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 (Arjovsky et al., 2017; Genevay et al., 2017; Liu et al., 2019), domain adaptation (Courty et al., 2014; 2017), bayesian inference (Kim et al., 2013), regression (Janati et al., 2019), clustering (Ye et al., 2017), learning from graphs (Kolouri et al., 2020) and point sets (Naderializadeh et al., 2021; Nguyen et al., 2023), to name a few. These metrics define a powerful geometry for comparing probability measures with numerous desirable properties, for instance, parameterized geodesics (Ambrosio et al., 2005), barycenters (Cuturi & Doucet, 2014), and a weak Riemannian structure (Villani, 2003).

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 (Cuturi, 2013; Chizat et al., 2020; Scetbon & Marco Cuturi, 2022). 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 (Chizat et al., 2015; 2018b; Liero et al., 2018) and optimal partial transport (Caffarelli & McCann, 2010; Figalli, 2010; Figalli & Gigli, 2010). Such unbalanced/partial optimal transport formulations have been recently used to improve minibatch optimal transport (Nguyen et al., 2022) and for point-cloud registration (Bai et al., 2022).

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, Wang et al. (2013) 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 (Moosmüller & Cloninger, 2023; Aldroubi et al., 2021), 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 (Cai et al., 2020) and comparing medical images (Basu et al., 2014; Kundu et al., 2018) to permutation invariant pooling for comparing graphs (Kolouri et al., 2020) and point sets (Naderializadeh et al., 2021).
Many applications in ML involve comparing non-negative measures (often empirical measures) with varying total amounts of mass, e.g., domain adaptation (Fatras et al., 2021). 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 (Caffarelli & McCann, 2010; Figalli, 2010; Figalli & Gigli, 2010), Kantorovich–Rubinstein norm (Guittet, 2002; Lellmann et al., 2014), and the Hellinger–Kantorovich distance (Chizat et al., 2018a; Liero et al., 2018). These methods fall under the broad category of “unbalanced optimal transport” (Chizat et al., 2018b; Liero et al., 2018). 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, Cai et al. (2022) 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 HK 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. Outer masses (located at -1.2 for initial time $t=0$ 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.

\textbf{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.

\textbf{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 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

\begin{figure}[h]
    \centering
    \includegraphics[width=\textwidth]{hk_opt.png}
    \caption{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.}
\end{figure}

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 $\Omega$ of $\mathbb{R}^d$, and consider $\mu^0, \mu^1 \in \mathcal{P}(\Omega)$. The Optimal Transport (OT) problem between $\mu^0$ and $\mu^1$ is that of finding the cheapest way to transport all the mass distributed according to the reference measure $\mu^0$ onto a new distribution of mass determined by the target measure $\mu^1$. Mathematically, it was stated by Kantorovich as the minimization problem

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

for $C(\gamma; \mu^0, \mu^1) := \int_{\Omega^2} \|x^0 - x^1\|^2 d\gamma(x^0, x^1)$,
where \( \Gamma(\mu^0, \mu^2) \) is the set of all joint probability measures in \( \Omega^2 \) with marginals \( \mu^0 \) and \( \mu^2 \). A measure \( \gamma \in \Gamma(\mu^0, \mu^2) \) is called a transportation plan, and given measurable 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 \( \|x^0 - x^2\|^2 \) is interpreted as the cost of transporting a unit mass located at \( x^0 \) to \( x^2 \). Therefore, \( C(\gamma; \mu^0, \mu^2) \) represents the total cost of moving \( \mu^0 \) to \( \mu^2 \) according to \( \gamma \). Finally, we will denote the set of all plans that achieve the infimum in (1), which is non-empty (Villani, 2003), as \( \Gamma^*(\mu^0, \mu^2) \).

Under certain conditions (e.g., when \( \mu^0 \) has continuous density), an optimal plan \( \gamma \) 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 \( \gamma \) 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 \( \gamma = (1\times T)_# \mu^0 \). Hence, (1) reads as

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

The function \( T : \Omega \to \Omega \) is called a Monge map, and when \( \mu^0 \) is absolutely continuous it is unique (Brenier, 1991).

Finally, the square root of the optimal value \( OT(\cdot, \cdot) \) is exactly the so-called Wasserstein distance, \( W_2 \), in \( P(\Omega) \) (Villani, 2003, Th.7.3), and we will call it also as OT squared distance. In addition, with this distance, \( P(\Omega) \) is not only a metric space but also a Riemannian manifold (Villani, 2003). In particular, the tangent space of any \( \mu \in P(\Omega) \) is \( \mathcal{T}_\mu = L^2(\Omega; \mathbb{R}^d, \mu) = \{ u : \Omega \to \mathbb{R}^d : \|u\|^2_\mu < \infty \} \), where

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

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 \( \rho \in P([0,1] \times \Omega) \) a curve of measures parametrized in time that describes the distribution of mass \( \rho_t := \rho(t, \cdot) \in P(\Omega) \) at each instant \( 0 \leq t \leq 1 \). We will require the curve to be sufficiently smooth, to have boundary conditions \( \rho_0 = \mu^0, \rho_1 = \mu^2 \), 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 \( \rho_t \) satisfies the continuity equation\(^1\) with boundary conditions

\[
\partial_t \rho + \nabla \cdot \rho v = 0, \quad \rho_0 = \mu^0, \quad \rho_1 = \mu^2. \tag{5}
\]

The length\(^3\) of the curve can be stated as

\[
\int_{[0,1]} \|v_t\|^2 d\rho_t := \int_{[0,1]} \|v_t\|^2 d\rho_t dt, \text{ for } \| \rho_t \|_2 \text{ as in (4), and } OT(\mu^0, \mu^2) \text{ coincides with the length of the shortest curve between } \mu^0 \text{ and } \mu^2 \text{ (Benamou \\& Brenier, 2000). Hence, the dynamical formulation of the OT problem (1) reads as}
\]

\[
OT(\mu^0, \mu^2) = \inf_{(\rho,v) \in CE(\mu^0, \mu^2)} \int_{[0,1]} \|v_t\|^2 d\rho_t, \tag{6}
\]

where \( CE(\mu^0, \mu^2) \) is the set of pairs \( (\rho, v) \), where \( \rho \in P([0,1] \times \Omega) \) 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

\[
v_t(x) = T(x) - x_0, \quad \text{for } x = T_t(x_0). \tag{8}
\]

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

\[
\rho_t = (T_t)_# \mu^0, \quad 0 \leq t \leq 1. \tag{9}
\]

The pair \( (\rho, v) \) defined by (9) and (8) satisfies the continuity equation (5) and solves (6). Moreover, the curve \( \rho_t \) is a constant speed geodesic in \( P(\Omega) \) between \( \mu^0 \) and \( \mu^2 \) (Figalli & Glaudo, 2021), i.e., it satisfies that for all \( 0 \leq s \leq t \leq 1 \)

\[
\sqrt{OT(\rho_s, p_t)} = (t-s) \sqrt{OT(\rho_0, \rho_1)}. \tag{10}
\]

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

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

which tells us that we only need the speed at the initial time to compute the total length of the curve. Moreover, \( OT(\mu^0, \mu^2) \) coincides with the squared norm of the tangent vector \( v_0 \) in the tangent space \( \mathcal{T}_{\mu^0} \) of \( P(\Omega) \) at \( \mu^0 \).

\(^1\)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 (Villani, 2003; Santambrogio, 2015).

\(^3\)Precisely, the length of the curve \( \rho \), with respect to the Wasserstein distance, should be \( \int_0^1 \|v_t\|_2 dt \), 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 (Santambrogio, 2015, Lemma 5.29).
2.3. Linear Optimal Transport Embedding

Inspired by the induced Riemannian geometry of the OT squared distance, Wang et al. (2013) proposed the so-called Linear Optimal Transportation (LOT) framework. Given two target measures \( \mu^0, \mu^1 \), the main idea relies on considering a reference measure \( \mu^0 \) and embed these target measures into the tangent space \( T_{\mu^0} \). This is done by identifying each measure \( \mu^j \) 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 \( \mu^0 \). Then, the LOT embedding (Moosmüller & Cloninger, 2023) is defined as

\[
\mu^j \mapsto u^j := T^j - \text{id} \quad \forall \mu^j \in \mathcal{P}(\Omega) \quad (12)
\]

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

\[
\|u^j\|^2_{\mu^0} = OT(\mu^0, \mu^j). \quad (13)
\]

After this embedding, one can use the distance in \( T_{\mu^0} \) between the projected measures to define a new distance in \( \mathcal{P}(\Omega) \) that can be used to approximate \( OT(\mu^1, \mu^2) \). The LOT squared distance is defined as

\[
LOT_{\mu^0}(\mu^1, \mu^2) := \|u^1 - u^2\|^2_{\mu^0}. \quad (14)
\]

2.4. LOT in the Discrete Setting

For discrete probability measures \( \mu^0, \mu^1 \) of the form

\[
\mu^0 = \sum_{n=1}^{N_0} p_n^0 \delta_{x_n^0}, \quad \mu^1 = \sum_{n=1}^{N_1} p_n^1 \delta_{x_n^1}, \quad (15)
\]

a Monge map \( T^j \) for \( OT(\mu^0, \mu^j) \) may not exist. Following Wang et al. (2013), in this setting, the target measure \( \mu^j \) can be replaced by a new measure \( \tilde{\mu}^j \) for which an optimal transport Monge map exists. For that, given an optimal plan \( \gamma_j \in \Gamma^e(\mu^0, \mu^j) \), it can be viewed as a \( N_0 \times N_1 \) matrix whose value at position \((n, m)\) represents how much mass from \( x_n^0 \) should be taken to \( x_m^1 \). Then, we define the OT barycentric projection\(^3\) of \( \mu^j \) with respect to \( \mu^0 \) as

\[
\tilde{\mu}^j := \sum_{n=1}^{N_0} p_n^0 \delta_{\tilde{x}_n^j}, \quad \text{where} \quad \tilde{x}_n^j := \frac{1}{p_n^0} \sum_{m=1}^{N_1} \gamma_{n,m} x_m^1. \quad (16)
\]

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

**Lemma 2.1.** Let \( \mu^0 \) and \( \mu^1 \) be two discrete probability measures as in (15), and consider an OT barycentric projection \( \tilde{\mu}^j \) of \( \mu^1 \) with respect to \( \mu^0 \) as in (16). Then, the map \( x_n^0 \mapsto \tilde{x}_n^j \) given by (16) solves the OT problem \( OT(\mu^0, \tilde{\mu}^j) \).

It is easy to show that if the optimal transport plan \( \gamma^j \) is induced by a Monge map, then \( \tilde{\mu}^j = \mu^j \). 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 \( \mu^0 \), we can define the LOT embedding for a discrete measure \( \mu^j \) as the rule

\[
\mu^j \mapsto u^j := [(\tilde{x}_1^j - x_1^0), \ldots, (\tilde{x}_{N_0}^j - x_{N_0}^0)]. \quad (17)
\]

The range \( T_{\mu^0} \) of this application is identified with \( \mathbb{R}^{d \times N_0} \) with the norm \( \|u^j\|_{\mu^0} := \sum_{n=1}^{N_0} \|u(n)\|^2_{\mu^0} \), where \( u(n) \in \mathbb{R}^d \) denotes the \( n \)-th 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 \( \gamma^j \) for problem \( OT(\mu^0, \mu^j) \) is induced by a Monge map, then the discrete embedding is consistent with (13) in the sense that

\[
\|u^j\|^2_{\mu^0} = OT(\mu^0, \mu^j) = OT(\mu^0, \mu^j). \quad (18)
\]

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 \( OT(\mu^1, \mu^2) \). The LOT discrepancy\(^4\) is defined as

\[
LOT_{\mu^0}(\mu^1, \mu^2) := \|u^1 - u^2\|^2_{\mu^0}. \quad (19)
\]

We call it a discrepancy because it is not a squared metric between discrete measures. It does not necessarily satisfy that \( LOT(\mu^1, \mu^2) \neq 0 \) for every distinct \( \mu^1, \mu^2 \). Nevertheless, \( \|u^1 - u^2\|_{\mu^0} \) is a metric in the embedding space.

2.5. OT and LOT Geodesics in Discrete Settings

Let \( \mu^j, \mu^j \) be discrete probability measures as in (15) (with ‘i’ in place of ‘0’). If an optimal Monge map \( T \) for \( OT(\mu^1, \mu^j) \) exists, a constant speed geodesic \( \rho_t \) between \( \mu^1 \) and \( \mu^j \), for the OT squared distance, can be found by mimicking (9). Explicitly, with \( T_t \) as in (7),

\[
\rho_t = (T_t) \# \mu^j = \sum_{n=1}^{N_1} p_n^j \delta_{(1-t)x_n^1 + T_t(x_n^1)}. \quad (20)
\]

In practice, one replaces \( \mu^j \) by its OT barycentric projection with respect to \( \mu^1 \) (and so, the existence of an optimal Monge map is guaranteed by Lemma 2.1).

\(^3\)We refer to (Ambrosio et al., 2005) for the rigorous definition.

\(^4\)In (Wang et al., 2013), the embedded vector is defined as the element-wise multiplication \( u^j \bigotimes \sqrt{p^j} \). In addition, \( \tilde{\mu}^j, u^j \) are determined by \( \gamma^j \), but we ignore the subscript ‘\( \gamma^j \)’ for convenience.

\(^5\)In (Wang et al., 2013), LOT is defined by the infimum over all possible optimal pairs \((\gamma^1, \gamma^j)\). We do not distinguish these two formulations for convenience in this paper. Additionally, (19) is determined by the choice of \((\gamma^1, \gamma^j)\).
Now, given a discrete reference $\mu^0$, the LOT discrepancy provides a new structure to the space of discrete probability densities. Therefore, we can provide a substitute for the OT geodesic (20) between $\mu^1$ and $\mu^2$. Assume we have the embeddings $\mu^i \mapsto w^i$, $\mu^j \mapsto w^j$ as in (17). The geodesic between $w^i$ and $w^j$ in the LOT embedding space $\mathbb{R}^{d \times N_0}$ has the simple form $w_j = (1 - t)w^i + tw^j$. This correlates with the curve $\hat{\rho}_t$ in $\mathcal{P}(\Omega)$ induced by the map $\hat{T}_t : \hat{x}_n^i \mapsto \hat{x}_n^j$ as

$$\hat{\rho}_t := (\hat{T}_t)_{\#} \hat{\mu}^t = \sum_{n=0}^{N_0} \hat{p}_n^0 \delta_{\hat{x}_n^0 + u_n(t)}.$$  

(21)

By abuse of notation, we call this curve the LOT geodesic between $\mu^i$ and $\mu^j$. 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 $0 \leq s \leq t \leq 1$

$$\sqrt{\text{LOT}_\mu(\hat{\rho}_s, \hat{\rho}_1)} = (t - s) \sqrt{\text{LOT}_\mu(\hat{\rho}_0, \hat{\rho}_1)}.$$  

(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 $\Omega$. For $\lambda \geq 0$ the OPT problem between $\mu^0, \mu^j \in \mathcal{M}_+(\Omega)$ can be formulated as

$$\text{OPT}_\lambda(\mu^0, \mu^j) := \inf_{\gamma \in \Gamma^\lambda(\mu^0, \mu^j)} C(\gamma; \mu^0, \mu^j, \lambda)$$

(23)

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

(24)

where $|\mu^0 - \gamma_0|$ is the total mass of $\mu^0 - \gamma_0$ (resp. $|\mu^j - \gamma_1|$), and $\Gamma^\lambda(\mu^0, \mu^j)$ denotes the set of all measures in $\Omega^2$ with marginals $\gamma_0$ and $\gamma_1$ satisfying $\gamma_0 \leq \mu^0$ (i.e., $\gamma_0(E) \leq \mu^0(E)$ for all measurable set $E$), and $\gamma_1 \leq \mu^j$. Here, the mass destruction and creation penalty is linear, parametrized by $\lambda$. The set of minimizers $\Gamma^\lambda(\mu^0, \mu^j)$ of (23) is non-empty (Figalli, 2010). One can further restrict $\Gamma^\lambda(\mu^0, \mu^j)$ to the set of partial transport plans $\gamma$ such that $\|x^0 - x^j\|^2 < 2\lambda$ for all $(x^0, x^j) \in \text{supp}(\gamma)$ (Bai et al., 2022, Lemma 3.2). This means that if the usual transportation cost is greater than $2\lambda$, it is better to create/destroy mass.

3.2. Dynamic Formulation of Optimal Partial Transport

Adding a forcing term $\zeta$ 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 of mass law. Thus, it is natural that the minimization problem (23) can be rewritten (Chizat et al., 2018b, Th. 5.2) into a dynamic formulation as

$$\text{OPT}_\lambda(\mu^0, \mu^j) = \inf_{(\rho, v, \zeta) \in \mathcal{FCE}(\mu^0, \mu^j)} \int_{[0, 1] \times \Omega} |v|^2 dp + \lambda |\zeta|$$

(25)

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

$$\partial_t \rho + \nabla \cdot \rho v = \zeta, \quad \rho_0 = \mu^0, \quad \rho_1 = \mu^j.$$  

(26)

As in the case of OT, under certain conditions on the minimizers $\gamma$ of (23), one curve $\rho_t$ that minimizes the dynamic formulation (25) is quite intuitive. We show in the next proposition that it consists of three parts $\gamma_t, (1 - t)v_0 + tv^j$ (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 $|v_0|, |v^j|$, respectively.

Proposition 3.1. Let $\gamma^* \in \gamma^t \mu^0, \mu^j$ be of the form $\gamma^* = (i\times T)_{\#} \gamma^0$ for $T : \Omega \rightarrow \Omega$ a (measurable) map. Let

$$\nu_0 := \rho^0 - \gamma_0^*, \quad \nu^j := \rho^j - \gamma_1^*,$$

(27)

$$T_t(x) := (1 - t)x + tT(x), \quad \gamma_t := (T_t)_{\#} \gamma_0^*.$$  

(28)

Then, an optimal solution $(\rho, v, \zeta)$ for (25) is given by

$$\rho_t := \gamma_t + (1 - t)v_0 + tv^j,$$

(29)

$$v_t(x) := T(x_0) - x_0, \quad \zeta_t := \nu^j - \nu_0.$$  

(30)

Moreover, plugging in $(\rho, v, \zeta)$ into (25), it holds that

$$\text{OPT}_\lambda(\mu^0, \mu^j) = \|v_0\|^2_{\gamma_0^0, 2\lambda} + \lambda (|v_0| + |v^j|),$$

(32)

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

$$|v|^2_{\mu^0, 2\lambda} := \int_{\Omega} \text{min}(\|v\|^2, 2\lambda) d\mu, \quad \text{for} \ v : \Omega \rightarrow \mathbb{R}^d.$$  

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

$$\mu^j \mapsto (w^j, \mu^j, v^j) := (v_0, \gamma_0, v_1)$$

(33)

where $v_0, \gamma_0, v^j$ 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 $\gamma_0$ is necessary since we need to specify what part of the reference is being transported. The third component $\nu^j$ 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 $\mu^0 \land \mu^1$ be the minimum measure between $\mu^0$ and $\mu^1$. By the above definition, $\mu^0 \mapsto (u^0, \bar{\mu}^0, \nu^0) = (0, \mu^0, 0)$. Therefore, (32) can be rewritten

$$OPT_{\gamma}(\mu^0, \mu^1) = \left\| u^0 - u^1 \right\|^2_{\bar{\mu}^0 \land \bar{\mu}^1, 2\lambda} + \lambda(\| \bar{\mu}^0 - \bar{\mu}^1 \| + |\nu_0 - \nu^1|)$$  (34)

This motivates the definition of the LOT discrepancy.  

**Definition 3.3.** Consider a reference $\mu^0 \in M_+(\Omega)$ and target measures $\mu^1, \mu^2 \in M_+(\Omega)$ such that $OPT_{\gamma}(\mu^0, \mu^1)$ and $OPT_{\gamma}(\mu^0, \mu^2)$ can be solved by plans induced by mappings as in the hypothesis of Proposition 3.1. Let $(u^1, \bar{\mu}^1, \nu^1)$ and $(u^2, \bar{\mu}^2, \nu^2)$ be the LOT embeddings of $\mu^1$ and $\mu^2$ with respect to $\mu^0$. The **LOT discrepancy** between $\mu^1$ and $\mu^2$ with respect to $\mu^0$ is defined as

$$LOPT_{\mu^0, \gamma}(\mu^1, \mu^2) := \left\| u^1 - u^2 \right\|^2_{\bar{\mu}^1 \land \bar{\mu}^2, 2\lambda} + \lambda(\| \bar{\mu}^1 - \bar{\mu}^2 \| + |\nu^1 - \nu^2|)$$  (35)

Similar to the LOT framework, by equation (34), LOT can recover OPT when $\mu^1 = \mu^0$. That is,

$$LOPT_{\mu^0, \gamma}(\mu^0, \mu^1) = OPT_{\gamma}(\mu^0, \mu^1).$$

### 3.4. LOT in the Discrete Setting

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

$$\min_{\gamma \in \Gamma_{\leq}(\mu^0, \mu^1)} \sum_{n,m} \left\| x_n - x_m \right\|^2_{\gamma}, \gamma n,m + \lambda(\| \gamma \| + \| \mu \| - 2|\gamma|)$$

where the set $\Gamma_{\leq}(\mu^0, \mu^1)$ can be viewed as the subset of $\mathcal{N}_0 \times \mathcal{N}_j$ matrices with non-negative entries

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

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

$^3$Formally, $\mu^0 \land \mu^1(B) := \inf \{ \mu^0(B_1) + \mu^1(B_2) \}$ for every Borel set $B$, where the infimum is taken over all partitions of $B_1$, 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.

$^9$Notice that in (16) we had $\bar{p}^0 = \sum_{m=1}^{N_0} \gamma_{0,m}$. This leads to introducing $\bar{p}_n^0$ as in (37). That is, $\bar{p}_n^0$ plays the role of $\bar{p}^0$ in the LOT framework. However, here $\bar{p}_n^0$ depends on $\gamma^j$ (on its first marginal $\gamma_{0,m}^j$) and not only on $p^0$, and so we add a superscript ‘j’.

The second consequence is that the LOT embedding of $\mu^j$ will always have a null third component. That is,

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

Therefore, we represent this embedding as $\hat{\mu}^j \mapsto (u^j, \bar{\mu}^j, \nu^j)$, for $u^j = [\hat{x}_1^j - x_1^j, \ldots, \hat{x}_{N_0}^j - x_{N_0}^j]$ and $\bar{\mu}^j = [p_1^j, \ldots, p_{N_0}^j]$. The last consequence is given in the next result.

**Theorem 3.5.** In the same setting of Definition 3.4, the map $x_n^j \mapsto \hat{x}_n^j$ given by (38) solves the problem $OPT_{\gamma}(\mu^0, \hat{\mu}^1)$, in the sense that induces the partial optimal plan $\hat{\gamma}^j = \text{diag}(\bar{p}_1^j, \ldots, \bar{p}_{N_0}^j)$.

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 $\mu^j$ that is not transported from the reference $\mu^0$. This has some minor consequences.

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

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

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

The second consequence is that the LOT embedding of $\mu^j$ will always have a null third component. That is,

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

Therefore, we represent this embedding as $\hat{\mu}^j \mapsto (u^j, \bar{\mu}^j, \nu^j)$, for $u^j = [\hat{x}_1^j - x_1^j, \ldots, \hat{x}_{N_0}^j - x_{N_0}^j]$ and $\bar{\mu}^j = [p_1^j, \ldots, p_{N_0}^j]$. The last consequence is given in the next result.
Proposition 3.7. If \( \mu^0, \mu^1, \mu^2 \) are discrete and satisfy the conditions of Definition 3.3, then

\[
LOPT_{\mu^0, \lambda}(\mu^1, \mu^2) = LOPT_{\mu^0, \lambda}(\tilde{\mu}^1, \tilde{\mu}^2) + \lambda C_{i,j}
\]

where \( C_{i,j} = |\mu^1| - |\tilde{\mu}^1| + |\mu^2| - |\tilde{\mu}^2| \).

As a byproduct we can define the LOPT discrepancy for any pair of discrete measures \( \mu^1, \mu^2 \) as the right-hand side of (41). In practice, unless to approximate \( OPT_{\lambda}(\mu^1, \mu^2) \), we set \( C_{i,j} = 0 \) in (41). That is,

\[
LOPT_{\mu^0, \lambda}(\mu^1, \mu^2) := LOPT_{\mu^0, \lambda}(\tilde{\mu}^1, \tilde{\mu}^2).
\]

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 \( \mu^1, \mu^2 \) for which exists \( \gamma \in \Gamma^1(\mu^1, \mu^2) \) of the form \( \gamma = (id \times T) \# \gamma_0 \), a natural candidate is the solution \( \rho_\gamma \) of the dynamic formulation of \( OPT_{\lambda}(\mu^1, \mu^2) \). The exact expression is given by Proposition 3.1. When working with general discrete measures \( \mu^1, \mu^2 \) (as in (15), with ‘i’ in place of 0) such \( \gamma \) is not guaranteed to exist. Therefore, to compute \( \rho_\gamma \) with its OPT barycentric projection with respect to \( \mu^0 \). And by Theorem 3.5 the map \( T : x_n^i \rightarrow \hat{x}_n^j \) solves \( OPT_{\lambda}(\mu^1, \mu^2) \) and the OPT interpolating curve is\(^{13}\)

\[
t \rightarrow \sum_{n=1}^{N} \hat{p}_n^1 \delta_{(1-t)x_n^i + tT(x_n)} + (1-t) \sum_{n=1}^{N} (\hat{p}_n^0 - \hat{p}_n^1) \delta_{x_n^i}.
\]

When working with a multitude of measures, it is convenient to consider a reference \( \mu^0 \) 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 \( \mu^0, \mu^1, \mu^2 \), with \( \mu^0 \) as the reference, let \( (u^i, \tilde{p}^i), (u^j, \tilde{p}^j) \) be the LOPT embeddings of \( \mu^1, \mu^2 \). Let \( \hat{p}^{ij} := \tilde{p}^i \wedge \tilde{p}^j \), and \( u_t := (1-t)u^i + tu^j \). We define the LOPT interpolating curve between \( \mu^0 \) and \( \mu^1 \) by

\[
t \rightarrow \sum_{k \in D_T} (\hat{p}_k^0 \delta_{x^i_k} + u_k) + (1-t) \sum_{k \in D_D} (\hat{p}_k^1 - \hat{p}_k^{ij}) \delta_{x^j_k} + (1-t) u_{ij} + t \sum_{k \in D_C} (\hat{p}_k^2 - \hat{p}_k^{ij}) \delta_{x^j_k} + u_{ij}
\]

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

\(^{13}\) \( \hat{p}_n^1 \) are the coefficients of \( \tilde{\mu}^1 \) with respect to \( \tilde{\mu}^1 \) analogous to (36).

4. Applications

Approximation of OPT Distance: Similar to LOT (Wang et al., 2013), and Linear Hellinger Kantorovich (LHK) (Cai et al., 2022), we test the approximation performance of OPT using LOPT. Given \( K \) empirical measures \( \{ \mu^i \}_{i=1}^K \), for each pair \( (\mu^i, \mu^j) \), 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 \( (\mu^i, \mu^j) \) 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 \( \mu^0 \) is critical for the accurate approximation of OPT. If \( \mu^0 \) is far away from \( \{ \mu^i \}_{i=1}^K \), the linearization is a poor approximation because the mass in \( \mu^i \) and \( \mu^0 \) would only be destroyed or created. In practice, one candidate for \( \mu^0 \) is the barycenter of the set of measures \( \{ \mu^i \} \). The OPT can be converted into OT problem (Caffarelli & McCann, 2010), and one can use OT barycenter (Cuturi & Doucet, 2014) to find \( \mu^0 \).

For our experiments, we created \( K \) point sets of size \( N = 500 \) for \( K \) different Gaussian distributions in \( \mathbb{R}^2 \). In particular, \( \mu^i \sim N(m^i, I) \), where \( m^i \) is randomly selected such that \( \|m^i\| = \sqrt{3} \) for \( i = 1, \ldots, K \). For the reference, we picked an \( N \) point representation of \( \mu^0 \sim N(m^0, I) \) with \( m^0 = \sum m^j / K \). We repeated each experiment 10 times. To exhibit the effect of the parameter \( \lambda \) in the approximation, the relative errors are shown in Figure 2. For the histogram of the relative errors for each value of \( \lambda \) and each number of measures \( K \), we refer to Figure 6 in the Appendix H. For large \( \lambda \), 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 (Karmarkar, 1984) to solve each OPT problem with a cost of \( O(N^3 \log(N)) \) each. Thus, computing the OPT distance pair-wisely for \( \{ \mu^i \}_{i=1}^K \) requires \( O(K^2 N^3 \log(N)) \). 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 the embeddings is linear. Thus, the total
computational cost is $O(KN^3 \log(N) + K^2 N)$. 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, HK geodesic, LHK 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 \( \{x_j^n, p_j^n\}_{n=1}^{N_j} \), \( j = 1, 2 \), that we consider as a discrete measure of the form \( \mu_j = \sum_{n=1}^{N_j} p_j^n \delta_{x_j^n} + 1/N_j \sum_{m=1}^{N_m} \delta_{y_m^j} \), 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 \( \eta \). Each \( x_j^n \) is a point in the square region \([0, 20]^2\) and each \( y_m^j \) is a point in the square region \([-10, 30]^2\). For HK, OPT, LOT and LHK, we use the distributions \( \mu_j \) 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 LOT and OPT do not eliminate noise points. HK, OPT still retains much of the noise because interpolation is essentially between \( \mu^1 \) and \( \hat{\mu}^2 \) (with respect to \( \mu^1 \)). So \( \mu^1 \) 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 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 weighted point set \( \{x_n^j, p_n^j\}_{n=1}^{N_j} \) that we consider as a discrete measure with added noise as in the previous experiment. The reference, \( \mu^0 \), is set to the OT barycenter of 30 samples from the clean data. For LOT we re-normalize each \( \mu^j \) 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 \( S_\eta \) 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.
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.

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 computational 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


Appendix

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.
- \(\text{diag}(p_1, \ldots, p_N)\), diagonal matrix.
- \(w^T\), transpose of a matrix \(w\).
- \(\mathbb{R}_+\), non-negative real numbers.
- \(\Omega\), convex and compact subset of \(\mathbb{R}^d\). \(\Omega^2 = \Omega \times \Omega\).
- \(\mathcal{P}(\Omega)\), set of probability Borel measures defined in \(\Omega\).
- \(\mathcal{M}_+(\Omega)\), set of positive finite Borel measures defined in \(\Omega\).
- \(\mathcal{M}\), 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.
- \(\Gamma_{\#}\mu\), push-forward of the measure \(\mu\) by the function. \(\Gamma_{\#}\mu(B) = \mu(F^{-1}(B))\) for all measurable set \(B\), where \(F^{-1}(B) = \{ x : F(x) \in B \}\).
- \(\delta_x\), 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 \(\mu\).
- \(\text{supp}(\mu)\), support of the measure \(\mu\).
- \(\mu^i \land \mu^j\) minimum measure between \(\mu^i\) and \(\mu^j\); \(p^i \land p^j\) vector having at each \(n\) entry the minimum value \(p_n^i\) and \(p_n^j\).
- \(\mu^0\) reference measure. \(\mu^i, \mu^j\) target measures. In the OT framework, they are in \(\mathcal{P}(\Omega)\). In the OPT framework, they are in \(\mathcal{M}_+(\Omega)\).
- \(\Gamma(\mu^0, \mu^j)\) set of Kantorovich transport plans.
- \(\mathcal{C}(\gamma; \mu^0, \mu^j) = \int_{\Omega^2} \| x^0 - x^j \|^2 d\gamma(x^0, x^j)\), Kantorovich cost given by the transportation plan \(\gamma\) between \(\mu^0\) and \(\mu^j\).
- \(\Gamma^+(\mu^0, \mu^j)\), set of optimal Kantorovich transport plans.
- \(T\), optimal transport Monge map.
- \(\text{id}\), identity map \(\text{id}(x) = x\).
- \(T_t(x) = (1 - t)x + tT(x)\) 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 \(0 \leq t \leq 1\), \(\rho_t \in \mathcal{P}(\Omega)\). 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.
- \(\nabla : V\), 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 \( \mu^0 \) and \( \mu^j \).

• \( \partial_t \rho + \nabla \cdot \rho v = \zeta \), continuity equation with forcing term \( \zeta \).

• \( \mathcal{FCE}(\mu^0, \mu^j) \), set solutions \((\rho, v, \zeta)\) of the continuity equation with forcing term with boundary conditions \( \mu^0 \) and \( \mu^j \).

• \( \text{OPT}(\mu^0, \mu^j) \), optimal transport minimization problem. See (1) for the Kantorovich formulation, and (6) for the dynamic formulation.

• \( W_2(\mu^0, \mu^j) = \sqrt{\text{OT}(\mu^0, \mu^j)} \), 2-Wasserstein distance.

• \( T_\mu = L^2(\Omega; \mathbb{R}^d, \mu) \), where \( \| u \|_\mu^2 = \int_\Omega \| u(x) \|^2 d\mu(x) \) (if \( \mu \) is discrete, \( T_\mu \) is identified with \( \mathbb{R}^{d \times N} \) and \( \| u \|_\mu = \sum_{n=1}^N \| u(n) \|^2 p_n \)).

• \( \text{LOT}_{\mu^0}(\mu^j, \mu^j) \), 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 \( \mu \) is dominated by \( \nu \).

• \( \Gamma_\leq(\mu^0, \mu^j) = \{ \gamma \in \mathcal{M}_+(\Omega^2) : \pi_0 \# \gamma \leq \mu^0, \pi_1 \# \gamma \leq \mu^j \} \), 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 \( \mu^0 \)), \( \nu^j = \mu^j - \gamma_1 \) (for the target \( \mu^j \)).

• \( C(\gamma; \mu^0, \mu^j, \lambda) = \int \| x^0 - x^j \|^2 d\gamma(x^0, x^j) + \lambda(|\mu^0 - \gamma_0| + |\mu^j - \gamma_1|) \), partial transport cost given by the partial transportation plan \( \gamma \) between \( \mu^0 \) and \( \mu^j \) with penalization \( \lambda \).

• \( \Gamma_\leq^\ast(\mu^0, \mu^j) \), set of optimal partial transport plans.

• \( \text{OPT}(\mu^0, \mu^j) \), partial optimal transport minimization problem between \( \mu^0 \) and \( \mu^j \). See (23) for the static formulation, and (25) for the dynamic formulation.

• \( \| v \|^2_{\mu, 2\lambda} := \int_\Omega \min(\| v \|^2, 2\lambda) d\mu \)

• \( \text{LOPT}_{\lambda}(\mu^j, \mu^j) \), 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 \( \mu^0 \) has continuous density, \( u^j := v^j = T^j - \text{id} \), and so it is a map from measures \( \mu^j \in \mathcal{P}(\Omega) \) to vector fields \( u^j \) defined in \( \Omega \), see (12). For discrete measures (\( \mu^0, \mu^j \) 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).

• \( \mu \), OT and OPT barycentric projection of \( \mu \) with respect to a reference \( \mu^0 \) (in \( \mathcal{P} \) or \( \mathcal{M}_+ \), resp.), see (16) and (36), respectively. \( \tilde{x}_n \) denote the new locations where \( \mu \) is concentrated. \( p_n^0 \) and \( \tilde{p}_n \) denote the wights of \( \mu \) for OT and OPT, respectively.

• \( u_t = (1 - t)u^i + tu^j \), geodesic in LOT embedding space.

• \( \hat{p}_t \), LOT geodesic.

• \( \text{LOPT} \) embedding, see (33), and (40).

• \( \text{OPT} \) interpolating curve, and \( \text{LOPT} \) 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 \( x_0 n \mapsto \hat{x}_j m \) according to (16). This map induces a transportation plan \( \hat{\gamma}^* := \text{diag}(p_1^0, \ldots, p_{N_0}^0) \in \mathbb{R}^{N_0 \times N_0}_+ \). We will prove that \( \hat{\gamma}^* \in \Gamma^*(\mu^0, \hat{\mu}^j) \). By definition, it is easy to see that its marginals are \( \mu^0 \) 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 \( \mu^0 \) and \( \mu^j \) (i.e. \( \gamma 1_{N_0} = \gamma^* 1_{N_0} = p^0 \)). We need to show

\[
C(\hat{\gamma}^*; \mu^0, \mu^j) \leq C(\hat{\gamma}; \mu^0, \hat{\mu}^j).
\]

Since \( \hat{\gamma}^* \) is a diagonal matrix with positive diagonal, its inverse matrix is \( (\gamma^*)^{-1} = \text{diag}(1/p_1^0, \ldots, 1/p_{N_0}^0) \). We define

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

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

\[
\gamma 1_{N_j} = \hat{\gamma}(\gamma^*)^{-1} \gamma^* 1_{N_j} = \hat{\gamma}(\gamma^*)^{-1} p^0 = \hat{\gamma} 1_{N_0} = p^0 \quad (44)
\]

\[
\gamma^T 1_{N_0} = (\gamma^*)^T (\gamma^*)^{-1} \gamma^T 1_{N_0} = (\gamma^*)^T (\gamma^*)^{-1} p^0 = (\gamma^*)^T 1_{N_0} = p^j, \quad (45)
\]

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 \( \gamma^* \) 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 \( \gamma^* \) and \( \gamma \), respectively. We write

\[
C(\gamma^*; \mu^0, \mu^j) = \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} \| x_0 n - x_m^j \|^2 \gamma^* n,m
\]

\[
= \sum_{n=1}^{N_0} \| x_0 n \|^2 p_n^0 + \sum_{m=1}^{N_j} \| x_m^j \|^2 p_n^0 - 2 \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} x_0 n \cdot x_m^j \gamma^* n,m
\]

\[
= K_1 - 2 \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} \hat{x}_n \cdot x_m^j \gamma^* n,m,
\]

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

\[
C(\gamma; \mu^0, \mu^j) = K_1 - 2 \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} x_0 n \cdot x_m^j \gamma n,m
\]

\[
= K_1 - 2 \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} \tilde{\gamma}_n \cdot \tilde{\gamma}_m x_0 n \cdot x_m^j
\]

Analogously, we have

\[
C(\hat{\gamma}^*; \mu^0, \hat{\mu}^j) = \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} \| \hat{x}_n - \hat{x}_m^j \|^2 \hat{\gamma}^* n,m
\]

\[
= \sum_{n=1}^{N_0} \| x_0 n \|^2 p_n^0 + \sum_{m=1}^{N_j} \| x_m^j \|^2 p_n^0 - 2 \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} x_0 n \cdot \hat{x}_m^j \hat{\gamma}^* n,m
\]

\[
= K_2 - 2 \sum_{n=1}^{N_0} \sum_{m=1}^{N_j} x_0 n \cdot \hat{x}_m^j \gamma^* n,m
\]

\[
= K_2 - K_1 + C(\gamma^*; \mu^0, \mu^j) \quad (46)
\]

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

14
where $K_2$ is constant (which only depends on $\mu^0$ and $\mu^f$) and similarly
\[
C(\gamma^*; \mu^0, \mu^f) = K_2 - 2 \sum_{n=1}^{N_0} \sum_{m=1}^{N_1} \sum_{\ell=1}^{N_0} \frac{\gamma_n \gamma_{\ell,m}}{P^0_{\ell,n}} x_n^0 \cdot x_m^f
\]
\[
= K_2 - K_1 + C(\gamma; \mu^0, \mu^f)
\] (47)

Therefore, by (46) and (47), we have that $C(\gamma^*; \mu^0, \mu^f) \leq C(\gamma; \mu^0, \mu^f)$ if and only if $C(\gamma^*; \mu^0, \mu^f) \leq C(\gamma; \mu^0, \mu^f)$. So, we conclude the proof by the fact $\gamma^*$ is optimal for $OT(\mu_0, \mu^f)$.

**C. Proof of Proposition 2.2**

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

**Proof.** Let $\gamma^* = \text{diag}(p_1^0, \ldots, p_{N_0}^0)$ 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
\[
C(\gamma^*; \mu^0, \mu^f) \leq C(\gamma; \mu^0, \mu^f).
\]

We have
\[
C(\gamma^*; \mu^0, \mu_t) = \sum_{k,k'=1}^{N_0} \left\Vert x_k^0 - (1-t)x_k^1 - tx_k^f \right\Vert^2 \gamma^*_{kk'}
\]
\[
= \sum_{k=1}^{N_0} \left\Vert x_k^0 \right\Vert^2 p_k^0 + \sum_{k=1}^{N_0} \left\Vert (1-t)x_k^1 + tx_k^f \right\Vert^2 p_k^0 - 2 \sum_{k,k'=1}^{N_0} x_k^0 \cdot ((1-t)x_k^1 + tx_k^f)
\]
\[
= K - 2 \left( (1-t) \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^1 \gamma^*_{kk'} + t \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^f \gamma^*_{kk'} \right),
\] (48)

where in third equation, $K$ is a constant which only depends on the marginals $\mu^0, \mu_t$. Similarly,
\[
C(\gamma; \mu^0, \mu^f) = K - 2 \left( (1-t) \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^1 \gamma_{kk'} + t \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^f \gamma_{kk'} \right).
\] (49)

By the fact that $\gamma^* \in \Gamma(\hat{\mu}, \mu_t) = \Gamma(\mu^0, \mu^1)$, and that $\gamma^*$ is optimal for $OT(\hat{\mu}, \mu^1)$, $OT(\hat{\mu}, \mu^2)$, we have
\[
\sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^1 \gamma^*_{kk'} \geq \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^1 \gamma_{kk'} \quad \text{and} \quad \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^f \gamma^*_{kk'} \geq \sum_{k,k'=1}^{N_0} \hat{x}_k \cdot x_k^f \gamma_{kk'}.
\]

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

**Proof of Proposition 2.2.** Consider $0 \leq s \leq t \leq 1$. By Lemma 2.1, the maps $T_i^0: x_k^0 \mapsto \hat{x}_k^i$, $T_j^0: x_k^0 \mapsto \hat{x}_k^j$ 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
\[
x_k^0 \mapsto (1-s)\hat{x}_k^i + s\hat{x}_k^i = x_k^0 + (1-s)u_k^i + su_k^j, \quad 1 \leq k \leq N_0
\]
solves $OT(\mu^0, \hat{\rho}_s)$, and similarly
\[
x_k^0 \mapsto (1-t)\hat{x}_k^i + t\hat{x}_k^j = x_k^0 + (1-t)u_k^i + tu_k^j, \quad 1 \leq k \leq N_0
\]

15
solves $OT(\mu^0, \hat{\rho}_t)$.

Thus,

$$\text{LOT}^\nu(\hat{\rho}_s, \hat{\rho}_t) = \sum_{k=1}^{N_0} \|((1-s)\hat{x}_k^i + s\hat{x}_k^j) - ((1-t)\hat{x}_k^i + t\hat{x}_k^j)\|^2 p_k^0$$

$$= \sum_{k=1}^{N_0} (t-s)^2 \|\hat{x}_k^i - \hat{x}_k^j\|^2 p_k^0$$

$$= (t-s)^2 \text{LOT}^\nu(\mu^i, \mu^j),$$

and this finishes the proof. \(\square\)

**D. Proof of proposition 3.1**

This result relies on the definition (29) of the curve $\rho_t$, 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 $\rho_t$ as a solution of a (homogeneous) continuity equation ($\gamma_t$ defined in (28)), and the second term ($\gamma_t := (1-t)\nu_0 + t\nu^\dagger$) as an appropriate particular solution of the equation (26) with forcing term $\zeta$ defined in (31). Moreover, the curve $\rho_t$ will become optimal for (25) since both $\gamma_t$ and $\gamma_t$ are "optimal" in different senses. On the one hand, $\gamma_t$ is optimal for a classical optimal transport problem\(^\ddagger\). On the other hand, $\gamma_t$ 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

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

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

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

since $\gamma_0$ and $\gamma_1$ are dominated by $\mu^0$ and $\mu^j$, 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 $(\rho, v, \zeta)$ is a solution for equation (26), we mean that it is a weak solution or, equivalently, it is a solution in the distributional sense (Santambrogio, 2015, Section 4.4). That is, for any test function $\psi : \Omega \to \mathbb{R}$ continuous differentiable with compact support, $(\rho, v, \zeta)$ satisfies

$$\partial_t \left( \int_{\Omega} \psi \, d\rho_t \right) - \int_{\Omega} \nabla \psi \cdot v \, d\rho_t = \int_{\Omega} \psi \, d\zeta,$$

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

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

**Lemma D.1.** If $\gamma^*$ is optimal for $\text{OPT}_\lambda(\mu^0, \mu^1)$, and has marginals $\gamma_0^*$ and $\gamma_1^*$, then $\gamma^*$ is optimal for $\text{OPT}_\lambda(\mu^0, \gamma_1^*)$, $\text{OPT}(\gamma_0^*, \mu^1)$ and $\text{OT}(\gamma_0^*, \gamma_1^*)$.

\(^\ddagger\)Here we will strongly use that the fixed partial optimal transport plan $\gamma^* \in \Gamma^*_p(\mu^0, \mu^1)$ in the hypothesis of the proposition has the form $\gamma^* = (I \times T)_#\gamma_0^*$, for a map $T : \Omega \to \Omega$. 

16
Linear Optimal Partial Transport Embedding

**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

$$C(\gamma; \mu^0, \mu^1, \lambda) = \int_{\Omega^2} \|x^0 - x^1\|^2 d\gamma(x^0, x^1) + \lambda(||\mu^0 - \gamma|| + ||\mu^1 - \gamma||)$$

$$= C(\gamma; \gamma^0, \gamma^1) + \lambda(||\mu^0 - \gamma|| + ||\mu^1 - \gamma||).$$

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, $\gamma^*$ is optimal for $OT(\gamma^0, \gamma^1)$.

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

**Lemma D.2.** If $\gamma^* \in \Gamma_\leq(\mu^0, \mu^1)$, then $d(\text{supp}(\nu^0), \text{supp}(\nu^1)) \geq \sqrt{2\lambda}$, 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 $d(\text{supp}(\nu^0), \text{supp}(\nu^1)) < \sqrt{2\lambda}$. Then, there exist $\bar{\nu}^0 \in \text{supp}(\nu^0)$ and $\bar{\nu}^1 \in \text{supp}(\nu^1)$ such that $||\nu^0 - \nu^1|| < \sqrt{2\lambda}$. Moreover, we can choose $\varepsilon > 0$ such $\nu^0(B(\bar{\nu}^0, \varepsilon)) > 0$ and $\nu^1(B(\bar{\nu}^1, \varepsilon)) > 0$, where $B(\bar{\nu}^i, \varepsilon)$ denotes the ball in $\mathbb{R}^d$ of radius $\varepsilon$ centered at $\bar{\nu}^i$.

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

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

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

$$\pi_i \# \bar{\gamma} = \gamma_i + \delta \bar{\nu}^i \leq \gamma_i + \nu^i = \mu^i \quad \text{for } i = 0, 1.$$

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

$$C(\bar{\gamma}, \mu^0, \mu^1, \lambda) = \int_{\Omega^2} ||x^0 - x^1||^2 d\bar{\gamma} + \lambda(||\mu^0|| + ||\mu^1|| - 2||\gamma||)$$

$$= \int_{\Omega^2} ||x^0 - x^1||^2 d\gamma + \delta \int_{\Omega^2} ||x^0 - x^1||^2 d(\bar{\nu}^0 \times \bar{\nu}^1) + \lambda(||\mu^0|| + ||\mu^1|| - 2||\gamma|| - 2\delta||\bar{\nu}^0 \times \bar{\nu}^1||)$$

$$= OPT_\lambda(\mu^0, \mu^1) + \delta \int_{B(\bar{\nu}^0, \varepsilon) \times B(\bar{\nu}^1, \varepsilon)} ||x^0 - x^1||^2 d(\bar{\nu}^0 \times \bar{\nu}^1) - 2\lambda\delta$$

$$< OPT_\lambda(\mu^0, \mu^1) + 2\lambda\delta - 2\lambda\delta = OPT_\lambda(\mu^0, \mu^1),$$

which is a contradiction.

**Lemma D.3.** Using the notation and hypothesis of Proposition 3.1, for $t \in (0, 1)$ let $D_t := \{x : \nu_t(x) \neq 0\}$. Then, $\gamma_t \wedge \nu_0 \equiv \gamma_t \wedge \nu^3 \equiv 0$ 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 $\gamma^*$ is optimal for $OT(\gamma^*_0, \gamma^*_1)$. Since we are also assuming that $\gamma$ is induced by a map $T$, i.e. $\gamma^* = (I \times T)\# \gamma^*_0$, then $T$ is an optimal Monge map between $\gamma^*_0$ and $\gamma^*_1$. Therefore, the map $T_t : \Omega \rightarrow \Omega$ is invertible (see for example (Santambrogio, 2015, Lemma 5.29)). For ease of notation we will denote $\nu^0 := \gamma_t \wedge \nu_0$ as a measure in $D_t$.

The idea of the proof is to exhibit a plan $\bar{\gamma}$ with smaller $OPT_\lambda$ cost than $\gamma^*$, which will be a contradiction since $\gamma^* \in \Gamma_\leq(\mu^0, \mu^1)$. Let us define

$$\bar{\gamma} := \gamma^* + (I, T \circ T_t^{-1})\# \nu^0 - (I, T)\# (T_t^{-1})\# \nu^0.$$
Then, \( \tilde{\gamma} \in \Gamma_{\leq}^{\leq} \) since

\[
\pi_{0} \# \tilde{\gamma} = \pi_{0} \# \gamma^* + \tilde{\nu}^0 - (T^{-1}_t) \# \tilde{\nu}^0 \leq \gamma^*_0 + \tilde{\nu}^0 \leq \mu^0,
\pi_{1} \# \tilde{\gamma} = \pi_{1} \# \gamma^* + (T \circ T^{-1}_t) \# \tilde{\nu}^0 - (T \circ T^{-1}_t) \# \tilde{\nu}^0 = \gamma^*_1. \quad \text{(and we know } \gamma^*_1 \leq \mu^1 \).
\]

From the last equation we can also conclude that \( |\tilde{\gamma}| = |\gamma^*_1| = |\gamma^*| \). Therefore, the partial transportation cost for \( \tilde{\gamma} \) is

\[
\int_{\Omega^2} \| x - y \|^2 d\tilde{\gamma}(x, y) + \lambda(|\mu^0| + |\mu^1| - 2|\tilde{\gamma}|)
= \int_{\Omega^2} \| x - y \| d\gamma^*(x, y) + \lambda(|\mu^0| + |\mu^1| - 2|\gamma|) + \int_{D_t} \| x - T(\gamma_{-1}(x)) \|^2 - \| \gamma_{-1}(x) - T(\gamma_{-1}(x)) \|^2 \) \( d\tilde{\nu}^0(x) \)
< OPT_{\lambda}(\mu^0, \mu^1)
\]

where in the last inequality we used that if \( y = T_{-1}(x) \) we have

\[
\| T_{1}(y) - T(y) \| = \| (1 - t)y + tT(y) - T(y) \| = (1 - t)\| y - T(y) \| < \| y - T(y) \| \quad \text{for } t \in (0, 1).
\]

\begin{lemma}
Using the notation and hypothesis of Proposition 3.1, for \( t \in (0, 1) \) the vector-valued measures \( v_t \cdot \nu_0 \) and \( v_t \cdot \nu^1 \) are exactly zero.
\end{lemma}

\begin{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 \( \Phi : \Omega \rightarrow \mathbb{R}^d \), 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 \( \gamma_t \cap \nu_0 \equiv 0 \) over \( D_t \). Therefore, \( \gamma_t \) and \( \nu_0 \) 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 \( \gamma_t \)-almost everywhere (up to sets of null measure with respect to \( \gamma_t \)), we can assume without loss of generality that \( v_t \equiv 0 \) on \( D_1 \).

Let \( \Phi : \Omega \rightarrow \mathbb{R}^d \) be a measurable vector-valued function over \( \Omega \). 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 \( \Phi \) was arbitrary, we can conclude that \( v_t \cdot \nu_0 \equiv 0 \).
\end{proof}

\begin{lemma}
Using the notation and hypothesis of Proposition 3.1, the measure \( \tau_t = (1 - t)\nu_0 + t\nu^1 \) satisfies the equation

\[
\begin{cases}
\partial_t \tau + \nabla \cdot \tau v = \zeta \\
\tau_0 = \nu_0, \quad \tau_1 = \nu^1.
\end{cases}
\]
\end{lemma}

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

\begin{proof}[Proof of Proposition 3.1]
First, we address that the \( v : [0, 1] \times \Omega \rightarrow \mathbb{R}^d \) is well defined. Indeed, by Lemma D.1, we have that \( \gamma^* = \text{id} \times T \# \gamma^* \) is optimal for \( OT(\gamma^*_0, \gamma^*_1) \). Thus, for each \((t, x) \in [0, 1] \times \Omega, \) by so-called \textit{cyclical monotonicity property} of the support of
We recall that we are denoting by \(v_t(x)\) the vector of weights of \(\nu_t\) such that \(v_t(x) = x\), see (Santambrogio, 2015, Lemma 5.29). So, \(v_t(x)\) in (30) is well defined by understanding its definition as

\[
v_t(x) = \begin{cases} T(x_0) - x_0 & \text{if } x = T_t(x_0), \\ 0 & \text{elsewhere.} \end{cases}
\]

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

\[
\rho_t = \gamma_t + \tau_t
\]

where \(\gamma_t\) is given by (28), and \(\tau_t\) is as in Lemma D.5. Then, from section 2.2, \((\gamma, v)\) 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 \(\gamma_0^*)\). On the other hand, from Lemma D.5, \(\tau_t\) solves (50). Therefore, by linearity,

\[
\partial_t \rho_t + \nabla \cdot \rho_t v_t = \partial_t \gamma_t + \nabla \cdot \gamma_t v_t + \nabla \cdot \gamma_t v_t = \zeta_t
\]

Finally, by plugging in \((\rho, v, \zeta)\) into the objective function in (25), we have:

\[
\int_{[0,1] \times \Omega} \|v_t\|^2 d \rho_t + \lambda|\zeta| = \int_{[0,1] \times \Omega} \|v_t\|^2 d \gamma_t dt + \lambda|\zeta|
\]

\[
= \int_{\Omega} \|T(x^0) - x^0\|^2 d \gamma_0^*(x^0) + \lambda|\nu_0| + |\nu^1|
\]

\[
= \int_{\Omega^2} \|x^0 - x^j\|^2 d \gamma^*(x^0, x^j) + \lambda|\mu_0 - \gamma_0^*| + |\mu^j - \gamma_1^*|
\]

\[
= OPT(\mu^0, \mu^j)
\]

since \(\gamma^* \in \Gamma^*(\mu^0, \mu^j)\). So, this shows that \((\rho, v, \zeta)\) is minimum for (25).

The ‘moreover’ part holds from the above identities, using that

\[
\int_{\Omega} \|v_0\|^2 d \gamma_0^*(x_0) + \lambda|\nu_0| + |\nu^1| = \int_{\Omega} \min(\|v_0\|^2, 2\lambda) d \gamma_0^*(x_0) + \lambda|\nu_0| + |\nu^1|
\]

for \(v_0(x^0) = T(x^0) - x^0\),

since \(\gamma^*\) is such that \(\|x^0 - x^j\|^2 < 2\lambda\) for all \((x^0, x^j) \in \text{supp}(\gamma)\) (Bai et al., 2022, Lemma 3.2).

\[
\square
\]

**E. Proof of Theorem 3.5**

The proof will be similar to that of Lemma 2.1.

**Proof.** We fixed \(\gamma^j \in \Gamma^*_\leq(\mu^0, \mu^j)\). We will understand the second marginal distribution \(\pi_{1\#} \gamma^j\) induced by \(\gamma^j\), either as the vector \(\gamma^j_1 := (\gamma^j)^T 1_{N_0}\) or as the measure \(\sum_{m=1}^{N_j} (\gamma^j_1)_m \delta_{x^j_m}\), and, by abuse of notation, we will write \(\pi_{1\#} \gamma^j = \gamma^j_1\) (analogously for \(\pi_{0\#} \gamma^j\)).

Let \(\hat{\gamma}^j \coloneqq \text{diag}(\tilde{p}^1_{k_1}, \ldots, \tilde{p}^j_{N_0})\) denote the transportation plan induced by mapping \(x^0_k \mapsto \tilde{x}^i_k\). Let \(D := \{k \in \{1, \ldots, N_0\} : \tilde{p}^j_k > 0\}\).

Given any \(\hat{\gamma} \in \Gamma_{\leq}(\mu^0, \tilde{\mu}^j)\), we need to show

\[
C(\hat{\gamma}^j; \mu^0, \tilde{\mu}^j, \lambda) \leq C(\hat{\gamma}; \mu^0, \tilde{\mu}^j, \lambda).
\]

(51)

We recall that we are denoting by \(p^0\) the vector of weights of the discrete measure \(\mu^0\) (analogously, \(p^j\) and \(\tilde{p}^j\) are the vectors of weights of \(\mu^j\) and \(\tilde{\mu}^j\), resp).
In general, if we consider any transportation plan \( \gamma' \in \Gamma \leq (\mu_0, \gamma^j_1) \), we have that
\[
C(\gamma'; \mu^0, \gamma^j_1, \lambda) = \sum_{k=1}^{N_0} \sum_{i=1}^{N_j} ||x^0_k - x^1_i||^2 \gamma'_{k,i} + \lambda(|p^0| + |\gamma^j_1| - 2|\gamma'|)
\]
where \( \gamma_{k,i} = \hat{\gamma}_{k,i} \) if \( k \in D \) and \( \gamma_{k,i} = 0 \) elsewhere. In fact, we have
\[
\begin{align*}
\gamma_{1,N_0} & = \hat{\gamma}^{j} \frac{1}{\gamma^j_0} \gamma^j_0 \Gamma \leq (\mu_0, \hat{\mu}^j) \\
\gamma^T 1_{N_0} & = (\gamma^j)^T \frac{1}{\gamma^j_0} \gamma^j_0 \Gamma \leq (\mu_0, \hat{\mu}^j)
\end{align*}
\]
where \( \beta \in \mathbb{R}^{N_0 \times 1} \) is defined with \( \beta_k = 1 \) if \( k \in D \) and \( \beta_k = 0 \) elsewhere. In fact, in (52), the first inequality follows from the fact each entry in the \( N_0 \times 1 \) vector \( \frac{1}{\gamma^j_0} \hat{\mu}^j \) is either 0 or 1; and the second inequality holds since \( \hat{\gamma} \in \Gamma \leq (\mu_0, \hat{\mu}^j) \). And in (53), the inequality follows from the fact \( \hat{\gamma}^T \in \Gamma \leq (\mu_0, \hat{\mu}^j) \).

In addition, since \( (\hat{\gamma}^j)^T 1_{N_0} \leq \hat{\mu}^j \), we have \( \hat{\gamma}_{k,k'} = 0, \forall k' \notin D, k \in \{1, \ldots, N_0\} \). Thus we have
\[
\gamma_{1,N_0} = \hat{\gamma}^j \beta = \hat{\gamma} 1_{N_0}
\]
that is the first marginals of \( \gamma \) and \( \hat{\gamma} \) are same and thus \( |\gamma| = |\hat{\gamma}| \).

We compute the transportation costs induced by \( \gamma^j, \gamma, \hat{\gamma}, \gamma' \):
\[
C(\gamma^j; \mu^0, \gamma^j_1, \lambda) = \sum_{i=1}^{N_j} \sum_{k \in D} ||x^0_k - x^1_i||^2 \gamma^j_{k,i} + \lambda(|p^0| + |\gamma^j_1| - 2|\gamma^j|)
\]
\[
= \sum_{k \in D} ||x^0_k||^2 (\gamma^j_0)_k + \sum_{i=1}^{N_j} ||x^1_i||^2 (\gamma^j_1)_i - 2 \sum_{k \in D} \sum_{i=1}^{N_j} x^0_k \cdot x^1_i \gamma^j_{k,i} + \lambda(|p^0| + |\gamma^j_1| - 2|\gamma^j|)
\]
\[
= \sum_{k \in D} ||x^0_k||^2 \hat{\mu}^j_k + \sum_{i=1}^{N_j} ||x^1_i||^2 (\gamma^j_1)_i - 2 \sum_{k \in D} \sum_{i=1}^{N_j} x^0_k \cdot x^1_i \gamma^j_{k,i} + \lambda(|p^0| + |\gamma^j_1| - 2|\gamma^j|)
\]
Similarly,

\[
C(\hat{\gamma}^j; \mu^0, \hat{p}^j, \lambda) = \sum_{k \in D} \|x_k^0\|^2(\hat{\gamma}^j_0)k + \sum_{k' \in D} \|\hat{x}_{k'}^j\|^2\hat{p}_{k'}^j - 2 \sum_{k \in D} \sum_{k' \in D} x_k^0 \cdot \hat{x}_{k'}^j \hat{k}_{k,k'} + \lambda(|p^0| + |\hat{p}^j| - 2|\hat{\gamma}^j|)
\]

\[
= \sum_{k \in D} \|x_k^0\|^2\hat{p}_{k}^j + \sum_{k' \in D} \|\hat{x}_{k'}\|^2\hat{p}_{k'}^j - 2 \sum_{k \in D} \sum_{k' \in D} x_k^0 \cdot \hat{x}_{k'}^j \hat{\gamma}^j_{k,k'} + \lambda(|p^0| + |\hat{\gamma}^j| - 2|\hat{\gamma}^j|)
\]

\[
= \sum_{k \in D} \|x_k^0\|^2\hat{p}_{k}^j + \sum_{k' \in D} \|\hat{x}_{k'}\|^2\hat{p}_{k'}^j - 2 \sum_{k \in D} \sum_{k' \in D} x_k^0 \cdot \hat{x}_{k'}^j \hat{\gamma}^j_{k,k'} + \lambda(|p^0| + |\hat{\gamma}^j| - 2|\hat{\gamma}^j|)
\]

\[
= C(\gamma^j; \mu^0, \gamma^j_1, \lambda) - \sum_{i=1}^N \|x_i^j\|^2(\gamma^j_1)i + \sum_{k' \in D} \|\hat{x}_{k'}^j\|^2\hat{p}_{k'}^j
\]  \hspace{1cm} (54)

And also,

\[
C(\gamma; \mu^0, \gamma^j_1, \lambda) = \sum_{k=1}^{N_0} \|x_k^0\|^2(\gamma^j_0)k + \sum_{i=1}^{N_1} \|x_i^j\|^2(\gamma^j_1)i - 2 \sum_{k=1}^{N_0} \sum_{k'=1}^{N_1} \sum_{k' \in D} x_k^0 \cdot x_{k'}^j \hat{\gamma}^j_{k,k'} \hat{\gamma}^j_{k,k'} + \lambda(|p^0| + |\hat{\gamma}^j| - 2|\hat{\gamma}^j|)
\]

\[
= \sum_{k=1}^{N_0} \|x_k^0\|^2\hat{p}_{k}^j + \sum_{k' \in D} \|\hat{x}_{k'}\|^2\hat{p}_{k'}^j - 2 \sum_{k=1}^{N_0} \sum_{k' \in D} x_k^0 \cdot \hat{x}_{k'}^j \hat{\gamma}^j_{k,k'} + \lambda(|p^0| + |\hat{\gamma}^j| - 2|\hat{\gamma}^j|)
\]

\[
= C(\gamma; \mu^0, \gamma^j_1, \lambda) - \sum_{i=1}^N \|x_i^j\|^2(\gamma^j_1)i + \sum_{k' \in D} \|\hat{x}_{k'}^j\|^2\hat{p}_{k'}^j
\]  \hspace{1cm} (55)

Let \(\vec{p} := \pi_1\hat{\gamma}\) (vector of weights). We have \(\vec{p} \leq \hat{p}^j\). Combining with (54) and (55), we have

\[
C(\hat{\gamma}^j; \mu^0, \hat{p}^j, \lambda) - C(\gamma; \mu^0, \hat{p}^j, \lambda)
\]

\[
= C(\gamma^j; \mu^0, \gamma^j_1, \lambda) - C(\gamma; \mu^0, \gamma^j_1, \lambda) + \sum_{i=1}^{N_1} \|x_i^j\|^2 \sum_{k \in D} (\hat{p}_k^j - \hat{p}_{k'}^j) + \sum_{k' \in D} \|\hat{x}_{k'}^j\|^2(\hat{p}_k^j - \vec{p})
\]

\[
\leq \sum_{i=1}^{N_1} \|x_i^j\|^2 \sum_{k \in D} (\hat{p}_k^j - \hat{p}_{k'}^j) + \sum_{k' \in D} \|\hat{x}_{k'}^j\|^2(\hat{p}_k^j - \vec{p})
\]  \hspace{1cm} (56)

where the inequality holds since \(\gamma^j\) is optimal for \(OPT_\lambda(\mu^0, \gamma^j_1)\).

Let

\[
F(\vec{p}) := \sum_{i=1}^{N_1} \|x_i^j\|^2 \sum_{k \in D} (\hat{p}_k^j - \hat{p}_{k'}^j) + \sum_{k' \in D} \|\hat{x}_{k'}^j\|^2(\hat{p}_k^j - \vec{p})
\]

Note that if \(k \notin D\), then \(\vec{p}_k = \hat{p}_k = 0\); and for \(k \in D\), we have

\[
\frac{\partial}{\partial \vec{p}_k} F = \sum_{i: \gamma_{i}^j > 0} \|x_i^j\|^2 \frac{\gamma_{i}^j}{\hat{p}_k^j} - \|\hat{x}_{k}^j\|^2 \geq 0
\]

where the inequality holds by the fact \(\sum_{i: \gamma_{i}^j > 0} \frac{\gamma_{i}^j}{\hat{p}_k^j} = 1\) and Jensen’s inequality (in particular, \(E(\|X\|^2) \geq \|E[X]\|^2\) for a random vector \(X\)). Combining with the fact \(F(\vec{p}) = 0\) when \(\vec{p}_k\) achieves its maximum \(\hat{p}_k^j\) for each \(k\), we have

\[
F(\vec{p}) \leq 0, \quad \forall 0 \leq \vec{p} \leq p^j.
\]

Thus \(C(\hat{\gamma}^j; \mu^0, \hat{p}^j, \lambda) \leq C(\gamma; \mu^0, \hat{p}^j, \lambda)\) by (56), and we conclude the proof. □
F. Proof of Theorem 3.6

Proof. Without loss of generality, we assume that $\gamma^j \in \Gamma^+_\infty(\mu^0, \mu^j)$ 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 $\mu^j$ when we get $\hat{\mu}^j$. For example, if two $x_m^0$ and $x_n^0$ are mapped to the same point $T(x_m^0) = T(x_n^0)$, 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, $\gamma^j \in \mathbb{R}^{N_0 \times N_0}$ is such that in each row and column there exists at most one positive entry, and all others are zero. Let $\hat{\gamma} := \text{diag}(\hat{p}_1^j, \ldots, \hat{p}_{N_0}^j)$. Then, by the definition of $\hat{\mu}^j$, we have $C(\gamma^j; \mu^0, \mu^j, \hat{\mu}) = C(\hat{\gamma}; \mu^0, \hat{\mu}^j, \lambda)$. Thus,

$$OPT(\mu^0, \mu^j) = C(\gamma^j; \mu^0, \mu^j, \lambda)$$

$$= C(\gamma^j, \mu^0, \nu^j) + \lambda(\|\nu^j| - |\gamma^j|)$$

$$= C(\gamma^j; \mu^0, \nu^j, \lambda) + \lambda(\|\nu^j| - |\gamma^j|)$$

$$= OPT(\mu^0, \nu^j) + \lambda(\|\nu^j| - |\gamma^j|),$$

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

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

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

\hfill \Box

G. Proof of Proposition 3.7

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

$$LOPT_{\lambda, \mu^0}(\hat{\mu}^j, \tilde{\mu}^j) = \|u^i - u^j\|_{\hat{p}^j \wedge \tilde{p}^j, 2\lambda} + \lambda(||\nu^j - \tilde{\nu}^j|)$$

and

$$LOPT_{\lambda, \mu^0}(\mu^j, \mu^j) = \|u^i - u^j\|_{\hat{p}^j \wedge \tilde{p}^j, 2\lambda} + \lambda(||\nu^j - \tilde{\nu}^j|),$$

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

\hfill \Box

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 (Wang et al., 2013) and (Moosmüller & Cloninger, 2023), 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 ($T_\mu^0$ space). Moreover, it does it in a way that allows us to use the $T_\mu^0$-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 $\lambda$ and each number of measures $K = 5, 10, 15$.  

22
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, $\mu^i \sim N(m^i, I)$, where $m^i$ is randomly selected such that $\|m^i\| = \sqrt{3}$ for $i = 1, \ldots, K$. For the reference, we picked an $N$ point representation of $\mu^0 \sim N(m, I)$ with $m = \sum m^i / K$. For figures 2 and 6, the sample size $N$ was set equal to 500.

In what follows, we include tests for $N = 200, 250, 300, 350, \ldots, 900, 950, 1000$ and $K = 2, 4$. For each $(N, K)$, we repeated each experiment 10 times. The relative errors are shown in Figure 7. For large $\lambda$, most mass is transported and $\text{OT}_{\lambda}(\mu_i, \mu_j) \approx \text{OPT}_{\lambda}(\mu_i, \mu_j)$, the performance of LOPT is close to that of LOT, and the relative error is small. For small $\lambda$, almost no mass is transported, $\text{OPT}_{\lambda}(\mu_i, \mu_j) \approx \lambda(|\mu_i| + |\mu_j|) \approx \lambda(|\mu_i| - |\hat{\mu}_i| + |\mu_j| - |\hat{\mu}_j|) \approx \text{LOPT}_{\mu^0, \lambda}(\mu_i, \mu_j)$, 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.

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

**Figure 7.** The average relative error between $\text{OPT}_{\lambda}$ and $\text{LOPT}_{\lambda, \mu_0}$ between all pairs of discrete measures $\mu^i$ and $\mu^j$.

**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 \( \eta \) are included here for completeness. Later, in Section H.4, Figure 11 will add the plots for the HK technique and its linearized version.

![Figure 8. Interpolation between two point-clouds at different times \( t \in \{0, 0.25, 0.5, 0.75, 1\} \). Different values of noise \( \eta \) 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.](image-url)
H.4. Preliminary comparisons between LHK and LOPT

The contrasts between the Linearized Hellinger-Kantorovich (LHK) (Cai et al., 2022) 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.

![Figure 9](image)

*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.*

In addition, in Figure 10 we not only compare HK and OPT, but also LHK and LOPT. The top subfigure 10a shows the measures \( \mu_1 \) (blue dots) and \( \mu_2 \) (orange crosses) to be interpolated with the different techniques in the next subfigures 10b and 10c. Also, in 10a we plot the reference \( \mu_0 \) (green triangles) which is going to be used only in experiment 10c. The measures \( \mu_1 \) and \( \mu_2 \) 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 \( \mu_1 \) and \( \mu_2 \). 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 \( \mu_1 \) and \( \mu_2 \) for LHK and LOPT. The reference measure \( \mu_0 \) 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 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.
Linear Optimal Partial Transport Embedding

(a) Plot of two measures ($\mu_1$ as blue circles and $\mu_2$ as orange crosses, where each point has mass one), and a reference measure $\mu_0$ (concentrated on the three green triangular locations, unit mass each).

(b) HK and OPT interpolation between the two measures $\mu_1$ and $\mu_2$ at different times. The color code is the same as for Figure 9.

(c) LHK and LOPT interpolation between the two measures $\mu_1$ and $\mu_2$ at different times using for both cases the same reference $\mu_0$.

Figure 10. HK vs OPT and LHK vs LOPT
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.

![Figure 11](image)

(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.

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 $\lambda$ and the choice of the initial reference measure are critical in this process.

---

**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) $\mu_1$ (digit 0), $\mu_2$ (digit 1), and $\mu_3$ (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 $\mu_1$, $\mu_2$, $\mu_3$ 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.