ.
- $\mathcal{FCE}(\mu^0, \mu^j)$, set solutions (ρ, v, ζ) of the continuity equation with forcing term with boundary conditions μ^0 and μ^j.
- $OT(\mu^0, \mu^j)$, optimal transport minimization problem. See (1) for the Kantorovich formulation, and (6) for the dynamic formulation.
- $W_2(\mu^0, \mu^j) = \sqrt{OT(\mu^0, \mu^j)}$, 2-Wasserstein distance.
- $\mathcal{T}_\mu = L^2(\Omega; \mathbb{R}^d, \mu)$, where $\|u\|_\mu^2 = \int_\Omega \|u(x)\|^2 d\mu(x)$ (if μ is discrete, $\mathcal{T}_\mu$ is identified with $\mathbb{R}^{d \times N}$ and $\|u\|_\mu^2 = \sum_{n=1}^N \|u(n)\|^2 p_n$).
- $LOT_{\mu^0}(\mu^i, \mu^j)$, see (14) for continuous densities, and (19) for discrete measures.
- $\lambda > 0$, penalization in OPT.
- $\mu \leq \nu$ if $\mu(B) \leq \nu(E)$ for all measurable set E and we say that μ is dominated by ν.
- $\Gamma_\leq(\mu^0, \mu^j) = \{\gamma \in \mathcal{M}_+(\Omega^2) : \pi_{0\#}\gamma \leq \mu^0, \pi_{1\#}\gamma \leq \mu^j\}$, set of partial transport plans.
- $\gamma_0 := \pi_{0\#}\gamma$, $\gamma_1 := \pi_{1\#}\gamma$, marginals of $\gamma \in \mathcal{M}_+(\Omega^2)$.
- $\nu_0 = \mu^0 - \gamma_0$ (for the reference μ^0), $\nu^j = \mu^j - \gamma_1$ (for the target μ^j).

- $C(\gamma;\mu^0,\mu^j,\lambda)=\int\|x^0-x^j\|^2d\gamma(x^0,x^j)+\lambda(|\mu^0-\gamma_0|+|\mu^j-\gamma_1|)$, partial transport cost given by the partial transportation plan γ between μ^0 and μ^j with penalization λ.
- $\Gamma^*_{\leq}(\mu^0,\mu^j)$, set of optimal partial transport plans.
- $OPT(\mu^0,\mu^j)$, partial optimal transport minimization problem between μ^0 and μ^j. See (23) for the static formulation, and (25) for the dynamic formulation.
- $\|v\|^2_{\mu,2\lambda}:=\int_\Omega\min(\|v\|^2,2\lambda)d\mu$
- $LOPT_\lambda(\mu^i,\mu^j)$, see (35), and (42) and the discussion above.
- $\mu^j\mapsto u^j$, LOT embedding (fixing first a reference $\mu^0\in\mathcal{P}(\Omega)$). If μ^0 has continuous density, $u^j:=v^j_0=T^j-\mathrm{id}$, and so it is a map from measures $\mu^j\in\mathcal{P}(\Omega)$ to vector fields u^i defined in Ω, see (12). For discrete measures (μ^0, μ^j discrete), LOT embedding is needed first, and in this case, the embedding is a map from discrete probability to $\mathbb{R}^{d\times N_0}$, see (17).
- μ, OT and OPT barycentric projection of μ with respect to a reference μ^0 (in $\mathcal{P}$ or $\mathcal{M}_+$, resp.), see (16) and (36), respectively. $\hat{x}_n$ denote the new locations where $\hat{\mu}$ is concentrated. p^0_n and $\hat{p}_n$ denote the wights of $\hat{\mu}$ for OT and OPT, respectively.
- $u_t=(1-t)u^i+tu^j$, geodesic in LOT embedding space.
- $\hat{\rho}_t$, LOT geodesic.
- LOPT embeding, see (33), and (40).
- OPT interpolating curve, and LOPT interpolating curve are defined in section 3.5.

B Proof of lemma 2.1

Proof. We fix $\gamma^*\in\Gamma^*(\mu^0,\mu^j)$[14], and then we compute the map $x^0_n\mapsto\hat{x}^j_n$ according to (16). This map induces a transportation plan $\hat{\gamma}^*:=\mathrm{diag}(p^0_1,\dots,p^0_{N_0})\in\mathbb{R}_+^{N_0\times N_0}$. We will prove that $\hat{\gamma}^*\in\Gamma^*(\mu^0,\hat{\mu}^j)$. By definition, it is easy to see that its marginals are μ^0 and $\hat{\mu}^j$. The complicated part is to prove optimality. Let $\hat{\gamma}\in\Gamma(\mu^0,\hat{\mu}^j)$ be an arbitrary transportation plan between μ^0 and μ^j (i.e. $\gamma 1_{N_0}=\gamma^T1_{N_0}=p^0$). We need to show

$$C(\hat{\gamma}^*;\mu^0,\hat{\mu}^j)\leq C(\hat{\gamma};\mu^0,\hat{\mu}^j). \tag{43}$$

Since $\hat{\gamma}^*$ is a diagonal matrix with positive diagonal, its inverse matrix is $(\hat{\gamma}^*)^{-1}=\mathrm{diag}(1/p^0_1,\dots,1/p^0_{N_0})$. We define

$$\gamma:=\hat{\gamma}(\hat{\gamma}^*)^{-1}\gamma^*\in\mathbb{R}_+^{N_0\times N_j}.$$

[14]Although in subsection 2.4 we use the notation $\gamma^j\in\Gamma^*(\mu^0,\mu^j)$, here we use the notation γ^* to emphasize that we are fixing an **optimal** plan.

We claim $\gamma \in \Gamma(\mu^0, \mu^j)$. Indeed,

$$\gamma 1_{N_j} = \hat{\gamma}(\hat{\gamma}^*)^{-1}\gamma^* 1_{N_j} = \hat{\gamma}(\hat{\gamma}^*)^{-1}p^0 = \hat{\gamma}1_{N_0} = p^0 \tag{44}$$

$$\gamma^T 1_{N_0} = (\gamma^*)^T(\hat{\gamma}^*)^{-1}\hat{\gamma}^T 1_{N_0} = (\gamma^*)^T(\hat{\gamma}^*)^{-1}p^0 = (\gamma^*)^T 1_{N_0} = p^j, \tag{45}$$

where the second equality in (44) and the third equality in (45) follow from the fact $\gamma^* \in \Gamma(\mu^0, \mu^1)$, and the fourth equality in (44) and the second equality in (45) hold since $\hat{\gamma} \in \Gamma(\mu^0, \hat{\mu}^j)$.

Since γ^* is optimal for $OT(\mu, \mu^j)$,we have $C(\gamma^*; \mu^0, \mu^j) \leq C(\gamma; \mu^0, \mu^j)$ to denote the transportation cost for γ^* and γ, respectively. We write

$$\begin{aligned} C(\gamma^*; \mu^0, \mu^j) &= \sum_{n=1}^{N_0}\sum_{m=1}^{N_j} \|x_n^0 - x_m^j\|^2 \gamma^*_{n,m} \\ &= \sum_{n=1}^{N_0} \|x_n^0\|^2 p_n^0 + \sum_{m=1}^{N_j} \|x_m^j\|^2 p_m^j - 2\sum_{n=1}^{N_0}\sum_{m=1}^{N_j} x_n \cdot x_m^j \, \gamma^*_{n,m} \\ &= K_1 - 2\sum_{n=1}^{N_0}\sum_{m=1}^{N_j} \hat{x}_n \cdot x_m^j \, \gamma^*_{n,m}, \end{aligned}$$

where K_1 is a constant (which only depends on μ^0, μ^j). Similarly,

$$\begin{aligned} C(\gamma; \mu^0, \mu^j) &= K_1 - 2\sum_{n=1}^{N_0}\sum_{m=1}^{N_j} x_n^0 \cdot x_m^j \, \gamma_{n,m} \\ &= K_1 - 2\sum_{n=1}^{N_0}\sum_{m=1}^{N_j}\sum_{\ell=1}^{N_0} \frac{\hat{\gamma}_{n,\ell}\gamma^*_{\ell,m}}{p_\ell^0} \, x_n^0 \cdot x_m^j \end{aligned}$$

Analogously, we have

$$\begin{aligned} C(\hat{\gamma}^*; \mu^0, \hat{\mu}^j) &= \sum_{n=1}^{N_0}\sum_{m=1}^{N_0} \|\hat{x}_n - \hat{x}_m^j\|^2 \hat{\gamma}^*_{n,m} \\ &= \sum_{n=1}^{N_0} \|x_n^0\|^2 p_n^0 + \sum_{m=1}^{N_0} \|\hat{x}_m^j\|^2 p_m^0 - 2\sum_{n=1}^{N_0}\sum_{m=1}^{N_0} x_n^0 \cdot \hat{x}_m^j \, \hat{\gamma}^*_{n,m} \\ &= K_2 - 2\sum_{n=1}^{N_0}\sum_{m=1}^{N_j} x_n^0 \cdot x_m^j \, \gamma^*_{n,m} \\ &= K_2 - K_1 + C(\gamma^*; \mu^0, \mu^j) \end{aligned} \tag{46}$$

where K_2 is constant (which only depends on μ^0 and $\hat{\mu}^j$) and similarly

$$\begin{aligned} C(\hat{\gamma}; \mu^0, \hat{\mu}^j) &= K_2 - 2\sum_{n=1}^{N_0}\sum_{m=1}^{N_j}\sum_{\ell=1}^{N_0} \frac{\hat{\gamma}_{n,\ell}\gamma^*_{\ell,m}}{p_\ell^0} \, x_n^0 \cdot x_m^j \\ &= K_2 - K_1 + C(\gamma; \mu^0, \mu^1) \end{aligned} \tag{47}$$

Therefore, by (46) and (47), we have that $C(\gamma^*;\mu^0,\mu^j) \le C(\gamma;\mu^0,\mu^j)$ if and only if $C(\hat{\gamma}^*;\mu^0,\hat{\mu}^j) \le C(\hat{\gamma};\mu^0,\hat{\mu}^j)$. So, we conclude the proof by the fact γ^* is optimal for $OT(\mu_0,\mu^j)$. □

C Proof of Proposition 2.2

Lemma C.1. *Given discrete measures $\mu^0 = \sum_{k=1}^{N_0} p_k^0 \delta_{x_k^0}, \mu^1 = \sum_{k=1}^{N_0} p_k^0 \delta_{x_k^1}, \mu^2 = \sum_{k=1}^{N_0} p_k^0 \delta_{x_k^2}$, suppose that the maps $x_k^0 \mapsto x_k^1$ and $x_k^0 \mapsto x_k^2$ solve $OT(\mu^0,\mu^1)$ and $OT(\mu^0,\mu^2)$, respectively. For $t \in [0,1]$, define $\mu_t := \sum_{k=1}^{N_0} p_k^0 \delta_{(1-t)x_k^1 + t x_k^2}$. Then, the mapping $T_t : x_k^0 \mapsto (1-t)x_k^1 + t x_k^2$ solves the problem $OT(\mu^0,\mu_t)$.*

Proof. Let $\gamma^* = \mathrm{diag}(p_1^0,\dots,p_{N_0}^0)$ be the corresponding transportation plan induced by T_t. Consider an arbitrary $\gamma \in \mathbb{R}_+^{N_0 \times N_0}$ such that $\gamma \in \Gamma(\mu^0,\mu_t)$. We need to show

$$C(\gamma^*;\mu^0,\mu_t) \le C(\gamma;\mu^0,\mu_t).$$

We have

$$\begin{aligned} C(\gamma^*;\mu^0,\mu_t) &= \sum_{k,k'=1}^{N_0} \|x_k^0 - (1-t)x_{k'}^1 - t x_{k'}^2\|^2 \gamma_{k,k'}^* \\ &= \sum_{k=1}^{N_0} \|x_k^0\|^2 p_k^0 + \sum_{k=1}^{N_0} \|(1-t)x_{k'}^1 + t x_{k'}^2\|^2 p_k^0 - 2 \sum_{k,k'=1}^{N_0} x_k^0 \cdot ((1-t)x_{k'}^1 + t x_{k'}^2) \\ &= K - 2\left[(1-t)\sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^1 \gamma_{k,k'}^* + t \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^2 \gamma_{k,k'}^*\right], \end{aligned} \tag{48}$$

where in third equation, K is a constant which only depends on the marginals μ^0,μ_t. Similarly,

$$C(\gamma;\mu^0,\mu_t) = K - 2\left[(1-t)\sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^1 \gamma_{k,k'} + t \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^2 \gamma_{k,k'}\right]. \tag{49}$$

By the fact that $\gamma,\gamma^* \in \Gamma(\hat{\mu},\mu_t) = \Gamma(\mu^0,\mu^1) = \Gamma(\mu^0,\mu^2)$, and that γ^* is optimal for $OT(\hat{\mu},\mu^1), OT(\hat{\mu},\mu^2)$, we have

$$\sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^1 \gamma_{k,k'}^* \ge \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^1 \gamma_{k,k'} \qquad \text{and} \qquad \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^2 \gamma_{k,k'}^* \ge \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_{k'}^2 \gamma_{k,k'}.$$

Thus, by (48) and (49), we have $C(\gamma^*;\hat{\mu},\mu_t) \le C(\gamma;\hat{\mu},\mu_t)$, and this completes the proof. □

Proof of Proposition 2.2. Consider $0 \leq s \leq t \leq 1$. By Lemma 2.1, the maps $T^i : x_k^0 \mapsto \hat{x}_k^i$, $T^j : x_k^0 \mapsto \hat{x}_k^j$ solve $OT(\mu^0, \hat{\mu}^i), OT(\mu^0, \hat{\mu}^j)$, respectively. Moreover, by Lemma C.1 (under the appropriate renaming), we have the mapping

$$x_k^0 \mapsto (1-s)\hat{x}_k^i + s\hat{x}_k^j = x_k^0 + (1-s)u_k^i + su_k^j, \qquad 1 \leq k \leq N_0$$

solves $OT(\mu^0, \hat{\rho}_s)$, and similarly

$$x_k^0 \mapsto (1-t)\hat{x}_k^i + t\hat{x}_k^j = x_k^0 + (1-t)u_k^i + tu_k^j, \qquad 1 \leq k \leq N_0$$

solves $OT(\mu^0, \hat{\rho}_t)$.

Thus,

$$\begin{aligned} LOT_{\mu^0}(\hat{\rho}_s, \hat{\rho}_t) &= \sum_{k=1}^{N_0} \|((1-s)\hat{x}_k^i + s\hat{x}_k^j) - ((1-t)\hat{x}_k^i + t\hat{x}_k^j)\|^2 \, p_k^0 \\ &= \sum_{k=1}^{N_0} (t-s)^2 \|\hat{x}_k^i - \hat{x}_k^j\|^2 \, p_k^0 \\ &= (t-s)^2 LOT_{\mu^0}(\mu^i, \mu^j), \end{aligned}$$

and this finishes the proof. □

D Proof of proposition 3.1

This result relies on the definition (29) of the curve ρ_t, which is inspired by the fact that solutions of non-homogeneous equations are given by solutions of the associated homogeneous equation plus a particular solution of the non-homogeneous. We choose the first term of ρ_t as a solution of a (homogeneous) continuity equation (γ_t defined in (28)), and the second term ($\overline{\gamma}_t := (1-t)\nu_0 + t\nu^j$) as an appropriate particular solution of the equation (26) with forcing term ζ defined in (31). Moreover, the curve ρ_t will become optimal for (25) since both γ_t and $\overline{\gamma}_t$ are 'optimal' in different senses. On the one hand, γ_t is optimal for a classical optimal transport problem[15]. On the other hand, $\overline{\gamma}_t$ is defined as a *line* interpolating the new part introduced by the framework of partial transportation, 'destruction and creation of mass'.

Although this is the core idea behind the proposition, to prove it we need several lemmas to deal with the details and subtleties.

First, we mention that, the push-forward measure $F_\# \mu$ can be defined satisfying the formula of change of variables

$$\int_A g(x)\, dF_\#\mu(x) = \int_{F^{-1}(A)} g(F(x))\, d\mu(x)$$

[15]Here we will strongly use that the fixed partial optimal transport plan $\gamma^* \in \Gamma^*_{\leq}(\mu^0, \mu^j)$ in the hypothesis of the proposition has the form $\gamma^* = (I \times T)_\# \gamma_0^*$, for a map $T : \Omega \to \Omega$

for all measurable set A, and all measurable function g, where $F^{-1}(A) = \{x : F(x) \in A\}$. This is a general fact, and as an immediate consequence, we have conservation of mass

$$|F_{\#}\mu| = |\mu|.$$

Then, in our case, the second term in the cost function (24) we can be written as

$$\lambda(|\mu^0 - \gamma_0| + |\mu^j - \gamma_1| = \lambda(|\mu^0| + |\mu^1| - 2|\gamma|)$$

since γ_0 and γ_1 are *dominated by* μ^0 and μ^j, respectively, in the sense that $\gamma_0 \leq \mu^0$ and $\gamma_1 \leq \mu^j$.

Finally, we would like to point out that when we say that (ρ, v, ζ) is a solution for equation (26), we mean that it is a *weak solution* or, equivalently, it is a *solution in the distributional sense* [38, Section 4 4]. That is, for any test function $\psi : \Omega \to \mathbb{R}$ continuous differentiable with compact support, (ρ, v, ζ) satisfies

$$\partial_t \left(\int_\Omega \psi \, d\rho_t \right) - \int_\Omega \nabla\psi \cdot v_t \, d\rho_t = \int_\Omega \psi \, d\zeta_t,$$

or, equivalently, for any test function $\phi : [0,1] \times \Omega \to \mathbb{R}$ continuous differentiable with compact support, (ρ, v, ζ) satisfies

$$\int_{[0,1]\times\Omega} \partial_t\phi \, d\rho + \int_{[0,1]\times\Omega} \nabla\phi \cdot v \, d\rho + \int_{[0,1]\times\Omega} d\zeta = \int_\Omega \phi(1,\cdot) \, d\mu^j - \int_\Omega \phi(0,\cdot) \, d\mu^0.$$

Lemma D.1. *If γ^* is optimal for $OPT_\lambda(\mu^0, \mu^1)$, and has marginals γ_0^* and γ_1^*, then γ^* is optimal for $OPT_\lambda(\mu^0, \gamma_1^*)$, $OPT_\lambda(\gamma_0^*, \mu^1)$ and $OT(\gamma_0^*, \gamma_1^*)$.*

Proof. Let $C(\gamma; \mu^0, \mu^1, \lambda)$ denote the objective function of the minimization problem $OPT_\lambda(\mu^0, \mu^1)$, and similarly consider $C(\gamma; \gamma_0^*, \mu^1, \lambda), C(\gamma; \mu^0, \gamma_1^*, \lambda)$. Also, let $C(\gamma, \gamma_0^*, \gamma_1^*)$ be the objective function of $OT(\gamma_0^*, \gamma_1^*)$.

First, given an arbitrary plan $\gamma \in \Gamma(\gamma_0^*, \gamma_1^*) \subset \Gamma_\leq(\mu^0, \mu^1)$, we have

$$\begin{aligned} C(\gamma; \mu^0, \mu^1, \lambda) &= \int_{\Omega^2} \|x^0 - x^1\|^2 d\gamma(x^0, x^1) + \lambda(|\mu^0 - \gamma_0| + |\mu^1 - \gamma_1|) \\ &= C(\gamma; \gamma_0^*, \gamma_1^*) + \lambda(|\mu^0 - \gamma_0| + |\mu^1 - \gamma_1|). \end{aligned}$$

Since $C(\gamma^*; \mu^0, \mu^1, \lambda) \leq C(\gamma; \mu^0, \mu^1, \lambda)$, we have $C(\gamma^*; \gamma_0^*, \gamma_1^*) \leq C(\gamma; \gamma_0^*, \gamma_1^*)$. Thus, γ^* is optimal for $OT(\gamma_0^*, \gamma_1^*)$.

Similarly, for every $\gamma \in \Gamma_\leq(\gamma_0^*, \mu^1)$, we have $C(\gamma; \mu^0, \mu^1, \lambda) = C(\gamma; \gamma_0^*, \mu^1) + \lambda|\mu^0 - \gamma_0^*|$ and thus γ^* is optimal for $OPT_\lambda(\gamma_0^*, \mu^1)$. Analogously, γ^* is optimal for $OPT_\lambda(\mu^0, \gamma_1^*)$. □

Lemma D.2. *If $\gamma^* \in \Gamma_\leq^*(\mu^0, \mu^1)$, then $d\left(\mathrm{supp}(\nu^0), \mathrm{supp}(\nu^1)\right) \geq \sqrt{2\lambda}$, where $\nu^0 = \mu^0 - \gamma_0$, $\nu^1 = \mu^1 - \gamma_1$, and d is the Euclidean distance in $\mathbb{R}^d$.*

Proof. We will proceed by contradiction. Assume that $d\left(\mathrm{supp}(\nu^0), \mathrm{supp}(\nu^1)\right) < \sqrt{2\lambda}$. Then, there exist $\widetilde{x}^0 \in \mathrm{supp}(\nu^0)$ and $\widetilde{x}^1 \in \mathrm{supp}(\nu^1)$ such that $\|\widetilde{x}^0 - \widetilde{x}^1\| < \sqrt{2\lambda}$. Moreover, we can choose $\varepsilon > 0$ such $\nu^0(B(\widetilde{x}^0, \varepsilon)) > 0$ and $\nu^1(B(\widetilde{x}^1, \varepsilon)) > 0$, where $B(\widetilde{x}^i, \varepsilon)$ denotes the ball in $\mathbb{R}^d$ of radius ε centered at $\widetilde{x}^i$.

We will construct $\widetilde{\gamma} \in \Gamma_{\leq}(\mu^0, \mu^1)$ with transportation cost strictly less than $OPT_\lambda(\mu^0, \mu^1)$, leading to a contradiction. For $i = 0, 1$ we denote by $\widetilde{\nu}^i$ the probability measure given by $\frac{1}{\nu^i(B(\widetilde{x}^i, \varepsilon))}\nu^i$ restricted to $B(\widetilde{x}^i, \varepsilon)$. Let $\delta = \min\{\nu^0(B(\widetilde{x}^0, \varepsilon)), \nu^1(B(\widetilde{x}^1, \varepsilon))\}$. Now, we consider

$$\widetilde{\gamma} := \gamma^* + \delta(\,\widetilde{\nu}^0 \times \widetilde{\nu}^1).$$

Then, $\widetilde{\gamma} \in \Gamma_{\leq}(\mu^0, \mu^1)$ because

$$\pi_{i\#}\widetilde{\gamma} = \gamma_i + \delta\widetilde{\nu}^i \leq \gamma_i + \nu^i = \mu^i \qquad \text{for } i = 0, 1.$$

The partial transport cost of $\widetilde{\gamma}$ is

$$\begin{aligned}
&C(\widetilde{\gamma}, \mu^0, \mu^1, \lambda) \\
&= \int_{\Omega^2} \|x^0 - x^1\|^2 \, d\widetilde{\gamma} + \lambda(|\mu^0| + |\mu^1| - 2|\widetilde{\gamma}|) \\
&= \int_{\Omega^2} \|x^0 - x^1\|^2 \, d\gamma + \delta \int_{\Omega^2} \|x^0 - x^1\|^2 \, d(\widetilde{\nu}^0 \times \widetilde{\nu}^1) + \lambda(|\mu^0| + |\mu^1| - 2|\gamma| - 2\delta|\widetilde{\nu}^0 \times \widetilde{\nu}^1|) \\
&= OPT_\lambda(\mu^0, \mu^1) + \delta \int_{B(\widetilde{x}^0, \varepsilon) \times B(\widetilde{x}^1, \varepsilon)} \|x^0 - x^1\|^2 \, d(\widetilde{\nu}^0 \times \widetilde{\nu}^1) - 2\lambda\delta \\
&< OPT_\lambda(\mu^0, \mu^1) + 2\lambda\delta - 2\lambda\delta = OPT_\lambda(\mu^0, \mu^1),
\end{aligned}$$

which is a contradiction.

□

Lemma D.3. *Using the notation and hypothesis of Proposition 3.1, for $t \in (0, 1)$ let $D_t := \{x : v_t(x) \neq 0\}$. Then, $\gamma_t \wedge \nu_0 \equiv \gamma_t \wedge \nu^j \equiv 0$ over D_t.*

Proof. We will prove this for $\gamma_t \wedge \nu_0$. The other case is analogous. Suppose by contradiction that $\gamma_t \wedge \nu_0$ is not the null measure. By Lemma D.1, we know that γ^* is optimal for $OT(\gamma_0^*, \gamma_1^*)$. Since we are also assuming that γ is induced by a map T, i.e. $\gamma^* = (I \times T)_\# \gamma_0^*$, then T is an optimal Monge map between γ_0^* and γ_1^*. Therefore, the map $T_t : \Omega \to \Omega$ is invertible (see for example [38, Lemma 5.29]). For ease of notation we will denote $\widetilde{\nu}^0 := \gamma_t \wedge \nu_0$ as a measure in D_t.

The idea of the proof is to exhibit a plan $\widetilde{\gamma}$ with smaller OPT_λ cost than γ^*, which will be a contradiction since $\gamma^* \in \Gamma^*_{\leq}(\mu^0\mu^j)$. Let us define

$$\widetilde{\gamma} := \gamma^* + (I, T \circ T_t^{-1})_\# \widetilde{\nu}^0 - (I, T)_\# (T_t^{-1})_\# \widetilde{\nu}^0.$$

Then, $\widetilde{\gamma} \in \Gamma_{\leq}(\mu^0, \mu^j)$ since

$$\begin{aligned}
&\pi_{0\#}\widetilde{\gamma} = \pi_{0\#}\gamma^* + \widetilde{\nu}^0 - (T_t^{-1})_\#\widetilde{\nu}^0 \leq \gamma_0^* + \widetilde{\nu}^0 \leq \mu^0, \\
&\pi_{1\#}\widetilde{\gamma} = \pi_{1\#}\gamma^* + (T \circ T_t^{-1})_\#\widetilde{\nu}^0 - (T \circ T_t^{-1})_\#\widetilde{\nu}^0 = \gamma_1^*. \qquad \text{(and we know } \gamma_1^* \leq \mu^j\text{)}.
\end{aligned}$$

From the last equation we can also conclude that $|\widetilde{\gamma}| = |\gamma_1^*| = |\gamma^*|$. Therefore, the partial transportation cost for $\widetilde{\gamma}$ is

$$\begin{aligned}
&\int_{\Omega^2} \|x-y\|^2 d\widetilde{\gamma}(x,y) + \lambda(|\mu^0| + |\mu^j| - 2|\widetilde{\gamma}|) \\
&\quad = \underbrace{\int_{\Omega^2} \|x-y\| d\gamma^*(x,y) + \lambda(|\mu^0| + |\mu^j| - 2|\gamma|)}_{OPT_\lambda(\mu^0,\mu^j)} \\
&\qquad + \int_{D_t} \left(\|x - T(T_t^{-1}(x))\|^2 - \|T_t^{-1}(x) - T(T_t^{-1}(x))\|^2 \right) d\widetilde{\nu}^0(x) \\
&\quad < OPT_\lambda(\mu^0, \mu^1)
\end{aligned}$$

where in the last inequality we used that if $y = T_t^{-1}(x)$ we have

$$\|T_t(y) - T(y)\| = \|(1-t)y + tT(y) - T(y)\| = (1-t)\|y - T(y)\| < \|y - T(y)\| \qquad \text{for } t \in (0,1).$$

□

Lemma D.4. *Using the notation and hypothesis of Proposition 3.1, for $t \in (0,1)$ the vector-valued measures $v_t \cdot \nu_0$ and $v_t \cdot \nu^j$ are exactly zero.*

Proof. We prove this for $v_t \cdot \nu_0$. The other case is analogous. We recall that the measure $v_t \cdot \nu_0$ is determined by

$$\int_\Omega \Phi(x) \cdot v_t(x)\, d\nu_0(x)$$

for all measurable functions $\Phi : \Omega \to \mathbb{R}^d$, where $\Phi(x) \cdot v_t(x)$ denotes the usual dot product in $\mathbb{R}^d$ between the vectors $\Phi(x)$ and $v_t(x)$.

From Lemma D.3 we know that $\gamma_t \wedge \nu_0 \equiv 0$ over D_t. Therefore, γ_t and ν_0 are mutually singular in that set. This implies that we can decompose D_t into two disjoint sets D_1, D_2 such that $\gamma_t(D_1) = 0 = \nu_0(D_2)$. Since v_t is a function defined γ_t–almost everywhere (up to sets of null measure with respect to γ_t), we can assume without loss of generality that $v_t \equiv 0$ on D_1.

Let $\Phi : \Omega \to \mathbb{R}^d$ be a measurable vector-valued function over Ω. Using that $v_t \equiv 0$ in $\Omega \setminus D_t$ and in D_1, and that $\nu_0(D_2) = 0$, we obtain

$$\int_\Omega \Phi(x) \cdot v_t(x)\, d\nu_0(x) = \int_{\Omega \setminus D_t} \Phi(x) \cdot v_t(x)\, d\nu_0(x) + \int_{D_1} \Phi(x) \cdot v_t(x)\, d\nu_0(x) + \int_{D_2} \Phi(x) \cdot v_t(x)\, d\nu_0(x) = 0.$$

Since Φ was arbitrary, we can conclude that $v_t \cdot \nu_0 \equiv 0$. □

Lemma D.5. *Using the notation and hypothesis of Proposition 3.1, the measure $\overline{\gamma}_t = (1-t)\nu_0 + t\nu^1$ satisfies the equation*

$$\begin{cases} \partial_t \overline{\gamma} + \nabla \cdot \overline{\gamma} v = \zeta \\ \overline{\gamma}_0 = \nu_0, \quad \overline{\gamma}_1 = \nu^1. \end{cases} \tag{50}$$

Proof. From Lemma D.4 we have that $v_t \cdot \overline{\gamma}_t = (1-t)v_t \cdot \nu_0 + t v_t \cdot \nu^1 = 0$, then $\nabla \cdot \overline{\gamma} v = 0$. It is easy to see that $\overline{\gamma}$ satisfies the boundary conditions by replacing $t = 0$ and $t = 1$ in its definition. Also, $\partial_t \overline{\gamma} = \nu^1 - \nu_0$. Then, since $\zeta_t = \nu^1 - \nu_0$ for every t, we get $\partial_t \overline{\gamma} + \nabla \cdot \overline{\gamma} v = \nu^1 - \nu_0 + 0 = \zeta_t$. □

Proof of Proposition 3.1.

First, we address that the $v : [0,1] \times \Omega \to \mathbb{R}^d$ is well defined. Indeed, by Lemma D.1, we have that $\gamma^* = (\mathrm{id} \times T)_\# \gamma_0^*$ is optimal for $OT(\gamma_0^*, \gamma_1^*)$. Thus, for each $(t, x) \in [0,1] \times \Omega$, by so-called *cyclical monotonicity property* of the support of classical optimal transport plans, there exists at most one $x_0 \in \mathrm{supp}(\gamma_0^*)$ such that $T_t(x_0) = x$, see [38, Lemma 5.29]. So, $v_t(x)$ in (30) is well defined by understanding its definition as

$$v_t(x) = \begin{cases} T(x_0) - x_0 & \text{if } x = T_t(x_0), \quad \text{for } x_0 \in \mathrm{supp}(\gamma_0^*) \\ 0 & \text{elsewhere.} \end{cases}$$

Now, we check that (ρ, v, ζ) defined in (29), (30), (31) is a solution for (26). In fact,

$$\rho_t = \gamma_t + \overline{\gamma}_t$$

where γ_t is given by (28), and $\overline{\gamma}_t$ is as in Lemma D.5. Then, from section 2.2, (γ, v) solves the continuity equation (5) since $\gamma^* = (I \times T)_\# \gamma_0^* \in \Gamma^*(\gamma_0^*, \gamma_1^*)$ (Lemma D.1), and from its definition $\gamma_t = (T_t)_\# \gamma_0^*$ coincides with (9) in this context, as well as v_t coincides with (8) (the support of v_t lies on the support of γ_0^*). On the other hand, from Lemma D.5, $\overline{\gamma}_t$ solves (50). Therefore, by linearity,

$$\partial_t \rho_t + \nabla \cdot \rho_t v_t = \underbrace{\partial_t \gamma_t + \nabla \cdot \gamma_t v_t}_{0} + \underbrace{\partial_t \overline{\gamma}_t + \nabla \cdot \overline{\gamma}_t v_t}_{\zeta_t} = \zeta_t$$

Finally, by plugging in (ρ, v, ζ) into the objective function in (25), we have:

$$\begin{aligned} \int_{[0,1]\times\Omega} \|v\|^2 d\rho + \lambda|\zeta| &= \int_{[0,1]\times\Omega} \|v\|^2 \, d\gamma_t \, dt + \lambda|\zeta| \\ &= \int_\Omega \|T(x^0) - x^0\|^2 \, d\gamma_0^*(x^0) + \lambda(|\nu_0| + |\nu^1|) \\ &= \int_{\Omega^2} \|x^0 - x^j\|^2 \, d\gamma^*(x^0, x^j) + \lambda(|\mu^0 - \gamma_0^*| + |\mu^j - \gamma_1^*|) \\ &= OPT_\lambda(\mu^0, \mu^j) \end{aligned}$$

since $\gamma^* \in \Gamma^*_{\leq}(\mu^0, \mu^j)$. So, this shows that (ρ, v, ζ) is minimum for (25).

The 'moreover' part holds from the above identities, using that

$$\begin{aligned} &\int_\Omega \|v_0\|^2 d\gamma_0^*(x_0) + \lambda(|\nu_0| + |\nu^1|) = \int_\Omega \min(\|v_0\|^2, 2\lambda) d\gamma_0^*(x_0) + \lambda(|\nu_0| + |\nu^1|) \\ &\text{for } v_0(x^0) = T(x^0) - x^0, \end{aligned}$$

since γ^* is such that $\|x^0 - x^j\|^2 < 2\lambda$ for all $(x^0, x^j) \in \mathrm{supp}(\gamma)$ [4, Lemma 3.2].

□

E Proof of Theorem 3.5

The proof will be similar to that of Lemma 2.1. To simplify the notation, in this proof, we use μ^1 (and $\hat{\mu}^1$) to replace μ^j (and $\hat{\mu}^j$).

E.1 Notation setup

We choose $\gamma^1 \in \Gamma^*_{\leq}(\mu^0, \mu^1)$. We will understand the second marginal distribution $\pi_{1\#}\gamma^1$ induced by γ^1, either as the vector $\gamma^1_1 := (\gamma^1)^T 1_{N_0}$ or as the measure $\sum_{m=1}^{N_1} (\gamma^1_1)_m \delta_{x^1_m}$, and, by abuse of notation, we will write $\pi_{1\#}\gamma^1 = (\gamma^1)^T 1_{N_0} = \gamma^1_1$ (analogously for $\pi_{0\#}\gamma^1$).

Let $\hat{\gamma}^1 := \text{diag}(\hat{p}^1_1, \ldots, \hat{p}^1_{N_0})$ denote the transportation plan induced by mapping $x^0_k \mapsto \hat{x}^1_k$.

Let

$$D := \{k \in \{1, \ldots N_0\} : \hat{p}^1_k > 0\}.$$

We select $\hat{\gamma} \in \Gamma_{\leq}(\mu^0, \hat{\mu}^1)$.

With a slight abuse of notation, we define

$$\gamma := \hat{\gamma}(\hat{\gamma}^1)^{-1}\gamma^j$$

where we mean that $(\hat{\gamma}^1)^{-1} \in \mathbb{R}^{N_0 \times N_0}$ is a digonal matrix with:

$$(\hat{\gamma}^1)_{kk} = \begin{cases} \frac{1}{\hat{p}^1_k} & \text{if } \hat{p}^1_k > 0 \\ 0 & \text{elsewhere} \end{cases}, \forall k \in [1 : N_0].$$

The goal is to show

$$C(\hat{\gamma}^j; \mu^0, \hat{\mu}^j, \lambda) \leq C(\hat{\gamma}; \mu^0, \hat{\mu}^j, \lambda). \tag{51}$$

E.2 Connect $OPT_\lambda(\mu^0, \mu^1)$ and $OPT_\lambda(\mu^0, \hat{\mu}^1)$.

Note, First, we claim $\gamma \in \Gamma_{\leq}(\mu^0, \gamma^1_1) \subset \Gamma_{\leq}(\mu^0, \mu^1)$. Indeed, we have

$$\gamma 1_{N_1} = \hat{\gamma}(\hat{\gamma}^1)^{-1}\gamma^j 1_{N_1} = \hat{\gamma}(\hat{\gamma}^1)^{-1}\hat{p}^j = \hat{\gamma}1_D = \hat{\gamma}_0 \leq p^0 \tag{52}$$

$$\gamma^T 1_{N_0} = (\gamma^1)^T(\hat{\gamma}^1)^{-1}\hat{\gamma}^T 1_{N_0} \leq (\gamma^1)^T(\hat{\gamma}^1)^{-1}\hat{p}^1 = (\gamma^1)^T 1_D = \gamma^1_1, \tag{53}$$

where $1_D \in \mathbb{R}^{N_0}$ is defined with $(1_D)_k = 1$ if $k \in D$ and $(1_D)_k = 0$ elsewhere.

In (52), the equality $\hat{\gamma}1_D = \hat{\gamma}_0$ follows from the following: $(\hat{\gamma}^1)^T 1_{N_0} \leq \hat{p}^1$, we have $\hat{\gamma}_{k,k'} = 0, \forall k' \notin D, k \in \{1, \ldots, N_0\}$. In (53), the inequality follows from the fact $\hat{\gamma} \in \Gamma_{\leq}(\mu^0, \hat{\mu}^1)$.

Note, from (52), we also have $\gamma^1_0 = \hat{\gamma}^1_0$.

We compute the transportation costs induced by $\gamma^1, \gamma, \hat{\gamma}^1, \hat{\gamma}$:

$$
\begin{aligned}
&C(\gamma^1; \mu^0, \mu^1, \lambda) \\
&= \sum_{i=1}^{N_1} \sum_{k \in D} \|x_k^0 - x_i^1\|^2 \gamma_{k,i}^1 + \lambda(|p^0| + |p^1| - 2|\gamma^1|) \\
&= \sum_{k \in D} \|x_k^0\|^2 (\gamma_0^1)_k + \sum_{i=1}^{N_1} \|x_i^1\|^2 (\gamma_1^1)_i - 2 \sum_{k \in D} \sum_{i=1}^{N_1} x_k^0 \cdot x_i^1 \, \gamma_{k,i}^1 + \lambda(|p^0| + |p^1| - 2|\gamma^1|)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&C(\hat{\gamma}^1; \mu^0, \hat{\mu}^1, \lambda) \\
&= \sum_{k \in D} \|x_k^0\|^2 (\hat{\gamma}_0^1)_k + \sum_{k' \in D} \|\hat{x}_{k'}^1\|^2 (\hat{\gamma}_1^1)_{k'} - 2 \sum_{k \in D} \sum_{k' \in D} x_k^0 \cdot \hat{x}_{k'}^1 \, \hat{\gamma}_{k,k'}^1 + \lambda(|p^0| + |\hat{p}^1| - 2|\hat{\gamma}^1|) \\
&= \sum_{k \in D} \|x_k^0\|^2 (\hat{\gamma}_0^1)_k + \sum_{k' \in D} \|\hat{x}_{k'}^1\|^2 (\hat{\gamma}_1^1)_{k'} - 2 \sum_{k \in D} \sum_{k' \in D} \sum_{i=1}^{N_1} x_k^0 \cdot x_i^1 \, \gamma_{k',i}^1 \frac{\hat{\gamma}_{k,k'}^1}{\hat{\gamma}_{k',k'}^1} + \lambda(|p^0| + |\hat{p}^1| - 2|\hat{\gamma}^1|) \\
&= \sum_{k \in D} \|x_k^0\|^2 (\hat{\gamma}_0^1)_k + \sum_{k' \in D} \|\hat{x}_{k'}^1\|^2 (\hat{\gamma}_1^1)_{k'} - 2 \sum_{k \in D} \sum_{i=1}^{N_1} x_k^0 \cdot x_i^1 \, \gamma_{k,i}^1 + \lambda(|p^0| + |\hat{p}^1| - 2|\hat{\gamma}^1|)
\end{aligned}
$$

Therefore, we obtain

$$
C(\hat{\gamma}^1; \mu^0, \hat{\mu}^1, \lambda) = C(\gamma^1; \mu^0, \gamma_1^1, \lambda) - \sum_{i=1}^{N_1} \|x_i^1\|^2 (\gamma_1^1)_i + \sum_{k' \in D} \|\hat{x}_{k'}^1\|^2 (\hat{\gamma}_1^1)_{k'} + \lambda(|\hat{p}^1| - |p^1|) \tag{54}
$$

And also,

$$
\begin{aligned}
&C(\gamma; \mu^0, \mu^1, \lambda) \\
&= \sum_{k=1}^{N_0} \|x_k^0\|^2 (\gamma_0)_k + \sum_{i=1}^{N_1} \|x_i^1\|^2 (\gamma_1)_i - 2 \sum_{k=1}^{N_0} \sum_{i=1}^{N_1} \sum_{k' \in D} x_k^0 \cdot x_i^1 \, \hat{\gamma}_{k,k'} \frac{\gamma_{k',i}^1}{\hat{\gamma}_{k',k'}^1} + \lambda(|p^0| + |p^1| - 2|\gamma|),
\end{aligned}
$$

$$
\begin{aligned}
&C(\hat{\gamma}; \mu^0, \hat{\mu}^1, \lambda) \\
&= \sum_{k=1}^{N_0} \|x_k^0\|^2 (\hat{\gamma}_0)_k + \sum_{k' \in D} \|\hat{x}_{k'}^1\|^2 (\hat{\gamma}_1)_{k'} - 2 \sum_{k=1}^{N_0} \sum_{k' \in D} \sum_{i=1}^{N_0} x_k^0 \cdot x_i^1 \, \hat{\gamma}_{k,k'}^1 \frac{\gamma_{k',i}^1}{\hat{\gamma}_{k',k'}^1} + \lambda(|p^0| + |\hat{p}^1| - 2|\hat{\gamma}|)
\end{aligned}
$$

Thus we obtain

$$
C(\hat{\gamma}; \mu^0, \hat{\mu}^1, \lambda) = C(\gamma; \mu^0, \gamma_1^1, \lambda) - \sum_{i=1}^{N_1} \|x_i^1\|^2 (\gamma_1)_i + \sum_{k' \in D} \|\hat{x}_{k'}^1\|^2 (\hat{\gamma}_1)_{k'} + \lambda(|\hat{p}^1| - |p^1|) \tag{55}
$$

Combining with (54) and (55), we have

$$
\begin{aligned}
&C(\hat{\gamma}^1;\mu^0,\hat{\mu}^1,\lambda) - C(\hat{\gamma};\mu^0,\hat{\mu}^1,\lambda) \\
&= C(\gamma^1;\mu^0,\gamma_1^1,\lambda) - C(\gamma;\mu^0,\gamma_1^1,\lambda) + \sum_{i=1}^{N_1}\|x_i^1\|^2((\gamma_1)_i - (\gamma_1^1)_i) + \sum_{k\in D}|\hat{x}_k^1|^2((\hat{\gamma}_1^1)_k - (\hat{\gamma}_1)_k) \\
&= \sum_{i=1}^{N_1}\|x_i^1\|^2((\gamma_1)_i - (\gamma_1^1)_i) - \sum_{k\in D}\|\hat{x}_k^1\|^2((\hat{\gamma}_1)_k - (\hat{\gamma}_1^1)_k) \qquad (56)
\end{aligned}
$$

where the inequality holds sine γ^1 is optimal for $OPT_\lambda(\mu^0,\gamma_1^1)$.

E.3 Verification of the inequality

It remains to show (56) is less than 0. By Jensen's inequality, for each $k \in D$, we obtain

$$\|\hat{x}_k^1\|^2 \le \sum_{i=1}^{N_1}\frac{\gamma_{k,i}^1}{\hat{p}_k^1}\|x_i^1\|^2.$$

Combined with the fact $\hat{\gamma}_1 \le \hat{\gamma}_1^1$, we obtain:

$$
\begin{aligned}
(56) &\le \sum_{i=1}^{N_1}\|x_i^1\|^2((\gamma_1)_i - (\gamma_1^1)_i) - \sum_{k\in D}\sum_{i=1}^{N_1}\frac{\gamma_{k,i}^1}{\hat{p}_k^1}\|x_i^1\|^2((\hat{\gamma}_1)_k - (\hat{\gamma}_1^1)_k) \\
&= \sum_{i=1}^{N_1}\|x_i^1\|^2\left((\gamma_1)_i - (\gamma_1^1)_i - \sum_{k\in D}\frac{\gamma_{k,i}^1((\hat{\gamma}_1)_k - \hat{p}_k^1)}{\hat{p}_k^1}\right) \qquad (57)
\end{aligned}
$$

Pick $i \in [1:N_0]$, we have:

$$
\begin{aligned}
&(\gamma_1)_i - (\gamma_1^1)_i - \sum_{k\in D}\frac{\gamma_{k,i}^1((\hat{\gamma}_1)_k - \hat{p}_k^1)}{\hat{p}_k^1} \\
&= (\gamma_1)_i - \hat{p}_i^1 - \sum_{k\in D}\frac{\gamma_{k,i}^1((\hat{\gamma}_1)_k)}{\hat{p}_k^1} + \hat{p}_i^1 \\
&= \sum_{k'=1}^{N}\gamma_{k',i} - \sum_{k\in D}\frac{\gamma_{k,i}^1((\hat{\gamma}_1)_k)}{\hat{p}_k^1} \\
&= \sum_{k'=1}^{N_0}\sum_{k\in D}\frac{\hat{\gamma}_{k',k}\gamma_{k,i}^1}{\hat{p}_k^1} - \sum_{k\in D}\sum_{k'=1}^{N_0}\frac{\gamma_{k,i}^1\hat{\gamma}_{k',k}}{\hat{p}_k^1} \qquad (58) \\
&= 0
\end{aligned}
$$

where (58) holds by the fact: for each $k' \in [1:N_0], i \in [1:N_1]$, we have

$$\begin{aligned}\gamma_{k',i} &= (\hat{\gamma}(\hat{\gamma}^1)^{-1}[k',:])\gamma^1[:,i] \\ &= \left[\hat{\gamma}_{k',1}\frac{1}{\hat{p}^1_1}, \dots \hat{\gamma}_{k',N^0}\frac{1}{\hat{p}^j_{N_0}}\right]^T [\gamma^1_{1,i}, \dots \gamma^1_{N_0,i}] \\ &= \sum_{k\in D}\frac{\hat{\gamma}_{k',k}\gamma^1_{k,i}}{\hat{p}^1_k}\end{aligned}$$

Thus (57) is 0. Therefore, (56) is upper bounded by 0, and we complete the proof.

F Proof of Theorem 3.6

Proof. Without loss of generality, we assume that $\gamma^j \in \Gamma^*_{\leq}(\mu^0, \mu^j)$ is induced by a 1-1 map. We can suppose this since the trick is that we will end up with an N_0-point representation of μ^j when we get $\hat{\mu}^j$. For example, if two x^0_n and x^0_m are mapped to the same point $T(x^0_n) = T(x^0_m)$, when performing the barycentric, we can split this into two different labels with corresponding labels. (The moral is that the labels are important, rather than where the mass is located.) Therefore, $\gamma^j \in \mathbb{R}^{N_0 \times N_j}$ is such that in each row and column there exists at most one positive entry, and all others are zero. Let $\hat{\gamma} := \mathrm{diag}(\hat{p}^j_1, \dots, \hat{p}^j_{N_0})$. Then, by the definition of $\hat{\mu}^j$, we have $C(\gamma^j; \mu^0, \mu^j, \lambda) = C(\hat{\gamma}; \mu^0, \hat{\mu}^j, \lambda)$. Thus,

$$\begin{aligned}OPT_\lambda(\mu^0, \mu^j) &= C(\gamma^j; \mu^0, \mu^j, \lambda) \\ &= C(\gamma^j; \mu^0, \gamma^j_1, \lambda) + \lambda(|\mu^j| - |\gamma^j_1|) \\ &= C(\hat{\gamma}; \mu^0, \hat{\mu}^j, \lambda) + \lambda(|\mu^j| - |\hat{\mu}^j|) \\ &= OPT_\lambda(\mu^0, \hat{\mu}^j) + \lambda(|\mu^j| - |\hat{\mu}^j|),\end{aligned}$$

where the last equality holds from Theorem 3.5. This concludes the proof.

Moreover, from the above identities (for discrete measure) we can express

$$OPT_\lambda(\mu^0, \mu^j) = \sum_{k=1}^{N_0} \hat{p}_k \|x^0_k - \hat{x}^j_k\|^2 + \lambda|p^0 - \hat{p}^j_k| + \lambda(|p^j| - |\hat{p}^j|).$$

□

G Proof of Proposition 3.7

Proof. The result follows from (40) and (35). Indeed, we have

$$\begin{aligned}LOPT_{\lambda,\mu_0}(\hat{\mu}^j, \hat{\mu}^j) &= \|u^i - u^j\|_{\hat{p}^i \wedge \hat{p}^j, 2\lambda} + \lambda(|\hat{p}^i - \hat{p}^j|) \qquad \text{and} \\ LOPT_{\lambda,\mu_0}(\mu^j, \mu^j) &= \|u^i - u^j\|_{\hat{p}^i \wedge \hat{p}^j, 2\lambda} + \lambda(|\hat{p}^i - \hat{p}^j|) + \lambda(|\nu^i| + |\nu^j|),\end{aligned}$$

since the LOPT barycentric projection of μ^i is basically μ^i without the mass that needs to be created from the reference (analogously for μ^j). □

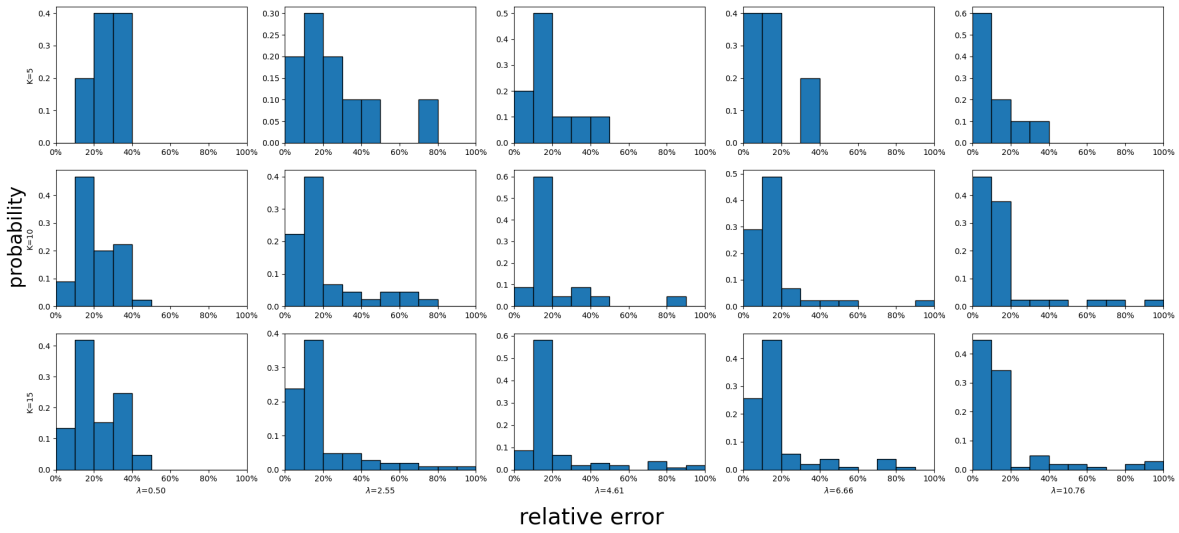


Figure 6: Histogram of relatives errors (depicted in Figure 2) between OPT_λ and $LOPT_{\lambda,\mu_0}$. (Number of samples $N = 500$. Number of repetitions $= 10$. Dimension $= 2$. –measures on $\mathbb{R}^2$–.)

H Applications

For completeness, we will expand on the experiments and discussion presented in Section 4, as well as on Figure 1 which contrasts the HK technique with OPT from the point of view of interpolation of measures.

First, we recall that similar to LOT [41] and [34], the goal of LOPT is not exclusively to approximate OPT, but to propose new transport-based metrics (or discrepancy measures) that are computationally less expensive and easier to work with than OT or OPT, specifically when many measures must be compared.

Also, one of the main advantages of the *linearization* step is that it allows us to embed sets of probability (resp. positive finite) measures into a linear space ($\mathcal{T}_{\mu_0}$ space). Moreover, it does it in a way that allows us to use the $\mathcal{T}_{\mu_0}$-metric in that space as a proxy (or replacement) for more complicated transport metrics while preserving the natural properties of transport theory. As a consequence, data analysis can be performed using Euclidean metric in a simple vector space.

H.1 Approximation of OPT Distance

For a better understanding of the errors plotted in Figure 2, the following Figure 6 shows the histograms of the relative errors for different values of λ and each number of measures $K = 5, 10, 15$.

As said in Section 4, we recall that for these experiments, we created K point sets of size N for K different Gaussian distributions in $\mathbb{R}^2$. In particular, $\mu^i \sim \mathcal{N}(m^i, I)$, where m^i is randomly selected such that $\|m^i\| = \sqrt{3}$ for $i = 1, ..., K$. For the reference,

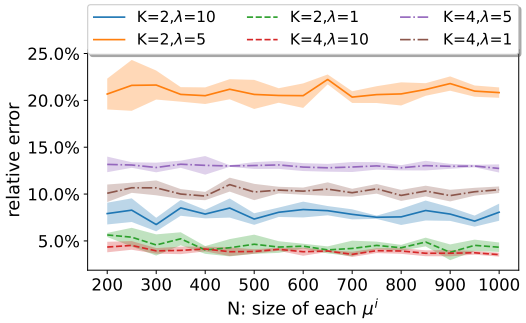


Figure 7: The average relative error between OPT_λ and $LOPT_{\lambda,\mu_0}$ between all pairs of discrete measures μ^i and μ^j.

we picked an N point representation of $\mu^0 \sim \mathcal{N}(\overline{m}, I)$ with $\overline{m} = \sum m^i/K$. For figures 2 and 6, the sample size N was set equal to 500.

In what follows, we include tests for $N = 200, 250, 300, 350, ..., 900, 950, 1000$ and $K = 2, 4$. For each (N, K), we repeated each experiment 10 times. The relative errors are shown in Figure 7. For large λ, most mass is transported and $OT(\mu^i, \mu^j) \approx OPT_\lambda(\mu^i, \mu^j)$, the performance of LOPT is close to that of LOT, and the relative error is small. For small λ, almost no mass is transported, $OPT_\lambda(\mu^i, \mu^j) \approx \lambda(|\mu^i| + |\mu^j|) \approx \lambda(|\mu^i| - |\hat{\mu}^i| + |\mu^j| - |\hat{\mu}^j|) \approx LOPT_{\mu^0,\lambda}(\mu^i, \mu^j)$, and we still have a small error. In between, e.g., $\lambda = 5$, we have the largest relative error. Similar results were obtained by setting the reference as the OT barycenter.

H.2 PCA analysis

For problems where doing pair-wise comparisons between K distributions is needed, in the classical optimal (partial) transport setting we have to solve $\binom{K}{2}$ OT (resp. OPT) problems. In the LOT (resp. LOPT) framework, however, one only needs to perform K OT (resp. OPT) problems (matching each distribution with a reference measure).

One very ubiquitous example is to do clustering or classification of measures. For this case, the number of target measures K representing different samples is usually very large. For the particular case of data analysis on Point Cloud MNIST, after using PCA in the embedding space (see Figure 5), it can be observed that the LOT framework is a natural option for separating different digits, but the equal mass requirement is too restrictive in the presence of noise. LOPT performs better since it does not have this restriction. This is one example where the number of measures $K = 900$ is much larger than 2.

H.3 Point Cloud Interpolation

Here, we add the results of the experiments mentioned in Section 4 which complete Figure 4. In the new Figure 8, in fact, Figure 4 corresponds to the subfigure 8b. We

conducted experiments using three digits (0, 1, and 9) from the PointCloud MNIST dataset, with 300 point sets per digit. We utilized LOPT embedding to calculate and visualize the interpolation between pairs of digits. We chose the barycenter between 0, 1, and 9 as the reference for our experiment. However, to avoid redundant results, in the main paper, we only demonstrated the interpolation between 0 and 9 in our main paper. The remaining plots for the other digit pairs using different levels of noise η are included here for completeness. Later, in Section H.4, Figure 11 will add the plots for the HK technique and its linearized version.

H.4 Preliminary comparisons between LHK and LOPT

The contrasts between the Linearized Hellinger-Kantorovich (LHK) [10] and the LOPT approaches come from the differences between HK (Hellinger-Kantorovich) and OPT distances.

The main issue is that for HK transportation between two measures, the transported portion of the mass does not resemble the OT-geodesic where mass is preserved. In other words, HK changes the mass as it transports it, while OPT preserves it.

This issue is depicted in Figure 1. In that figure, both the initial (blue) and final (purple) distributions of masses are two unit-mass delta measures at different locations. Mass decreases and then increases while being transported for HK, while it remains constant for OPT. For HK, the transported portion of the mass does not resemble the OT-geodesic where mass is preserved. In other words, HK changes the mass as it transports it, while OPT preserves it.

To illustrate better this point, we incorporate here Figure 9 which is the two-dimensional analog of Figure 1.

In addition, in Figure 10 we not only compare HK and OPT, but also LHK and LOPT. The top subfigure 10a shows the measures μ_1 (blue dots) and μ_2 (orange crosses) to be interpolated with the different techniques in the next subfigures 10b and 10c. Also, in 10a we plot the reference μ_0 (green triangles) which is going to be used only in experiment 10c. The measures μ_1 and μ_2 are interpreted as follows. The three masses on the interior, forming a triangle, can be considered as the signal information and the two masses on the corners can be considered noise. That is, we can assume they are just two noisy point cloud representations of the same distribution shifted. We want to transport the three masses in the middle without affecting their mass and relative positions too much. Subfigure 10b shows HK and OPT interpolation between the two measures μ_1 and μ_2. In the HK cases, we notice not only how the masses vary, but also how their relative positions change obtaining a very different configuration at the end of the interpolation. OPT instead returns a more natural interpolation because of the mass preservation of the transported portion and the decoupling between transport and destruction/creation of mass. Finally, subfigure 10c shows the interpolation of μ_1 and μ_2 for LHK and LOPT. The reference measure μ_0 is the same for both and we can see the *denoising* effect due to the fact that, for the three measures, the mass is concentrated in the triangular region in the center. However, while mass and structure-preserving

(a) $\eta = 0$ (b) $\eta = 0.5$

(c) $\eta = 0.75$ (d) $\eta = 0$

(e) $\eta = 0.5$ (f) $\eta = 0.75$

Figure 8: Interpolation between two point-clouds at different times $t \in \{0, 0.25, 0.5, 0.75, 1\}$. Different values of noise η were considered for the different interpolation approaches (OT, LOP, OPT, LOPT). For LOT and LOPT the reference measure is the barycenter between the PointClouds 0, 1, and 9 with no noise. Top row: Interpolation between digits 0 and 9. Bottom row: Interpolation between digits 0 and 1.

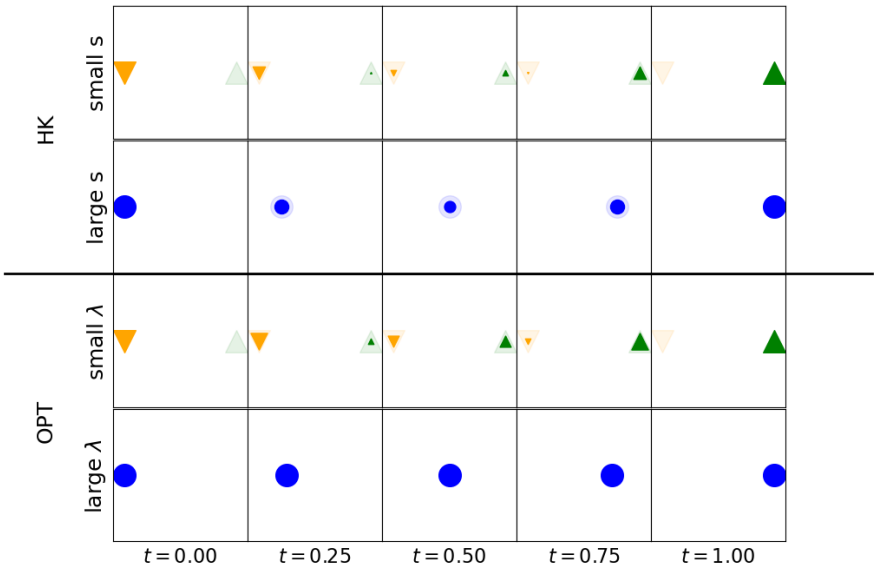


Figure 9: HK vs. OPT interpolation between two delta measures of unit mass located at different positions. Two scenarios are shown for each transport framework varying the respective parameters (s for HK, and λ for OPT). Blue circles stand for the cases when the geodesic transports mass. Triangular shapes represent destruction and creation of mass. Triangles pointing down in orange indicate the locations where mass will be destroyed. Triangles pointing up in green indicate the locations where mass will be created. On top of that shadows are added to emphasize the change of mass from initial and final configurations. The Top two plots exhibit two extreme cases when performing HK geodesic. When s is small, everything is created/destroyed, when s is large everything is transported *without mass-preservation*. On the other side, the Bottom two plots show the two analogous cases for OPT geodesics. When λ is small we observe creation/destruction, when λ is large we have *mass-preserving transportation*. The intermediate cases are treated in the following.

transportation can be seen for LOPT, for LHK the shape of the configuration changes.

On top of that, as is often the case on quantization, point cloud representations of measures are given as samples with uniform mass. OPT/LOPT interpolation will not change the mass of each transported point. Therefore, the intermediate steps of an algorithm using OPT/LOPT transport benefit from conserving the same type of representation. That is, as a uniform set of points.

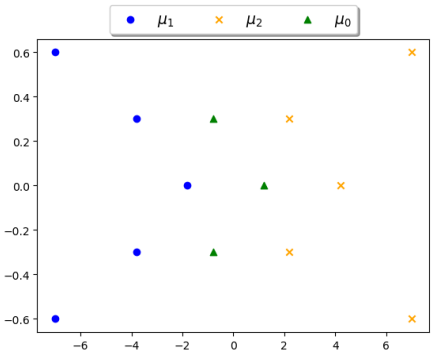
For comparison on PointCloud interpolation using MNIST, we include Figure 11 that illustrates OPT, LOPT, HK, and LHK interpolation for digit pairs (0,1) and (0,9). In the visualizations, the size of each circle is plotted according to amount of the mass at each location.

However, we do not claim the presented LOPT tool to be a one size fits all kind of tool. We are working on the subtle differences between LHK and LOPT and expect to have a complete and clear picture in the future. The aim of this article was to present a new tool with an intuitive introduction and motivation so that the OT community would benefit from it.

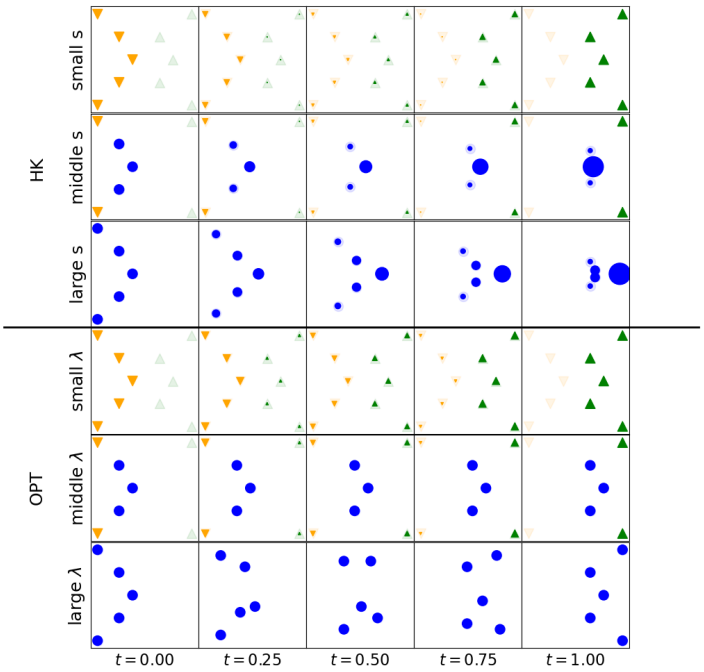
H.5 Barycenter computation

Can the linear embedding technique be used to compute the barycenter of a set of measures, e.g., by computing the barycenter of the embeddings and performing an inversion to the original space?

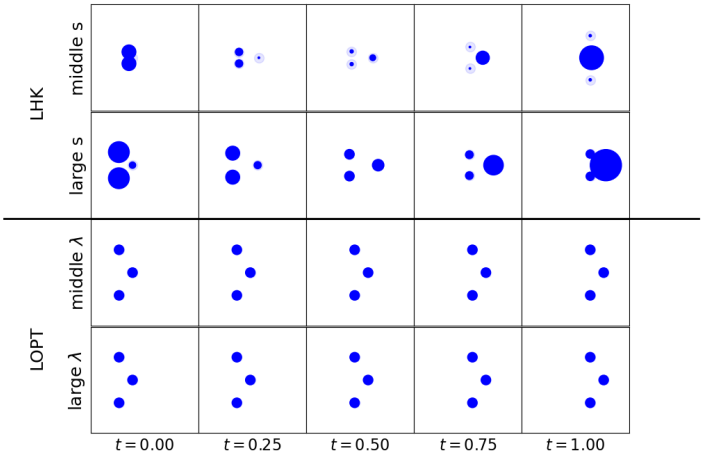
One can first calculate the mean of the embedded measures, and recover a measure from this mean. The recovered measure is not necessarily the OPT barycenter, however, one can repeat this process and obtain better approximations of the barycenter. Similar to LOT, we have numerically observed that such a process will converge to a measure that is close to the barycenter. However, there are several technical considerations that one needs to pay close attention. For instance, the choice of λ and the choice of the initial reference measure are critical in this process.



(a) Plot of two measures (μ_1 as blue circles and μ_2 as orange crosses, where each point has mass one), and a reference measure μ_0 (concentrated on the three green triangular locations, unit mass each).

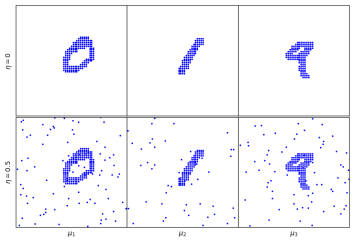


(b) HK and OPT interpolation between the two measures μ_1 and μ_2 at different times. The color code is the same as for Figure 9.



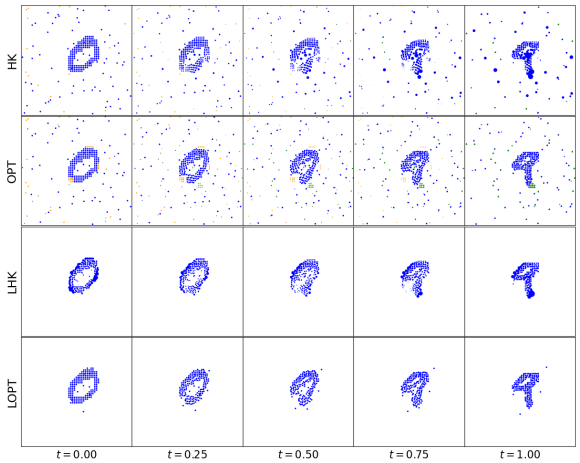
(c) LHK and LOPT interpolation between the two measures μ_1 and μ_2 at different times using for both cases the same reference μ_0.

Figure 10: HK vs OPT and LHK vs LOPT

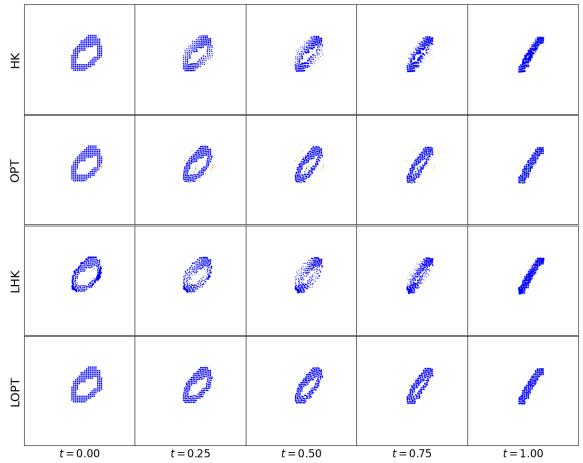


(a) Samples of digits 0, 1, and 9 from MNIST Data Set. Top row: clean point cloud. Bottom row: 50% of noise.

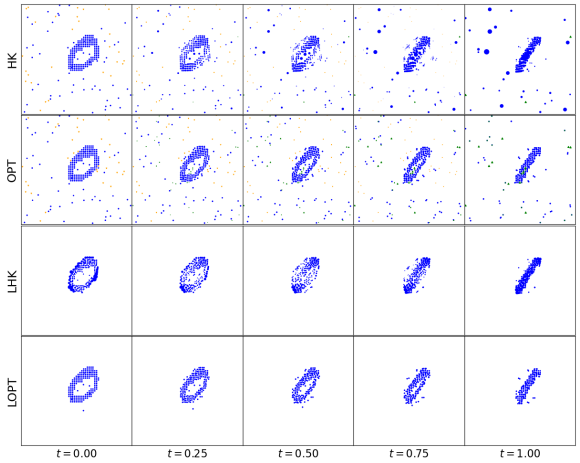
(b) Noise level $\eta = 0$.



(c) Noise level $\eta = 0.5$



(d) Noise level $\eta = 0$.



(e) Noise level $\eta = 0.5$.

Figure 11: Point cloud interpolation using all the techniques unbalanced transportation HK, OPT, LHK, and LOPT.

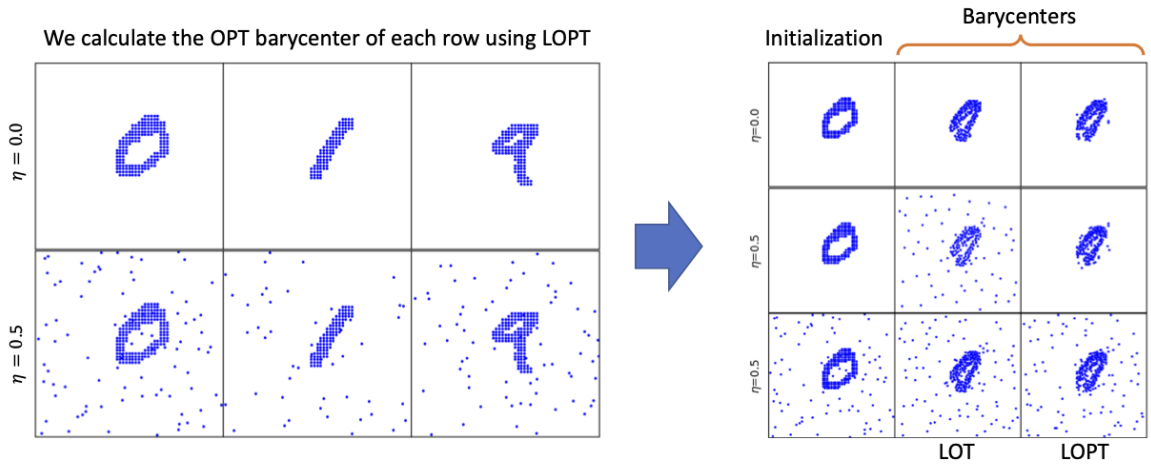


Figure 12: The depiction of barycenters between digits 0 and 1, and between 0 and 9 using the LOPT technique. Left panel: original measures (point-clouds) μ_1 (digit 0), μ_2 (digit 1), and μ_3 (digit 9). We considered them with no noise on the top left panel ($\eta = 0$), and with corrupted under noise on the bottom left panel (level of noise $\eta = 0.5$). Right panel: The first row is the result that both initial measure and data μ_1, μ_2, μ_3 are clean data; the second row is the result of clean initial measure and noise corrupted data; the third row is the result for noise corrupted initial measure and data.