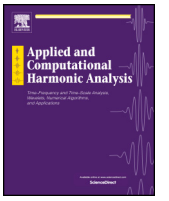
Applied and Computational Harmonic Analysis 74 (2025) 101718

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

journal homepage: www.elsevier.com/locate/acha



Full Length Article

# Linearized Wasserstein dimensionality reduction with approximation guarantees

Alexander Cloninger [a,b], Keaton Hamm [c,d,*], Varun Khurana [e], Caroline Moosmüller [f]

[a] Department of Mathematics, University of California, San Diego, CA, United States of America
[b] Halıcıoğlu Data Science Institute, University of California, San Diego, CA, United States of America
[c] Department of Mathematics, University of Texas at Arlington, Arlington, TX, United States of America
[d] Division of Data Science, College of Science, University of Texas at Arlington, Arlington, TX, United States of America
[e] Department of Applied Mathematics, Brown University, Providence, RI 02906, United States of America
[f] Department of Mathematics, University of North Carolina at Chapel Hill, NC, United States of America

ARTICLE INFO

Communicated by Massimo Fornasier

*MSC:*
49Q22
60D05
68T10

*Keywords:*
Optimal transport
Dimensionality reduction
Wasserstein space
Multidimensional scaling
Isomap

ABSTRACT

We introduce LOT Wassmap, a computationally feasible algorithm to uncover low-dimensional structures in the Wasserstein space. The algorithm is motivated by the observation that many datasets are naturally interpreted as probability measures rather than points in $\mathbb{R}^n$, and that finding low-dimensional descriptions of such datasets requires manifold learning algorithms in the Wasserstein space. Most available algorithms are based on computing the pairwise Wasserstein distance matrix, which can be computationally challenging for large datasets in high dimensions. Our algorithm leverages approximation schemes such as Sinkhorn distances and linearized optimal transport to speed-up computations, and in particular, avoids computing a pairwise distance matrix. We provide guarantees on the embedding quality under such approximations, including when explicit descriptions of the probability measures are not available and one must deal with finite samples instead. Experiments demonstrate that LOT Wassmap attains correct embeddings and that the quality improves with increased sample size. We also show how LOT Wassmap significantly reduces the computational cost when compared to algorithms that depend on pairwise distance computations.

**Contents**



* Corresponding author.
*E-mail addresses:* acloninger@ucsd.edu (A. Cloninger), keaton.hamm@uta.edu (K. Hamm), varun_khurana@brown.edu (V. Khurana), cmoosm@unc.edu (C. Moosmüller).

https://doi.org/10.1016/j.acha.2024.101718
Received 14 February 2023; Received in revised form 30 September 2024; Accepted 7 October 2024

Available online 15 October 2024

## 1. Introduction

A classical problem in analyzing large volume, high-dimensional datasets is to develop efficient algorithms that classify points based on a similarity measure, or based on a subset of preclassified training data points. Even when data points lie in high-dimensional Euclidean space, they can often be approximated by low-dimensional structures, such as subspaces or submanifolds. This observation has led to significant advances in the field, mostly through the development of *manifold learning algorithms*, which produce a low-dimensional representation of a given dataset; see for example [8,15,26,38]. In many of these frameworks, the data points are assumed to be sampled from a low-dimensional Riemannian manifold embedded in Euclidean space, and approximately preserve intrinsic properties such as geodesic distances.

In many applications however, data points are more naturally interpreted as distributions $\{\mu_i\}_{i=1}^N$ over $\mathbb{R}^n$, or finite samples $X_i = \{x_j^{(i)}\}_{j=1}^{N_i}$ with $x_j^{(i)} \sim \mu_i$. Examples include imaging data [36], text documents (the bag-of-word model uses word count within a text as features, creating a histogram for each document [45]), and gene expression data, which can be interpreted as a distribution over a gene network [14,28]. In this setting, a Euclidean embedding space with Euclidean distances locally approximating the intrinsic distance of the data may not be geometrically meaningful, and datasets are better modeled as probability measures in the *Wasserstein space* [39].

We assume that our data points $\{\mu_i\}_{i=1}^N$ belong to the quadratic Wasserstein space $W_2(\mathbb{R}^n)$ of probability measures with finite second moment, equipped with the Wasserstein distance

$$W_2(\mu,\nu) := \inf_{\pi\in\Gamma(\mu,\nu)} \left( \int_{\mathbb{R}^{2n}} \|x-y\|^2 d\pi(x,y) \right)^{\frac{1}{2}}, \tag{1}$$

where $\mathscr{P}(\mathbb{R}^{2n})$ is the set of all probability measures over $\mathbb{R}^{2n}$ and $\Gamma(\mu,\nu) := \{\gamma \in \mathscr{P}(\mathbb{R}^{2n}) : \gamma(A\times\mathbb{R}^n) = \mu(A),\ \gamma(\mathbb{R}^n\times A) = \nu(A) \text{ for all } A\subset\mathbb{R}^n\}$ is the set of all joint probability measures with marginals $\mu$ and $\nu$. Under regularity assumptions on $\mu$, the optimal coupling $\pi$ has the form $\pi = (\mathrm{id}, T)_\sharp\mu$, where $T \in \mathrm{L}^2(\mathbb{R}^n,\mu)$ is the "optimal transport map" [10,39].

The Wasserstein space and optimal transport have gained popularity in the machine learning community, as they are based on a solid theoretical foundation [39] (for example, (1) is a metric), while providing a versatile framework for applications (for example, as a cost function for generative models [6], in semi-supervised learning [37], and in pattern detection for neuronal data [31]).

In this paper, we are interested in uncovering low-dimensional submanifolds in the Wasserstein space in a *computationally feasible* manner as well as analyzing the quality of the embedding. To this end, we follow the idea of [21,40], which introduces the *Wassmap* algorithm (see Section 2.6 for more details), a version of the Multidimensional Scaling algorithm (MDS) [27] (see Algorithm 1), or more generally, the Isomap algorithm [38].

A central part of manifold learning algorithms like MDS or Isomap relies on the computation of the pairwise Euclidean distances. Wassmap uses the pairwise Wasserstein distance matrix instead, which leads to $O(N^2)$ Wasserstein distance computations, each of which is of the order $O(n^3\log(n))$ if one uses interior point methods to solve the linear program (1). If both $N$ and $n$ are large, computing all pairwise distances becomes infeasible. To deal with this issue, approximations of the Wasserstein distance can be considered. In this paper, we are interested in *entropic regularized* distances (Sinkhorn distances) [2,17], which deal with the computational issue involving $n$, and in *linearized optimal transport* (LOT) [20,40], to reduce the computational cost in $N$.

Our results are twofold:

(1) **Approximation guarantees**:
- We provide bounds on the embedding quality of the Multidimensional Scaling algorithm (MDS) [27] (see Algorithm 1) applied to a dataset in the Wasserstein space, where the pairwise Wasserstein distances are only available up to an error $\tau$.
- We study the size of $\tau$ in common approximation schemes such as entropic regularization and linearized approximations, *and* when explicit descriptions of the data points $\mu_i, i = 1,\dots,N$ are not available, and one must deal with finite samples instead.

(2) **Efficient algorithm (LOT Wassmap)**: We provide an algorithm, "LOT Wassmap", inspired by the Wassmap algorithm of [21]. It essentially uses linearized Wasserstein distance approximations through LOT in the Multidimensional Scaling algorithm, leveraging our approximation guarantees from (1). However, we *do not* compute the LOT-Wasserstein distance matrix and feed it into MDS, but instead compute the truncated SVD of centered transport maps. This is the same in theory, but computationally more efficient.

### 1.1. Previous work

The idea of replacing pairwise Euclidean distances with pairwise Wasserstein distances in common manifold learning algorithms has been explored in many settings; for example in [44] to study shape spaces of proteins, in [28,14] to analyze gene expression data, and in [40] for cancer detection.

Theoretical results on the reconstruction of certain submanifolds in $W_2(\mathbb{R}^n)$ through the MDS algorithm using pairwise Wasserstein distances are presented in [21]. The associated algorithm, Wassmap, is the basis for our LOT Wassmap algorithm.

Related to the idea of uncovering submanifolds in the Wasserstein space is "Wasserstein dictionary learning" as discussed in [33,42]. The authors propose to represent complex data in the Wasserstein space as Wasserstein barycenters of a dictionary.

### 1.2. Approximation guarantees

Using approximations of the Wasserstein distance in manifold learning algorithms such as MDS may change the embedding quality, and our main result provides theoretical bounds on the error:

**Theorem 1.1** *(Informal version of Theorem 3.3). Assume that data points $\{\mu_i\}_{i=1}^N$ are $\tau_1$-close to a $d$-dimensional submanifold $\mathscr{W}$ in the Wasserstein space, which is isometric to a subset $\Xi$ of Euclidean space $\mathbb{R}^d$. Furthermore assume that we only have access to approximations $\lambda_{ij}$ of the pairwise distances $W_2(\mu_i, \mu_j)$, and that the approximation error is $\tau_2$.*

*Then, under some technical assumptions, the Multidimensional Scaling algorithm using distances $\lambda_{ij}$ as input recovers data points $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$, which are $C_{N,\mathscr{W}}(\tau_1 + \tau_2)$-close to $\Xi$ up to rigid transformations.*

Some remarks on this result:

- The first source of error, $\tau_1$, depends on how close the data points are to the submanifold $\mathscr{W}$ isometric to a subspace of $\mathbb{R}^d$, which is completely determined by the dataset.
- The second source of error, $\tau_2$, depends on the approximation scheme used, and can be made arbitrarily small with sufficient computational time or good choice of parameters.

A significant part of this paper is dedicated to providing bounds for $\tau_2$, when common approximation schemes for $W_2(\mu_i, \mu_j)$ are used, and when $\{\mu_i\}_{i=1}^N$ are only available through samples, i.e. when $\mu_i \approx \widehat{\mu}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \delta_{Y_j^{(i)}}$ with $Y_j^{(i)} \sim \mu_i$ i.i.d. In particular, we introduce *empirical linearized Wasserstein-2 distance*, $\widehat{W}_{2,\sigma}^{\mathrm{LOT}}$, which uses two approximation schemes:

(a) *Entropic regularized formulation*: A very successful approximation framework for efficient Wasserstein distance computation is the entropic regularized formulation of (1), which depends on a parameter $\beta$, and leads to *Sinkhorn distances* [17]:

$$\min_{\pi\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^{2n}} \frac{1}{2}\|x-y\|^2 d\pi(x,y) + \beta D_{\mathrm{KL}}(\pi\|\mu\otimes\nu), \tag{2}$$

where $D_{\mathrm{KL}}$ is the Kullback–Leibler divergence of measures [23]. This formulation leads to a unique solution (in contrast to (1)), and to a significant computational speed-up in $n$, achieving $O(n^2\log(n))$ through matrix scaling algorithms (Sinkhorn's algorithm) [2,17].

(b) *Linearized Wasserstein distances*: Linearized optimal transport (LOT) [20,40] approximates Wasserstein distances by linear $\mathrm{L}^2$-distances in the tangent space at a chosen reference measure $\sigma$:

$$W_{2,\sigma}^{\mathrm{LOT}}(\mu,\nu) := \left(\int_{\mathbb{R}^n} \|T_\sigma^\mu(x) - T_\sigma^\nu(x)\|^2 \, d\sigma(x)\right)^{1/2}, \tag{3}$$

where $T_\sigma^\mu$ denotes the optimal transport map from $\sigma$ to $\mu$ (either computed through (1) or (2), and using barycentric projections to make a transport plan into a transport map). Instead of computing all pairwise optimal transport maps, in this framework, one computes $T_\sigma^{\mu_i}$ from $\sigma$ to $\mu_i$, and approximates pairwise maps between $\mu_i$ and $\mu_j$ as a composition of $T_\sigma^{\mu_i}$ and $T_\sigma^{\mu_j}$, reducing the computation in $N$ to $O(N)$. This framework has been successfully applied signal and image classification tasks [34,41], such as visualizing phenotypic differences between types of cells [7]. There furthermore exist error bounds for $W_{2,\sigma}^{\mathrm{LOT}}$ [9,19,20,25,29,32].

With these approximation schemes at hand, we define the *empirical linearized Wasserstein-2 distance*:

$$\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu}) := \left(\frac{1}{m}\sum_{j=1}^{m} \|T_\sigma^{\widehat{\mu}}(X_j) - T_\sigma^{\widehat{\nu}}(X_j)\|^2\right)^{1/2}, \tag{4}$$

where $X_j \sim \sigma$ i.i.d. and the transport maps are either computed by (1) or (2) (and with barycentric projections, if necessary).

We provide values for $\tau_2$ as in Theorem 1.1, by bounding $|W_2(\mu,\nu)^2 - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu},\widehat{\nu})^2|$, using either a linear program or Sinkhorn iterations to compute the transport plans. These bounds are derived by combining the following results:

- Estimation of optimal transport maps with plug-in estimators, i.e. bounds on $\|T^{\widehat{\nu}}_{\mu} - T^{\nu}_{\mu}\|_{\mu}$, which are provided by [18] for the linear program case, and by [35] in the regularized case. Both [18] and [35] assume compactly supported $\mu$ and $\nu$, while we are able to relax the compact support assumption on the target measure, as long as it can be approximated by compactly supported measures.
- Approximation results for $W^{\mathrm{LOT}}_{2,\sigma}$, which are provided in [25,32], and are based on the idea that $\mu_i$ are generated by almost compatible functions $\mathscr{H}$ applied to a fixed generator $\mu$. We also strengthen some of the approximation results in [25,32].

### 1.3. Efficient algorithm: LOT Wassmap

The Wassmap algorithm of [21] requires computing the pairwise Wasserstein distance matrix $W_2(\mu_i,\mu_j)$, $i,j = 1,\dots,N$, which leads to $O(N^2)$ expensive computations. We introduce *LOT Wassmap* (see Algorithm 2), which uses LOT distances (3) to linearly approximate $W_2(\mu_i,\mu_j)$ (since the input of our algorithm are empirical samples $\widehat{\mu}_i$, we actually use the empirical linearized Wasserstein-2 distance (4)). This results in only $O(N)$ optimal transport computations.

However, in practice, we avoid computing the pairwise LOT distance matrix. Instead, we compute the truncated SVD of the centered transport maps, which is computationally more efficient. We show that in theory this produces a result equivalent to Theorem 1.1:

**Corollary 1.2** *(Informal version of Corollary 3.4). Assume that data points $\{\mu_i\}_{i=1}^N$ are $\tau_1$-close to a $d$-dimensional submanifold $\mathscr{W}$ in the Wasserstein space, which is isometric to a subset $\Xi$ of Euclidean space $\mathbb{R}^d$. Choose a reference measure $\sigma$ and compute all transport maps $T^{\mu_i}_{\sigma}$ (either with a linear program (1) or with Sinkhorn approximations (2), and with barycentric projections, if necessary). Let $\tau_2$ be the error between the empirical linearized Wasserstein-2 distance $\widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu}_i,\widehat{\mu}_j)$ of (4) and the actual Wasserstein-2 distance $W_2(\mu_i,\mu_j)$.*

*Then, under some technical assumptions, the truncated SVD of the centered transport maps $T^{\mu_i}_{\sigma}$ (column-stacked) produces data points $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$, which are $C_{N,\mathscr{W}}(\tau_1+\tau_2)$-close to $\Xi$ up to rigid transformations.*

We note that Corollary 1.2 is a corollary of Theorem 1.1 and that the technical assumptions and constants are the same in both results.

In Section 8, we provide experiments demonstrating that LOT Wassmap does attain correct embeddings given finite samples without explicitly computing the pairwise LOT distance matrix. In particular, we show that the embedding quality improves with increased sample size and that LOT Wassmap significantly reduces the computational cost when compared to Wassmap.

### 1.4. Organization of the paper

This paper is organized as follows: We start by introducing important notation and background in Section 2. This includes discussion of the MDS and Wassmap algorithms, (linearized) optimal transport, and plug-in estimators. Section 3 introduces the LOT Wassmap algorithm and provides the main results. Sections 4 and 5 provide approximation guarantees for $\widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu},\widehat{\nu})$ for compactly and non-compactly supported target measures, respectively. The approximation guarantees come with many technical assumptions, and Sections 6 and 7 are dedicated to discussing settings in which these assumptions hold. The paper concludes with experiments in Section 8, which show the effectiveness of LOT Wassmap. Proofs are provided in Appendices A to D.

## 2. Notation and background

This paper has a significant amount of background and notation which is summarized categorically here. See Table 1 for an overview of notation used in the paper.

### 2.1. Linear algebra preliminaries

Given $A \in \mathbb{R}^{m\times n}$, its *Singular Value Decomposition* (SVD) is given by $A = U\Sigma V^{\top}$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m\times n}$ has non-zero entries along its main diagonal (singular values). The singular values are the square roots of the eigenvalues of $A^{\top}A$ and are taken in descending order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0$. The truncated SVD of order $d$ of $A$ is $A_d = U_d\Sigma_d V_d^{\top}$ where $U_d$ and $V_d$ consist of the first $d$ columns of $U$ and $V$, respectively, and $\Sigma_d = \mathrm{diag}(\sigma_1,\dots,\sigma_d) \in \mathbb{R}^{d\times d}$. The Moore–Penrose pseudoinverse of $A \in \mathbb{R}^{m\times n}$ is the $n\times m$ matrix denoted by $A^{\dagger}$ and defined by $A^{\dagger} = V\Sigma^{\dagger}U^{\top}$ where $\Sigma^{\dagger}$ is the $n\times m$ matrix with entries $\frac{1}{\sigma_1},\dots,\frac{1}{\sigma_{\min\{m,n\}}}$ along its main diagonal.

The Schatten $p$-norms ($1 \leq p \leq \infty$) are a general class of unitarily invariant, submultiplicative norms on $\mathbb{R}^{m\times n}$ and are defined to be the $\ell^p$ norms of the vector of singular values: $\|A\|_{S_p} := \|(\sigma_1,\dots,\sigma_{\min\{m,n\}})\|_{\ell^p}$. The Frobenius norm, which is the Schatten 2-norm is denoted by $\|\cdot\|_{\mathrm{F}}$, and the spectral norm, which is the Schatten $\infty$-norm is denoted simply by $\|\cdot\|$. We also use $\|\cdot\|$ to denote the Euclidean norm of a vector.

**Table 1**
Overview of notation used in the paper.

| Notation | Definition | Reference |
|---|---|---|
| $\Delta$ | Square Euclidean distance matrix | Algorithm 1 |
| $\Lambda$ | Perturbed distance matrix | Corollary 3.2 |
| $X^\dagger$ | Moore–Penrose pseudoinverse of matrix $X$ | Section 2.1 |
| $\mu$ | Template measure | Section 2.4 |
| $\widehat{\mu}$ | Empirical measure approximating $\mu$ | (7) |
| $\sigma$ | Reference measure for LOT | Section 2.4 |
| $\|\cdot\|_{S_p}$ | Schatten $p$-norm | Section 2.1 |
| $\|\cdot\|$ | Spectral norm of a matrix or Euclidean norm of a vector | Section 2.1 |
| $\|\cdot\|_F$ | Frobenius norm of a matrix | Section 2.1 |
| $\|\cdot\|_{\max}$ | (Entrywise) maximum norm of a matrix | Section 2.1 |
| $\|\cdot\|_\mu$ | Norm on $L^2(\mathbb{R}^n, \mu)$ | Section 2.3 |
| $n$ | Dimension of Euclidean space that probability measures are defined on | Section 2.3 |
| $\mathscr{P}(\mathbb{R}^n)$ | Probability measures on $\mathbb{R}^n$ | Section 2.3 |
| $\mathscr{P}_{ac}(\mathbb{R}^n)$ | Absolutely continuous probability measures on $\mathbb{R}^n$ | Section 2.3 |
| $W_2(\mathbb{R}^n)$ | Wasserstein-2 space over $\mathbb{R}^n$ | Section 2.3 |
| $W_2(\mu, \nu)$ | Wasserstein-2 distance between $\mu$ and $\nu$ | (5) |
| $W^{\mathrm{LOT}}_{2,\sigma}(\mu, \nu)$ | Linearized Wasserstein-2 distance between $\mu$ and $\nu$, with $\sigma$ as reference | (6) |
| $\widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\mu, \nu)$ | Empirical linearized Wasserstein-2 distance | (12) |
| $T^\mu_\sigma$ | Optimal transport (Monge) map from $\sigma$ to $\mu$ | Section 2.3 |
| $T_\sharp\mu$ | Pushforward of $\mu$ with respect to $T$ | Section 2.3 |
| $T^{\widehat{\mu}}_\sigma$ | Barycentric projection of an optimal transport plan (Kantorovich potential) | (10) |
| $d$ | Embedding dimension of MDS | Section 2.2 |
| $k$ | Sample size that generates $\widehat{\mu}$ | (7) |
| $m$ | Sample size that generates $\widehat{\sigma}$ | Algorithm 2 |
| $N$ | Number of data points | Algorithm 2 |
| $\varepsilon$ | Distance from compatibility | Definition 2.2 |
| $\beta$ | Regularizer for Sinkhorn OT | Section 4.2 |

### 2.2. Multidimensional scaling

Let $\mathbf{1}$ be the all-ones vector in $\mathbb{R}^N$, and $J := I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. Then Multidimensional Scaling (MDS) is summarized in Algorithm 1. For more details see [27].

**Algorithm 1:** Multidimensional Scaling (MDS) [27].

**Input:** Points $\{y_i\}_{i=1}^N \subset \mathbb{R}^D$; embedding dimension $d \ll D$.
**Output:** Low-dimensional embedding points $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$
Compute pairwise distance matrix $\Delta_{ij} = \|y_i - y_j\|^2$
$B = -\frac{1}{2} J \Delta J$
(Truncated SVD): $B_d = V_d \Sigma_d V_d^\top$
$z_i = (V_d \Sigma_d)(i, :)$, for $i = 1, \ldots, N$
**Return** $\{z_i\}_{i=1}^N$

MDS produces an isometric embedding $\mathbb{R}^D \to \mathbb{R}^d$ if and only if the matrix $B$ is symmetric positive semi-definite with rank $d$, a result that goes back to Young and Householder [43]. In this case, the embedding points $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$ satisfy $\|z_i - z_j\| = \|y_i - y_j\|$ and are unique up to rigid transformation.

### 2.3. Optimal transport preliminaries

Let $\mathscr{P}(\mathbb{R}^n)$ be the space of all probability measures on $\mathbb{R}^n$, with $\mathscr{P}_{ac}(\mathbb{R}^n)$ being the subset of all probability measures which are absolutely continuous with respect to the Lebesgue measure. Given $\mu \in \mathscr{P}_{ac}(\mathbb{R}^n)$, we denote its probability density function by $f_\mu$. The quadratic Wasserstein space $W_2(\mathbb{R}^n)$ is the subset of $\mathscr{P}(\mathbb{R}^n)$ of measures with finite second moment $\int_{\mathbb{R}^n} \|x\|^2 d\mu(x) < \infty$ equipped with the quadratic Wasserstein metric given by

$$W_2(\mu, \nu) := \inf_{\pi \in \Gamma(\mu,\nu)} \left( \int_{\mathbb{R}^{2n}} \|x - y\|^2 d\pi(x, y) \right)^{\frac{1}{2}}, \tag{5}$$

where $\Gamma(\mu, \nu) := \{\gamma \in \mathscr{P}(\mathbb{R}^{2n}) : \gamma(A \times \mathbb{R}^n) = \mu(A),\ \gamma(\mathbb{R}^n \times A) = \nu(A) \text{ for all } A \subset \mathbb{R}^n\}$ is the set of couplings, i.e., measures on the product space whose marginals are $\mu$ and $\nu$.

In [10], Brenier showed that if $\mu$ is absolutely continuous with respect to the Lebesgue measure, the optimal coupling of (5) takes the special form $\pi = (\mathrm{id}, T_\mu^\nu)_\sharp \mu$, where $\sharp$ is the pushforward operator ($S_\sharp \mu(A) = \mu(S^{-1}(A))$ for $A$ measurable) and $T_\mu^\nu \in \mathrm{L}^2(\mathbb{R}^n, \mu)$ solves

$$\min_{T : T_\sharp \mu = \nu} \int_{\mathbb{R}^n} \|T(x) - x\|^2 \, d\mu(x).$$

For simplicity, we denote the norm on $\mathrm{L}^2(\mathbb{R}^n, \mu)$ by $\|f\|_\mu^2 := \int_{\mathbb{R}^n} \|f(x)\|^2 d\mu(x)$. Note that if $T_\mu^\nu$ exists, then

$$W_2(\mu, \nu) = \|T_\mu^\nu - \mathrm{id}\,\|_\mu.$$

Furthermore, [10] shows that when $\mu$ is absolutely continuous with respect to the Lebesgue measure, the map $T_\mu^\nu$ is uniquely defined as the gradient of a convex function $\phi$, i.e. $T_\mu^\nu = \nabla \phi$ (up to an additive constant).

### 2.4. Linearized optimal transport

Linearized optimal transport (LOT) [20,29,34,41] defines an embedding of $\mathscr{P}(\mathbb{R}^n)$ into the linear space $\mathrm{L}^2(\mathbb{R}^n, \sigma)$, with $\sigma$ being a fixed reference measure. Under the assumption that the optimal transport map exists, the embedding is defined by $\mu \mapsto T_\sigma^\mu$. This embedding can be used as a feature space, for example, to classify subsets of $\mathscr{P}(\mathbb{R}^n)$, to linearly approximate the Wasserstein distance, or for fast Wasserstein barycenter computations [1,25,29,32,34].

In particular, the LOT embedding defines a linearized Wasserstein-2 distance:

$$W_{2,\sigma}^{\mathrm{LOT}}(\mu, \nu) := \|T_\sigma^\mu - T_\sigma^\nu\|_\sigma. \tag{6}$$

In certain settings, this linearized distance approximates the Wasserstein-2 distance. The strongest results can be obtained when the so-called *compatibility condition* is satisfied:

**Definition 2.1** *(Compatibility condition [1,32,34]).* Let $\sigma, \mu \in W_2(\mathbb{R}^n) \cap \mathscr{P}_{\mathrm{ac}}(\mathbb{R}^n)$. We say that the LOT embedding is compatible with the $\mu$-pushforward of a function $g \in \mathrm{L}^2(\mathbb{R}^n, \mu)$ if

$$T_\sigma^{g_\sharp \mu} = g \circ T_\sigma^\mu.$$

The compatibility condition describes an interaction between the optimal transport map and the pushforward operator, namely it requires invertability of the exponential map [20].

When the compatibility condition holds for two functions $g_1, g_2$, then LOT is an isometry, i.e. $W_{2,\sigma}^{\mathrm{LOT}}(g_{1\sharp}\mu, g_{2\sharp}\mu) = W_2(g_{1\sharp}\mu, g_{2\sharp}\mu)$ as shown in Lemma A.3 and [32,34]. In particular, this is the case when $g$ is either a shift or scaling, or a certain type of shearing [25,32,34].

We can furthermore consider a generalization to "almost compatible" functions, also termed $\varepsilon$-compatible:

**Definition 2.2** *($\varepsilon$-compatibility).* Let $\sigma, \mu \in W_2(\mathbb{R}^n) \cap \mathscr{P}_{\mathrm{ac}}(\mathbb{R}^n)$. We say that $\mathscr{H}$ is $\varepsilon$- compatible with respect to $\sigma$ and $\mu$, if for every $h \in \mathscr{H}$, there exists a compatible transformation $g$ such that $\|g - h\|_\mu < \varepsilon$, where $g \circ T_\sigma^\mu = T_\sigma^{g_\sharp \mu}$.

We remark that compatibility is stable. Similar to compatibility implying isometry, there exist results that imply $\varepsilon$-compatible transformations imply "almost"-isometry between $W_{2,\sigma}^{\mathrm{LOT}}$ and $W_2$. Some of these results are accounted for in [32, Proposition 4.1]; however, we also extend these almost-compatibility results in Theorem A.4. These results make use of the Hölder regularity bounds for $W_{2,\sigma}^{\mathrm{LOT}}$ of [20,29]. We note that the "isometry under compatibility" result mentioned above is a direct consequence of the preceding proposition, namely by setting $\varepsilon = 0$.

In this paper, we consider measures $\mu_i, i = 1, \ldots, N$ of the form $\mu_i = h_{i\sharp}\mu$, where $\mu$ is a fixed *template measure*, and $h \in \mathscr{H}$ with $\mathscr{H}$ a space of functions in $\mathrm{L}^2(\mathbb{R}^n, \mu)$. This is similar to assumptions in [1,25,32,34], where $\mathscr{H}$ consists of shifts and scalings, compatible maps, or has other properties, such as convexity and compactness. We will write $\mu_i \sim \mathscr{H}_\sharp \mu$ to indicate that $\mu_i$ is of such a form for all $i = 1, \ldots, N$, and $\mathscr{H}$ will be specified in the respective context. Note that [1] calls this data generation process an "algebraic generative model".

### 2.5. Optimal transport with plug-in estimators

Explicit descriptions of the measures $\mu$ are often unavailable in applications, and one must instead deal with finite samples of the measure. In this paper, we consider empirical distributions

$$\widehat{\mu} = \frac{1}{k} \sum_{i=1}^{k} \delta_{Y_i} \tag{7}$$

with $Y_i \sim \mu$ i.i.d. In what follows, we will consider approximations of both the target and reference distributions via empirical distributions.

The Kantorovich problem (5) has a (possibly non-unique) solution for transporting an absolutely continuous measure $\sigma$ to an empirical measure of the form (7). Following [18], we define the set of Kantorovich plans

$$\Gamma_{\min} := \underset{\pi\in\Gamma(\sigma,\widehat{\mu})}{\operatorname{argmin}} \int_{\mathbb{R}^{2n}} \|x-y\|^2 d\pi(x,y), \tag{8}$$

which may contain more than one transport plan. In practice, these optimal transport plans are exactly computed via linear programming to solve (8). We call optimal transport plans solved with linear programming $\gamma_{LP}$. It is much faster, however, to approximate the optimal transport plan by using an entropic regularized plan [17]. In particular, we get a unique solution by solving

$$\gamma_\beta := \underset{\pi\in\Gamma(\sigma,\widehat{\mu})}{\operatorname{argmin}} \int \frac{1}{2}\|x-y\|^2 d\pi(x,y) + \beta D_{\mathrm{KL}}(\pi\|\sigma\otimes\widehat{\mu}), \tag{9}$$

where $D_{\mathrm{KL}}$ is the Kullback–Leibler divergence of measures [23], $\sigma\otimes\widehat{\mu}$ is the direct product measure on the product space $\mathbb{R}^n\times\mathbb{R}^n$, and $\beta$ denotes the regularizer. We solve (9) with Sinkhorn's algorithm, which yields entropic potentials $f_\beta$ and $g_\beta$ corresponding to $\sigma$ and $\widehat{\mu}$, respectively.

Regardless of whether we solve the optimal transport plan using (8) or (9), we can make a transport plan $\gamma\in\Gamma$ into a map by defining the barycentric projection

$$T_\sigma^{\widehat{\mu}}(x;\gamma) := \frac{\int_y y\, d\gamma(x,y)}{\int_y d\gamma(x,y)}, \quad \text{for } x\in\operatorname{supp}(\sigma). \tag{10}$$

**Remark 2.3.** Notice that if $\mu=\widehat{\mu}$ (i.e. $\mu$ is a finite atomic measure), then $T_\sigma^{\mu}$ and $T_\sigma^{\widehat{\mu}}$ are the same $\sigma$-almost everywhere assuming $\sigma$ is the absolutely continuous. If $\sigma$ is absolutely continuous, then the Kantorovich solution can be written as a product measure with the Monge solution as one of the products. Thus, since no mass from $\sigma$ splits, the barycentric projection $T_\sigma^{\widehat{\mu}}$ is the same as $T_\sigma^{\mu}$ $\sigma$-almost everywhere.

This leads to a natural way to consider linearized Wasserstein-2 distances of the form (6) with absolutely continuous reference $\sigma$, and for empirical distributions:

$$W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu};\gamma) := \|T_\sigma^{\widehat{\mu}}(\,\cdot\,;\gamma_{\widehat{\mu}}) - T_\sigma^{\widehat{\nu}}(\,\cdot\,;\gamma_{\widehat{\nu}})\|_\sigma, \tag{11}$$

where $\gamma\in\{\gamma_{LP},\gamma_\beta\}$ denotes the method used to calculate the transport plans $\gamma_{\widehat{\mu}}$ and $\gamma_{\widehat{\nu}}$, which are transport plans from $\sigma$ to $\widehat{\mu}$ and $\widehat{\nu}$, respectively. We suppress this notation and will simply use $T_\sigma^{\widehat{\mu}}(\,\cdot\,;\gamma_{LP})$ or $T_\sigma^{\widehat{\mu}}(\,\cdot\,;\gamma_\beta)$ to denote the barycentric projection map computed via linear programming and Sinkhorn, respectively, so that $\gamma_{LP}$ and $\gamma_\beta$ are understood to be in $\Gamma(\sigma,\widehat{\mu})$. Notice that the solution $\gamma_{LP}$ is not necessarily unique. In this case, the results we derive for $W_{2,\sigma}^{\mathrm{LOT}}$ and $T_\sigma^{\widehat{\mu}}(\,\cdot\,;\gamma_{LP})$ still work with high probability as we use concentration inequalities and results that hold with high probability.

To account for $m$ finite samples of the reference distribution, we define the empirical linearized Wasserstein-2 distance by

$$\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu};\gamma) := \left(\frac{1}{m}\sum_{j=1}^{m}\|T_\sigma^{\widehat{\mu}}(X_j;\gamma_{\widehat{\mu}}) - T_\sigma^{\widehat{\nu}}(X_j;\gamma_{\widehat{\nu}})\|^2\right)^{1/2}, \tag{12}$$

where $X_j\sim\sigma$ i.i.d.

**Remark 2.4.** When we use $\gamma_\beta$ for a transport plan between $\widehat{\sigma}$ and $\widehat{\mu}$, note that our barycentric projection map is given by

$$T_{\widehat{\sigma}}^{\widehat{\mu}}(x;\gamma_\beta) := \frac{\frac{1}{k}\sum_{i=1}^{k} y_i \exp\left(\left(g_{\beta,k}(y_i) - \frac{1}{2}\|x-y_i\|^2\right)/\beta\right)}{\frac{1}{k}\sum_{i=1}^{k}\exp\left(\left(g_{\beta,k}(y_i) - \frac{1}{2}\|x-y_i\|^2\right)/\beta\right)}, \tag{13}$$

where $g_{\beta,k}$ denotes the entropic potential corresponding to $\widehat{\mu}$, $y_i\in\operatorname{supp}(\widehat{\mu})$, and $k$ is the sample size for both $\widehat{\sigma}$ and $\widehat{\mu}$.

**Remark 2.5.** Since our approximations will require us to use $m$ samples from the reference distributions, the barycentric projection map $T_\sigma^{\widehat{\mu}}(x)$ will only work for $x\in\operatorname{supp}(\widehat{\sigma})$; however, for general computation, we can just interpolate to calculate $T_\sigma^{\widehat{\mu}}(x)$ for $x\in\operatorname{supp}(\sigma)\setminus\operatorname{supp}(\widehat{\sigma})$.

In what follows, we are interested in bounds for

$$|W_2(\mu,\nu)^2 - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu};\gamma)^2|$$

for $\gamma \in \{\gamma_{LP}, \gamma_\beta\}$. In particular, we want similar results to Theorem A.4 (Wasserstein-2 compared to LOT) and results in [18] (Wasserstein-2 compared to Wasserstein-2 on empirical distributions). This requires comparisons between all of $W_2(\mu,\nu)$, $W_{2,\sigma}^{\mathrm{LOT}}(\mu,\nu)$, $W_{2,\sigma}^{\mathrm{LOT}}(\hat{\mu},\hat{\nu};\gamma)$, and $\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\hat{\mu},\hat{\nu};\gamma)$, which are discussed in Section 4 and Section 5.

### 2.6. Wassmap

Various generalizations of MDS have been explored [16] including stress minimization, which is useful in graph drawing [24,30], Isomap [38] which replaces pairwise distance by a graph estimation of manifold geodesics, and is useful for embedding data from $d$–dimensional nonlinear manifolds in $\mathbb{R}^D$. Wang et al. [40] utilized MDS with $\Delta_{ij} = W_2(\mu_i,\mu_j)^2$ for data considered as probability measures in Wasserstein space with applications to cell imaging and cancer detection. Subsequently, Hamm et al. [21] proved that several types of submanifolds of $W_2$ can be isometrically embedded via MDS with Wasserstein distances (as in [40]) and empirically studied Wassmap: a variant of Isomap that approximates nonlinear submanifolds of $W_2$. In particular, [21] shows that for some submanifolds of $W_2(\mathbb{R}^m)$ of the form $\mathscr{H}_\sharp\mu$ where $\mathscr{H} = \{h_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ which are isometric Euclidean space, the parameter set $\Theta \subset \mathbb{R}^d$ can be recovered up to rigid transformation via MDS with Wasserstein distances (e.g., translations and anisotropic dilations).

### 2.7. Other notations

For scalars $a$ and $b$ we use $a \vee b$ to denote the maximum and $a \wedge b$ to denote the minimum value of the pair. Throughout the paper, constants will typically be denoted by $C$ and may change from line to line, and subscripts will be used to denote dependence on a given set of parameters. We use $a \asymp b$ to mean that $ca \le b \le Ca$ for some absolute constants $0 < c, C < \infty$.

For a random variable $X_n$, we say that $X_n = O_p(a_n)$ if for every $\varepsilon > 0$ there exists $M > 0$ and $N > 0$ such that

$$\mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > M\right) < \varepsilon \quad \forall n \ge N.$$

We denote by $\mathscr{O}(d)$ the orthogonal group over $\mathbb{R}^d$, and the related Procrustes distance (in the Frobenius norm) between matrices $X, Y \in \mathbb{R}^{d\times N}$ is $\min_{Q\in\mathscr{O}(d)} \|X - QY\|_{\mathrm{F}}$.

## 3. LOT Wassmap algorithm and main theorem

Here we present our main algorithm which is an LOT approximation to the Wassmap embedding of [21], and our main theorem which describes the quality of the embedding using some existing perturbation bounds for MDS.

### 3.1. The LOT Wassmap embedding algorithm

The algorithm presented here (Algorithm 2) takes discretized samples of a set of measures $\{\mu_i\}_{i=1}^N \subset W_2(\mathbb{R}^n)$ and a discretized sample of a reference measure $\sigma \in W_2(\mathbb{R}^n)$, computes transport maps from the empirical reference measure $\widehat{\sigma}$ to each empirical target measure $\widehat{\mu}_i$ using optimal transport solvers and barycentric projections. Finally, the truncated right singular vectors and singular values of the centered transport map matrix are used to produce the low-dimensional embedding of the measures. Two things are important to note here: first, the output of the algorithm is the same as the output of multi-dimensional scaling using pairwise squared LOT distances (or Sinkhorn distances in the approximate case), but we use the same trick as the reduction of PCA to the SVD to avoid actually computing the distance matrix; second, in contrast to the Wassmap embedding of [21] which requires $O(N^2)$ Wasserstein distance computations, Algorithm 2 requires computation of only $O(N)$ optimal transport maps. Given the high cost of computing a single optimal transport map for densely sampled measures, this represents significant savings.

Note that the factor of $\frac{1}{\sqrt{m}}$ appearing in the computation of the final embedding is due to (12) where the $\frac{1}{m}$ appears in the definition of the empirical LOT distance. Lemma A.1 shows that $T^\top T$ where $T$ is as in Algorithm 2 is actually the MDS matrix $-\frac{1}{2}J\Lambda J$ where $\Lambda$ consists of the empirical LOT distances between the data, hence we absorb the $\frac{1}{m}$ into the norm in (12) to get the matrix $T$ in Algorithm 2. If we have a $\ell \times k$ matrix $A$, colstack($A$) is a $\ell \cdot k$ vector constructed by stacking the columns on top of each other.

### 3.2. MDS perturbation bounds

As stated above, the output of Algorithm 2 is equivalent to the output of MDS on the transport map matrix $T$ therein. Consequently, the analysis of the algorithm will require some results regarding MDS. On the road to stating our main result, we summarize some nice MDS perturbation results of [5].

**Theorem 3.1** *([5, Theorem 1]). Let $Y, Z \in \mathbb{R}^{d\times N}$ with $d < N$ such that* $\mathrm{rank}(Y) = d$, *and let $\varepsilon^2 := \|Z^\top Z - Y^\top Y\|_{S_p}$ for some $p \in [1,\infty]$. Then,*

**Algorithm 2:** LOT WassMap embedding.

**Input:** Reference point cloud $\{w_i\}_{i=1}^m \sim \sigma \in W_2(\mathbb{R}^n)$
Sample point clouds $\{x_j^k\}_{j=1}^{n_k} \sim \mu_k \in W_2(\mathbb{R}^n)$ $(k = 1, \dots, N)$
OT solver (with regularizer if Sinkhorn)
Embedding dimension $d$
**Output:** Low-dimensional embedding points $\{z_i\}_{i=1}^N \subseteq \mathbb{R}^d$
**for** $k = 1, \dots, N$ **do**
  Calculate cost matrix $C_{ij} = \|w_i - x_j^k\|^2$
  Compute OT plan $\gamma_k \in \mathbb{R}^{m \times n_k}$ between $\{w_i\}_{i=1}^m$ and $\{x_j^k\}_{j=1}^{n_k}$ using $C$ and OT solver
  Calculate barycentric projection $\widetilde{T}_k(w_i) = \left(\sum_{j=1}^{n_k} x_j^k (\gamma_k)_{ij}\right) / \left(\sum_{j=1}^{n_k} (\gamma_k)_{ij}\right)$
$\widehat{T} = \left[\text{colstack}\{\widetilde{T}_j(w_i)\}_{i=1}^m\right]_{j=1}^N$
**for** $k = 1, \dots, N$ **do**
  $T_{:k} = \frac{1}{\sqrt{m}}(\widehat{T}_{:k} - \frac{1}{N}\sum_{k=1}^N \widehat{T}_{:k})$
Compute the truncated SVD of $T$ as $T_d = U_d \Sigma_d V_d^\top$
**Return** $z_i = V_d \Sigma_d(i, :)$

$$\min_{Q \in \mathcal{O}(d)} \|Z - QY\|_{S_p} \leq \begin{cases} \|Y^\dagger\|\varepsilon^2 + \left((1 - \|Y^\dagger\|^2\varepsilon^2)^{-\frac{1}{2}}\|Y^\dagger\|\varepsilon^2\right) \wedge d^{\frac{1}{2p}}\varepsilon, & \|Y^\dagger\|\varepsilon < 1, \\ \|Y^\dagger\|\varepsilon^2 + d^{\frac{1}{2p}}\varepsilon, & \text{o.w.} \end{cases}$$

*Consequently, if* $\|Y^\dagger\|\varepsilon \leq \frac{1}{\sqrt{2}}$, *then*

$$\min_{Q \in \mathcal{O}(d)} \|Z - QY\|_{S_p} \leq (1 + \sqrt{2})\|Y^\dagger\|\varepsilon^2.$$

**Corollary 3.2.** *Let* $y_1, \dots, y_N \in \mathbb{R}^d$ *be centered, span* $\mathbb{R}^d$, *and have pairwise dissimilarities* $\Delta_{ij} = \|y_i - y_j\|^2$. *Let* $\{\Lambda_{ij}\}_{i,j=1}^N$ *be arbitrary real numbers and* $p \in [1, \infty]$. *If* $\|Y^\dagger\|\|\Lambda - \Delta\|_{S_p}^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}}$, *then MDS (Algorithm 1) with input dissimilarities* $\{\Lambda_{ij}\}_{i,j=1}^N$ *and embedding dimension* $d$ *returns a point set* $z_1, \dots, z_N \in \mathbb{R}^d$ *satisfying*

$$\min_{Q \in \mathcal{O}(d)} \|Z - QY\|_{S_p} \leq (1 + \sqrt{2})\|Y^\dagger\|\|\Lambda - \Delta\|_{S_p}.$$

**Proof of Corollary 3.2.** The proof follows along similar lines to that of [5, Corollary 2] with some modifications. First, note that the centering matrix $J$ in MDS satisfies $\|J\| = 1$ as it is an orthogonal projection. Then, by using the fact that $\|AB\|_{S_p} \leq \|A\|\|B\|_{S_p}$, we can estimate

$$\frac{1}{2}\|J(\Lambda - \Delta)J\|_{S_p} \leq \frac{1}{2}\|J\|^2\|\Lambda - \Delta\|_{S_p} \leq \frac{1}{2}\|\Lambda - \Delta\|_{S_p} < \sigma_d^2(Y), \tag{14}$$

where the final inequality follows by assumption.

Since $Y$ is a centered point set, we have $Y^\top Y = JY^\top YJ = -\frac{1}{2}J\Delta J$ (Lemma A.1). Thus by Weyl's inequality, the fact that $\|\cdot\| \leq \|\cdot\|_{S_p}$ for all $p$, and (14),

$$\begin{aligned} \sigma_d\left(-\frac{1}{2}J\Lambda J\right) &\geq \sigma_d\left(-\frac{1}{2}J\Delta J\right) - \frac{1}{2}\|J(\Lambda - \Delta)J\|_{S_p} \\ &= \sigma_d^2(Y) - \frac{1}{2}\|J(\Lambda - \Delta)J\|_{S_p} \\ &> 0. \end{aligned}$$

Consequently, $-\frac{1}{2}J\Lambda J$ has rank at least $d$, so if $Z$ contains the columns of the MDS embedding corresponding to $\Lambda$, then $Z^\top Z$ is the best rank-$d$ approximation of $-\frac{1}{2}J\Lambda J$ (by construction). It follows from Mirsky's inequality and the facts that $-\frac{1}{2}J\Delta J = Y^\top Y$ and $\text{rank}(Y) = d$ that

$$\left\|Z^\top Z + \frac{1}{2}J\Lambda J\right\|_{S_p} \leq \left\|\frac{1}{2}J(\Lambda - \Delta)J\right\|_{S_p}. \tag{15}$$

Combining (14) and (15), we have

$$\begin{aligned} \varepsilon^2 := \|Z^\top Z - Y^\top Y\|_{S_p} \leq \left\|Z^\top Z + \frac{1}{2}J\Lambda J\right\|_{S_p} + \left\|\frac{1}{2}J(\Lambda - \Delta)J\right\|_{S_p} &\leq \|J(\Lambda - \Delta)J\|_{S_p} \\ &\leq \|\Lambda - \Delta\|_{S_p}. \end{aligned}$$

Thus, $\|Y^\dagger\|\varepsilon \leq \|Y^\dagger\|\|\Lambda - \Delta\|_{S_p}^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}}$, so we may apply the final bound of Theorem 3.1 to yield the conclusion. □

### 3.3. Main theorem

The following theorem shows the quality of an MDS embedding of a discrete subset of $W_2(\mathbb{R}^n)$ when approximations of the pairwise $W_2(\mathbb{R}^n)$ distances are used (via, for example, LOT approximations, Sinkhorn regularization, or other approximation techniques). The embedding quality is understood in two parts: first, how far away the set is from a subset of $W_2(\mathbb{R}^n)$ that is isometric to $\mathbb{R}^d$, and second, how good an approximation to the Wasserstein distances one utilizes in MDS. The second source of error can always be made arbitrarily small given sufficient computation time or judicious choice of parameters (as in Sinkhorn, for example). However, the first source of error arises from the geometry of the set of points, and may or may not be small.

Note that using Corollary 3.2 outright would require computing a proxy distance matrix and applying MDS; however, to make Algorithm 2 computationally efficient, we instead compute the truncated SVD of the centered transport maps rather than on the distance matrix between the transport maps. These are the same in theory, but allow for significantly less computation in practice. Below, we state our main theorem, which is stated in terms of the output of MDS on an estimation of Wasserstein distances between measures; but we stress that we are able to easily transfer the bounds to the output of Algorithm 2, which does not require any distance matrix computation.

**Theorem 3.3.** *Let $\{\mu_i\}_{i=1}^N \subset W_2(\mathbb{R}^n)$. Suppose $\mathscr{W} \subset W_2(\mathbb{R}^n)$ is a subset of Wasserstein space that is isometric to a subset of Euclidean space $\Xi \subset \mathbb{R}^d$, and $\{\nu_i\}_{i=1}^N \subset \mathscr{W}$ and $\{y_i\} \subset \Xi$ are such that $|y_i - y_j| = W_2(\nu_i, \nu_j)$. Let $\Delta_{ij} := W_2(\nu_i, \nu_j)^2$, $\Gamma_{ij} := W_2(\mu_i, \mu_j)^2$, and $\Lambda_{ij} := \lambda_{ij}^2$ for some $\lambda_{ij} \in \mathbb{R}$. Let $\{z_i\}_{i=1}^N$ be the output of MDS (Algorithm 1) with input $\Lambda$.*

*If $|W_2(\mu_i, \mu_j)^2 - W_2(\nu_i, \nu_j)^2| \leq \tau_1$ and $|W_2(\mu_i, \mu_j)^2 - \lambda_{ij}^2| \leq \tau_2$ for some $\tau_1$ and $\tau_2$, and*

$$\|Y^\dagger\|\sqrt{N}\left(\tau_1 + \tau_2\right)^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}}, \tag{16}$$

*then $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$ satisfies*

$$\min_{Q \in \mathscr{O}(d)} \|Z - QY\|_F \leq (1 + \sqrt{2})\|Y^\dagger\|N\left(\tau_1 + \tau_2\right).$$

**Proof.** Note that

$$\|\Lambda - \Delta\|_F \leq \|\Gamma - \Delta\|_F + \|\Lambda - \Gamma\|_F \leq N(\tau_1 + \tau_2).$$

Consequently, (16) allows us to apply Corollary 3.2 to yield the conclusion. □

Specializing Theorem 3.3 to the case of Algorithm 2 yields the following corollary, which shows that the truncated SVD of the centered LOT transport matrix $T$ is equivalent to the output $z_i$ of MDS in Theorem 3.3.

**Corollary 3.4.** *Invoke the notations and assumptions of Theorem 3.3. Choose a reference measure $\sigma \in W_2(\mathbb{R}^n)$ and compute all transport maps $T_\sigma^{\mu_i}$. Let $T$ be the transport map matrix created by centering and column-stacking the transport maps $T_\sigma^{\mu_i}$ as in Algorithm 2. Let $U_d\Sigma_d V_d^\top$ be the truncated SVD of $T$, and let $z_i = V_d\Sigma_d(i, :)$ for $1 \leq i \leq N$ (i.e., $z_i$ is the output of Algorithm 2). Then $z_i$ is the output of MDS with $\Lambda$ being the empirical linearized Wasserstein-2 distance (12), and if (16) holds, then*

$$\min_{Q \in \mathscr{O}(d)} \|Z - QY\|_F \leq (1 + \sqrt{2})\|Y^\dagger\|N\left(\tau_1 + \tau_2\right).$$

**Proof.** Since $T$ is centered, Lemma A.1 implies that $T^\top T = JT^\top TJ = -\frac{1}{2}J\Lambda J$. Consequently, if $-\frac{1}{2}J\Lambda J = V\Sigma^2 V^\top = T^\top T$, then $T$ has truncated SVD $T_d = U_d\Sigma_d V_d^\top$, and therefore $z_i = V_d\Sigma_d(i, :)$ arises from the truncated SVD of $T$ and is also the output of MDS with input $\Lambda$. The conclusion follows by direct application of Theorem 3.3. □

In the rest of the paper, we will discuss how various LOT approximations to Wasserstein distances affect the value of the bound $\tau_2$ appearing in Theorem 3.3 and Corollary 3.4. In particular, we get different values of $\tau_2$ when we have compactly supported target measures (as in Theorem 4.2 for linear programming estimators and Theorem 4.8 for Sinkhorn estimators) and non-compactly supported target measures (as in Theorem 5.4 for linear programming estimators and Theorem 5.5 for Sinkhorn estimators).

## 4. Bounds for compactly supported target measures

To capture the bound $\tau_2$ of Theorem 3.3, we turn our attention to approximating the pairwise square-distance matrix $\left[W_2^2(\mu_i, \mu_j)\right]_{i,j=1}^N$ appearing in the theorem statement with the finite sample, discretized LOT distance matrix that comes from

differences between transport maps to a fixed reference, a finite sampling of $\mu_i$, and a discretization of the reference distribution $\sigma$. In particular, the main approximation argument consists of the following triangle inequality:

$$\begin{aligned}\left|W_2(\mu_1,\mu_2)^2 - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2\right| \le & \underbrace{\left|W_2(\mu_1,\mu_2)^2 - W^{\mathrm{LOT}}_{2,\sigma}(\mu_1,\mu_2)^2\right|}_{\text{LOT error}} \\ & + \underbrace{\left|W^{\mathrm{LOT}}_{2,\sigma}(\mu_1,\mu_2)^2 - W^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2\right|}_{\text{finite sample and optimization error}} \\ & + \underbrace{\left|W^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2 - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2\right|}_{\text{discretized } \sigma \text{ sampling error}}.\end{aligned}$$

There are four sources of error between these two distance matrices:

(1) approximating the Wasserstein distance with LOT distance,
(2) approximating LOT embeddings between $\mu_i$ and $\mu_j$ with the barycentric approximations computed using finite samples $\widehat{\mu_i}$ and $\widehat{\mu_j}$,
(3) approximating the integral with respect to the reference measure $\sigma$ by the discretized sampling $\widehat{\sigma}$, and
(4) optimization error in approximating the optimal transport map.

The errors from (1) and (3) are handled in Appendix B whilst the error from (2) gives us the main theorems of this section. Error from (4) is also implicitly considered by handling error from (2) since the optimization error for using a linear programming optimizer versus a Sinkhorn optimizer is seen in the error bounds of Theorem 4.2 and Theorem 4.8. We deal with each error separately and chain the bounds together at the end.

Before dealing with any of the details of the proofs, we need the following assumptions on $\sigma$, $\mu$, and $\mathscr{H}$:

**Assumption 4.1.** Consider the following conditions on $\sigma$, $\mu$, and $\mathscr{H}$

(i) $\sigma \in \mathscr{P}_{ac}(\Omega)$ for a compact convex set $\Omega \subseteq B(0,R) \subset \mathbb{R}^n$ with probability density $f_\sigma$ bounded above and below by positive constants.
(ii) $\mu$ has finite $p$-th moment with bound $M_p$ with $p > n$ and $p \ge 4$.
(iii) There exist $a, A > 0$ such that every $h \in \mathscr{H}$ satisfies $a\|x\| \le \|h(x)\| \le A\|x\|$ for every $x \in \mathrm{supp}(\mu)$.
(iv) $\mathscr{H}$ is compact with respect to the $L^2(\mu)$-norm and $\varepsilon$-compatible with respect to $\sigma, \mu \in W_2(\mathbb{R}^n)$. Moreover, $\sup_{h,h' \in \mathscr{H}} \|h - h'\|_\mu \le M$.
(v) $\mu_i \sim \mathscr{H}_\sharp \mu$ i.i.d.

These assumptions ensure that $\varepsilon$-compatible transformations are also "$\varepsilon$-isometric" as shown in Theorem A.4. Moreover, note that $M \ne M_p$ as $M_p$ is important for extending theory to non-compactly supported measures while $M$ handles bounds associated with the complexity of the function class $\mathscr{H}$ of pushforwards.

### 4.1. Using the linear program to compute transport maps

In this subsection, we assume that the classical linear program is used to compute the optimal transport maps from $\widehat{\mu_i}$ to the reference (and its discretization).

**Theorem 4.2.** *Let $\delta > 0$. Along with Assumption 4.1 and that $\mu \in \mathscr{P}_{ac}(\Omega)$ for the $\Omega$ in Assumption 4.1 with simply connected support, assume that*

*(i) $T_\sigma^{\mu_i}$ is Lipschitz.*
*(ii) We estimate $\mu_i$ with an empirical measure $\widehat{\mu_i}$ using $k$ samples and discretize $\sigma$ with $m$ samples. Let our estimator be given by (10) with $\gamma$ solved using linear programming.*

*Then with probability at least $1 - \delta$,*

$$\left|W_2(\mu_1,\mu_2)^2 - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu_1},\widehat{\mu_2};\gamma_{LP})^2\right| \le (M + 2AR)\left(C\varepsilon^{\frac{p}{6p+16n}} + 2O_p(r_n^{(k)}\log(1+k)^{t_n}) + R\sqrt{\frac{2\log(2/\delta)}{m}}\right). \tag{17}$$

*Here $C$ is the constant from Theorem A.4 depending on $n, p, \Omega, M_p$, the constants $a$ and $A$ come from Assumption 4.1 (iii), and*

$$r_n^{(k)} = \begin{cases} 2k^{-1/2} & n = 2,3 \\ 2k^{-1/2}\log(1+k) & n = 4 \\ 2k^{-2/n} & n \geq 5 \end{cases}, \quad t_n = \begin{cases} \frac{5}{2} & n \leq 4 \\ 2(1+n^{-1}) & n > 4 \end{cases},$$

*so that $r_n^{(k)}$ is on the order of $k^{-1/n}$ and $t_n$ is on the order of $2(1+n^{-1})$. In this case, $\tau_2$ of Corollary 3.4 is bounded above by the right-hand side of (17).*

**Proof.** Note that the transport plan that we are using for the following proof is $\gamma_{LP}$. Henceforth, we will suppress $\gamma_{LP}$ from the terms $\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma_{LP})$ and $T_\sigma^{\widehat{\mu_j}}(\cdot;\gamma_{LP})$ for simplicity.

Since $|x^2-y^2| = |x+y||x-y|$, we need to bound both

(a) $\left|W_2(\mu_1,\mu_2) + \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})\right|$,
(b) $\left|W_2(\mu_1,\mu_2) - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})\right|$.

*We start with* (a): Since both $\mu_1$ and $\mu_2$ are pushforwards of a fixed template distribution $\mu$, we know that $\mu_i = h_{i\sharp}\mu$, where by [3, Eq. 2.1] and our assumptions, it follows that

$$W_2(\mu_1,\mu_2) = W_2(h_{1\sharp}\mu, h_{2\sharp}\mu) \leq \|h_1 - h_2\|_\mu \leq M.$$

Moreover, since $\mu$ is compactly supported for $\Omega \subseteq B(0,R)$ and $\mu_i = (h_i)_\sharp\mu$ with $h_i \in \mathscr{H}$ and $a\|x\| \leq \|h_i(x)\| \leq A\|x\|$, we know that $\mu_i$ is compactly supported with $\mathrm{supp}(\mu_i) \subseteq B(0,AR)$ for all $i$. This implies that

$$\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2}) = \left(\frac{1}{m}\sum_{j=1}^{m} \underbrace{|T_\sigma^{\widehat{\mu_1}}(X_j) - T_\sigma^{\widehat{\mu_2}}(X_j)|^2}_{\leq (2AR)^2}\right)^{1/2} \leq 2AR.$$

Putting these estimates together, we have

$$\left|W_2(\mu_1,\mu_2) + \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})\right| \leq M + 2AR.$$

*We continue with* (b): From the triangle inequality we get

$$\begin{aligned}\left|W_2(\mu_1,\mu_2) - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})\right| \leq{}& \left|W_2(\mu_1,\mu_2) - W_{2,\sigma}^{\mathrm{LOT}}(\mu_1,\mu_2)\right| + \left|W_{2,\sigma}^{\mathrm{LOT}}(\mu_1,\mu_2) - W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})\right| \\ &+ \left|W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2}) - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})\right|.\end{aligned}$$

We now bound these three parts individually:

a) By Assumption 4.1, we can use $\varepsilon$-compatibility of $\mathscr{H}$ in Theorem A.4 to get that

$$\left|W_2(\mu_1,\mu_2) - W_{2,\sigma}^{\mathrm{LOT}}(\mu_1,\mu_2)\right| \leq C\varepsilon^{\frac{p}{6p+16n}},$$

where $C$ is from Theorem A.4.

b) For the second term, we again assume that any transport maps involving discrete measures are obtained from the linear program. In particular, we see that

$$\begin{aligned}W_{2,\sigma}^{\mathrm{LOT}}(\mu_1,\mu_2) &= \|T_\sigma^{\mu_1} - T_\sigma^{\mu_2}\|_\sigma \\ &\leq \|T_\sigma^{\mu_1} - T_\sigma^{\widehat{\mu}_1}\|_\sigma + \|T_\sigma^{\widehat{\mu}_1} - T_\sigma^{\widehat{\mu}_2}\|_\sigma + \|T_\sigma^{\widehat{\mu}_2} - T_\sigma^{\mu_2}\|_\sigma \\ &= \|T_\sigma^{\mu_1} - T_\sigma^{\widehat{\mu}_1}\|_\sigma + \|T_\sigma^{\widehat{\mu}_2} - T_\sigma^{\mu_2}\|_\sigma + W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2}).\end{aligned}$$

Notice that we can equivalently apply the triangle inequalities starting from $W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})$ to get

$$\begin{aligned}W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2}) &= \|T_\sigma^{\widehat{\mu}_1} - T_\sigma^{\widehat{\mu}_2}\|_\sigma \\ &\leq \|T_\sigma^{\mu_1} - T_\sigma^{\mu_2}\|_\sigma + \|T_\sigma^{\mu_1} - T_\sigma^{\widehat{\mu}_1}\|_\sigma + \|T_\sigma^{\widehat{\mu}_2} - T_\sigma^{\mu_2}\|_\sigma \\ &= \|T_\sigma^{\mu_1} - T_\sigma^{\widehat{\mu}_1}\|_\sigma + \|T_\sigma^{\widehat{\mu}_2} - T_\sigma^{\mu_2}\|_\sigma + W_{2,\sigma}^{\mathrm{LOT}}(\mu_1,\mu_2).\end{aligned}$$

Note that Assumption 4.1(i) implies that for every $t > 0$ and $\alpha > 0$, $\mathbb{E}_\sigma[\exp(t\|x\|^\alpha)] < \infty$ as $\sigma$ is compactly supported. Because $T_\sigma^{\mu_i}$ is $L$-Lipschitz, this allows us to use Theorem B.1 and optimize the exponents $t_{n,\alpha}$ over $\alpha > 0$ to conclude that

$$|W_{2,\sigma}^{\mathrm{LOT}}(\mu_1,\mu_2) - W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})| \leq \|T_\sigma^{\mu_1} - T_\sigma^{\widehat{\mu}_1}\|_\sigma + \|T_\sigma^{\widehat{\mu}_2} - T_\sigma^{\mu_2}\|_\sigma \leq 2\,O_p(r_n^{(k)}\log(1+k)^{t_n}).$$

c) From Theorem B.3 we know that with probability at least $1-\delta$,

$$\left|W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2}) - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2})\right| \leq R\sqrt{\frac{2\log(2/\delta)}{m}}.$$

Putting these bounds together yields the result. □

**Remark 4.3.** There exist conditions on $\mathscr{H}$ that ensure $T_\sigma^{\mu_i}$ is Lipschitz, which is the setting needed to apply Theorem B.1. We cover these conditions in Section 6. When $\mathscr{H}$ is not exactly compatible, we have Theorem 6.1 which requires more technical assumptions. For the case when $\mathscr{H}$ is exactly compatible, we have Theorem 6.2 which only requires that $\mathscr{H}$ is comprised of Lipschitz functions. Similar results in Section 6 hold for using Sinkhorn transport plans.

### 4.2. Using entropic regularization (Sinkhorn) to compute transport maps

Although [18] gives estimation rates in terms of a transport map constructed from solving the linear program associated to the optimal transport problem, solving the regularized optimal transport problem (9) and using the barycentric projection map (13) is much faster. For this section, we will assume that the target and reference measures are discretized with the same number of samples $k$.

**Remark 4.4.** Since we can choose $\sigma$ as well as the sample size for $\widehat{\sigma}$, we can allow $k=m$ in this case. We believe, however, that choosing a larger sample size for $\sigma$ than $\mu_i$ (i.e. $m>k$) will result in better approximation.

For the following results, we make use of the following quantity:

**Definition 4.5.** Consider the Wasserstein geodesic between $\sigma=\mu_0$ and $\mu=\mu_1$ with $\mu_t$ being the measure on the geodesic for $t\in(0,1)$. Let $f(t,x)$ be the density corresponding to $\mu_t$. Then the integrated Fisher information along the Wasserstein geodesic between $\sigma$ and $\mu$ is given by

$$I_0(\sigma,\mu) = \int_0^1 \int_{\mathbb{R}^n} \left\|\nabla_x \log f(t,x)\right\|_2^2 f(t,x)\,dx\,dt.$$

Moreover, recall that the convex conjugate of a function $\phi:\mathbb{R}^n\to\mathbb{R}$ is given by

$$\phi^*(x^*) = \sup_{x\in\mathbb{R}^n} {x^*}^{\top}x - \phi(x),$$

see, e.g., [4, p. 45]. Now by using Theorem 3 from [35], we will show that under suitable conditions the entropic map $T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta)$ is close to $T_\sigma^{\mu_i}$.

**Theorem 4.6** *([35, Theorem 3]). Assume that*

(A1) $\sigma,\mu_i\in\mathscr{P}_{ac}(\Omega)$ *for a compact set* $\Omega\subset\mathbb{R}^n$ *with densities satisfying* $f_\sigma, f_{\mu_i}\leq B$ *and* $f_{\mu_i}\geq b>0$ *for all* $x\in\Omega$.
(A2) $\phi\in C^2(\Omega)$ *and* $\phi^*\in C^{\alpha+1}(\Omega)$ *for* $\alpha>1$, *where* $\phi^*$ *denotes the convex conjugate of* $\phi$.
(A3) $T_\sigma^{\mu_i}=\nabla\phi$ *with* $mI\leq\nabla^2\phi(x)\leq LI$ *for* $m,L>0$ *for all* $x\in\Omega$.

*Then the entropic map* $T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta)$ *from* $\widehat{\sigma}$ *to* $\widehat{\mu}_i$ *with regularization parameter* $\beta\asymp k^{-\frac{1}{n'+\widetilde{\alpha}+1}}$ *satisfies*

$$\mathbb{E}\left\|T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta) - T_\sigma^{\mu_i}\right\|_\sigma^2 \leq \left(1+I_0(\sigma,\mu_i)\right)k^{-\frac{\widetilde{\alpha}+1}{2n'+\widetilde{\alpha}+1}}\log k,$$

*where* $n'=2\lceil n/2\rceil$, $\widetilde{\alpha}=\alpha\wedge 3$, $k$ *is the sample size for both* $\widehat{\sigma}$ *and* $\widehat{\mu}_i$, *and* $I_0(\sigma,\mu_i)$ *is the integrated Fisher information along the Wasserstein geodesic between* $\sigma$ *and* $\mu_i$.

Given the sample size $k$ for both $\widehat{\sigma}$ and $\widehat{\mu}_i$, if we let

$$Z_k = \left\|T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta) - T_\sigma^{\mu_i}\right\|_\sigma,$$

then by Jensen's inequality (for concave functions) and Theorem 4.6 we have that

$$\mathbb{E}[Z_k] \leq \mathbb{E}\left[Z_k^2\right]^{1/2} \leq \sqrt{\left(1+I_0(\sigma,\mu_i)\right)k^{-\frac{\widetilde{\alpha}+1}{2n'+\widetilde{\alpha}+1}}\log k}$$

$$= \sqrt{\log(k)(1 + I_0(\sigma, \mu_i))}k^{-\frac{\widetilde{\alpha}+1}{2(2n'+\widetilde{\alpha}+1)}}.$$

Now using Markov's inequality, we easily have the following corollary.

**Corollary 4.7.** *Assume that $\sigma$ and $\mu_i$ satisfy (A1)–(A3) of Theorem 4.6 and let $\delta > 0$. Then with probability at least $1 - \delta$, we have that*

$$\left\| T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta) - T_\sigma^{\mu_i} \right\|_\sigma \leq \frac{1}{\delta}\sqrt{\log(k)\big(1 + I_0(\sigma, \mu_i)\big)}k^{-\frac{\widetilde{\alpha}+1}{2(2n'+\widetilde{\alpha}+1)}}.$$

Now we can approximate $T_\sigma^{\mu_i}$ with the entropic map that is derived from using Sinkhorn's algorithm. Although the barycentric projection map and entropic map approximations have similar rates of convergence, the entropic map is computationally faster at the cost of more stringent assumptions in the theorem. The main difference in assumptions below is the addition of (A1)–(A3) from Theorem 4.6 and the asymptotic bound on the regularization parameter $\beta$ used in the entropic regularization.

**Theorem 4.8.** *Let $\delta > 0$. Along with Assumption 4.1 and $\mu \in \mathscr{P}_{ac}(\Omega)$ for the $\Omega$ in Assumption 4.1, assume that*

*(i) $\sigma$ and $\mu_i$ satisfy assumptions (A1)–(A3) from Theorem 4.6 for all $i$. Note that (A1), regularity of $\phi$ in (A2), and the upper bound of (A3) are satisfied under the conditions of Caffarelli's regularity theorem.*

*(ii) Given empirical distributions $\widehat{\sigma}$ and $\widehat{\mu}_i$ both with $k$ sample size, assume that we have associated entropic potentials $(f_{\beta,k}, g_{\beta,k})$, where $\beta \asymp k^{-\frac{1}{n'+\widetilde{\alpha}+1}}$ and $n'$ and $\widetilde{\alpha}$ are defined in Theorem 3 from [35]. Assume our estimator is $T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta)$ given by (13).*

*Then with probability at least $1 - \delta$,*

$$\left| W_2(\mu_i, \mu_j)^2 - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu}_i, \widehat{\mu}_j; \gamma_\beta)^2 \right| \leq (M + 2AR)\left( C\varepsilon^{\frac{p}{6p+16n}} + \frac{2}{\delta}\sqrt{\log(k)(1 + I_0(\sigma, \mu_i))}k^{-\frac{\widetilde{\alpha}+1}{2(2n'+\widetilde{\alpha}+1)}} + R\sqrt{\frac{2\log(2/\delta)}{k}}\right),$$

*where $C$ is from Theorem A.4 and $I_0(\sigma, \mu_i)$ is defined in Theorem 4.6. In this case, $\tau_2$ in Corollary 3.4 is bounded above by the right-hand side of the inequality above.*

**Proof.** Note that the transport plan that we are using for the following proof is $\gamma_\beta$. Henceforth, we will suppress $\gamma_\beta$ from the notation $\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1}, \widehat{\mu_2}; \gamma_\beta)$ for simplicity.

Using the same reasoning as in Theorem 4.2, we find that

$$\left( W_2(\mu_i, \mu_j) + \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu}_i, \widehat{\mu}_j) \right) \leq M + 2AR.$$

Similar to the proof of Theorem 4.2, we bound

$$\begin{aligned}\left| W_2(\mu_i, \mu_j) - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu}_i, \widehat{\mu}_j) \right| \leq{}& \left| W_2(\mu_i, \mu_j) - \left\| T_\sigma^{\mu_i} - T_\sigma^{\mu_j} \right\|_\sigma \right| \\ &+ \left\| T_\sigma^{\mu_i} - T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta) \right\|_\sigma + \left\| T_\sigma^{\mu_j} - T_{\widehat{\sigma}}^{\widehat{\mu}_j}(\,\cdot\,;\gamma_\beta) \right\|_\sigma \\ &+ \left| \left\| T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta) - T_{\widehat{\sigma}}^{\widehat{\mu}_j}(\,\cdot\,;\gamma_\beta) \right\|_\sigma - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu}_i, \widehat{\mu}_j) \right|.\end{aligned}$$

The first and is bounded the same way as in the proof of Theorem 4.2 above. For the last term, we apply Corollary B.4. Since assumption (i) of Assumption 4.1, implies assumption (A1) of Theorem 4.6, we get that with probability at least $1 - \delta$

$$\left\| T_\sigma^{\mu_\ell} - T_{\widehat{\sigma}}^{\widehat{\mu_\ell}}(\,\cdot\,;\gamma_\beta) \right\|_\sigma \leq \frac{1}{\delta}\sqrt{\log(k)(1 + I_0(\sigma, \mu_\ell))}k^{-\frac{\widetilde{\alpha}+1}{2(2n'+\widetilde{\alpha}+1)}}$$

for $\ell = i$ and $\ell = j$. Putting the bounds together, we get the result. □

Using Theorem 4.2 and Theorem 4.8, we see that as long as $\mu_i$ are $\varepsilon$-compatible push-forwards of $\mu$ and the number of samples used in the empirical distribution is large enough, then our LOT distance is a computationally efficient and a tractable approximation for the Wasserstein distance and the distortion of the LOT Wassmap embedding of $\{\mu_i\}$ is small with high probability.

## 5. Bounds for non-compactly supported target measures

In the last section, we saw that for compactly supported $\mu_i \sim \mathscr{H}_\sharp\mu$ (as well as a few other conditions), either the barycentric estimator $T_\sigma^{\widehat{\mu}_i}(\,\cdot\,;\gamma_{LP})$ or the entropic estimator $T_\sigma^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta)$ will allow for fast yet accurate approximation of the pairwise Wasserstein distances $W_2(\mu_i, \mu_j)$, which in turn allows for fast, accurate LOT approximation to the Wassmap embedding [21] via Algorithm 2. In this section, we show that we can adapt Theorem 4.2 and Theorem 4.8 to non-compactly supported measures as long as we can

approximate the non-compactly supported measure with a compactly supported and absolutely continuous measure. To this end, we use the main theorem of [19].

**Theorem 5.1** *([19]). Let $\Omega$ be a compact convex set and let $\sigma$ be a probability density on $\Omega$, bounded from above and below by positive constants. Let $p > n$ and $p \geq 4$. Assume that $\mu, \nu \in W_2(\mathbb{R}^n)$ have bounded $p$-th moment, and $\max(M_p(\mu), M_p(\nu)) \leq M_p < \infty$. Then*

$$\|T_\sigma^\mu - T_\sigma^\nu\|_\sigma \leq C_{n,p,\Omega,M_p} W_1(\mu,\nu)^{\frac{p}{6p+16n}}.$$

To achieve our purposes, we will assume that $\mu$ is a non-compactly supported measure that has a suitable tail decay rate, and then show that there exists a compactly supported absolutely continuous $\widetilde{\mu}$ that approximates $\mu$ well (i.e., $W_1(\mu,\widetilde{\mu}) < \eta$.). The particular compactly supported $\widetilde{\mu}$ will be formed by a pushforward that is the identity on $B(0,R)$ and applies a modified version of the $\frac{x}{1+\|x\|}$ map near the boundary of $B(0,R)$. We achieve this in the following lemma.

**Lemma 5.2.** *Fix $\eta > 0$, and let $\sigma$ satisfy the assumptions of Theorem 5.1. Moreover, let $\mu \in W_2(\mathbb{R}^n)$ with density $f_\mu$ have a bounded $p$-th moment for some $p > n$ and $p \geq 4$. Finally, assume that there exists some $R > 0$ such that for every $x \notin B(0,R)$, we have*

$$f_\mu(x) < \left(\frac{\eta}{\widetilde{C}_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{\|x\|^{n+2}},$$

*where $\widetilde{C}_{n,p,\Omega,M_p}$ denotes a combination of the constant from Theorem 5.1 and a constant from integrating over concentric $n$-spheres. Then there exists a compactly supported absolutely continuous measure $\widetilde{\mu}$ such that*

$$\|T_\sigma^\mu - T_\sigma^{\widetilde{\mu}}\|_\sigma < \eta.$$

For many of the results before, we require that our compactly supported measures have a density bounded away from 0 and $\infty$. The next lemma will be useful in establishing conditions on $\mathscr{H}$ and $\mu$ so that our truncated measure $\widetilde{\mu}$ has a density that is bounded away from 0. In particular, we construct $\widetilde{\mu}$ by constructing a density that is compactly supported and remains bounded away from 0 and $\infty$ rather than using a pushforward. The proof uses that a pushforward must exist between $\mu$ and $\widetilde{\mu}$.

**Lemma 5.3.** *Let $\sigma$ satisfy the assumptions of Theorem 5.1 and let $\mu \in W_2(\mathbb{R}^n)$ with density $f_\mu \leq C < \infty$ have a bounded $p$-th moment for some $p > n$ and $p \geq 4$. Moreover, assume that there exists some $R > 0$ and $\eta > 0$ such that for $x \in B(0,R)$, we have $f_\mu(x) \geq c > 0$; and for every $x \notin B(0,R)$, we have*

$$f_\mu(x) \leq \left(\frac{\eta}{\widetilde{C}'_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{\|x\|^{n+2}},$$

*where $\widetilde{C}'_{n,p,\Omega,M_p}$ comes from combining the constant from Theorem 5.1, a constant from integrating over concentric $n$-spheres, and another constant from our approximation method. Then there exists a compactly supported, absolutely continuous measure $\widetilde{\mu}$ with density $0 < c \leq b \leq f_{\widetilde{\mu}} \leq B < \infty$ such that*

$$\|T_\sigma^\mu - T_\sigma^{\widetilde{\mu}}\|_\sigma < \eta.$$

The proofs of both Lemma 5.2 and Lemma 5.3 are located in Appendix C. Notice that the condition on the density $f_\mu$ ensures that $W_1(\mu,\widetilde{\mu}) < \eta$ for the two different truncated measures $\widetilde{\mu}$. With these two lemmas above, we obtain the following theorems. Note that Theorem 5.4 replaces the assumption that $\mu$ is compactly supported with one of polynomial (in the ambient dimension) tail decay; while the second assumption below is the same as Theorem 4.2, the final assumption differs from that of Theorem 4.2 by requiring the discretizations of $\sigma$ and $\mu_i$ to have the same sample size to apply the lemmas above.

**Theorem 5.4.** *Let $\delta > 0$. Along with Assumption 4.1, assume that*

*(i) Every $\mu_i$ has bounded $p$-th moment for some $p > n$ and $p \geq 4$. Moreover, assume that for all $i$, there exists some $R > 0$ such that for every $x \notin B(0,AR)$, we have*

$$f_{\mu_i} < \left(\frac{\eta}{C}\right)^{\frac{6p+16n}{p}} \frac{1}{\|x\|^{n+2}},$$

*where $C = \widetilde{C}_{n,p,\Omega,M_p}$ or $C = \widetilde{C}'_{n,p,\Omega,M_p}$ depending on if we use the truncated measure $\widetilde{\mu}_i$ to be from Lemma 5.2 or Lemma 5.3 so that $W_1(\mu_i,\widetilde{\mu}_i) < \eta$.*

*(ii) $T_\sigma^{\widetilde{\mu}_i}$ is $L$-Lipschitz (this happens, e.g., if $\sigma$ and $\widetilde{\mu}_i$ are both compactly supported).*

*(iii) Given empirical distributions $\widehat{\sigma}$ and $\widehat{\mu}_i$ with $\operatorname{supp}(\widehat{\mu}_i) \subseteq B(0, AR)$ and sample sizes $m$ and $k$, respectively, let our estimator be the barycentric estimator (10), with $\gamma_{LP}$.*

*Then with probability at least $1-\delta$,*

$$\left|W_2(\mu_i,\mu_j)^2 - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu}_i,\widehat{\mu}_j;\gamma_{LP})^2\right| \leq (M+2AR)\left(C\varepsilon^{\frac{p}{6p+16n}} + 2\eta + 2O_p(r_n^{(k)}\log(1+k)^{t_{n,\alpha}}) + R\sqrt{\frac{2\log(2/\delta)}{m}}\right),$$

*where $r_n^{(k)}$ and $t_{n,\alpha}$ are defined in Theorem 4.2 and $C$ is a constant coming from Theorem A.4. In this case, $\tau_2$ of Corollary 3.4 is bounded above by the right-hand side of the inequality above.*

Similarly for the entropic map case we have the following. Note that the primary difference in assumption between Theorem 5.5 and Theorem 5.4 is the addition of (A1)–(A3) from Theorem 4.6 and the asymptotic assumption on the regularization parameter for the entropic map. The assumptions (i) and (ii) below are essentially the same as those of Theorem 4.8, but with $\widehat{\mu}_i$ replaced with $\widetilde{\mu}_i$ arising from Theorem 5.4, whereas the additional assumptions below are that $\mu_i$ have decaying tails as opposed to being compactly supported.

**Theorem 5.5.** *Let $\delta > 0$. Along with Assumption 4.1 and (i) of Theorem 5.4, assume that*

*(i) $\sigma$ and $\widetilde{\mu}_i$ satisfy assumptions (A1)–(A3) in 4.6 for all $i$, where $\widetilde{\mu}_i$ is the truncated measure from Theorem 5.4.*
*(ii) Given empirical distributions $\widehat{\sigma}$ and $\widehat{\mu}_i$ with $\operatorname{supp}(\widehat{\mu}_i) \subseteq B(0, AR)$ and sample size $k$ for both, assume that we have associated entropic potentials $(f_{\beta,k}, g_{\beta,k})$, where $\beta \asymp k^{-\frac{1}{n'+\widetilde{\alpha}+1}}$ and $n'$ and $\widetilde{\alpha}$ are defined in Theorem 4.6. Moreover, assume our estimator is given by (13).*

*Then with probability at least $1-\delta$,*

$$\left|W_2(\mu_i,\mu_j)^2 - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu}_i,\widehat{\mu}_j)^2\right| \leq (M+2AR)\left(C\varepsilon^{\frac{p}{6p+16n}} + 2\eta + \frac{2}{\delta}\sqrt{\log(k)(1+I_0(\sigma,\mu_i))}k^{-\frac{\widetilde{\alpha}+1}{2(2n'+\widetilde{\alpha}+1)}} + R\sqrt{\frac{2\log(2/\delta)}{k}}\right),$$

*where $I_0(\sigma,\mu_i)$ is defined in Theorem 4.6 and $C$ is a constant from Theorem A.4. In this case, $\tau_2$ of Corollary 3.4 is bounded above by the right-hand side of the inequality above.*

The following is a proof for both theorems above.

**Proof of Theorems 5.4 and 5.5.** In the following, we let $T_\sigma^{\widehat{\mu}_i}$ denote the optimal transport map estimator that we are considering (either the barycentric estimator with $\gamma_{LP}$ or the entropic estimator with $\gamma_\beta$) since the same proof works for both cases. The only difference in the compactly supported case and these theorems is that our approximation now becomes

$$\begin{aligned}\left|W_2(\mu_i,\mu_j) - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu}_i,\widehat{\mu}_j)\right| \leq{}& \left|W_2(\mu_i,\mu_j) - \|T_\sigma^{\mu_i} - T_\sigma^{\mu_j}\|_\sigma\right| \\ &+ \left|\|T_\sigma^{\mu_i} - T_\sigma^{\mu_j}\|_\sigma - \|T_\sigma^{\widetilde{\mu}_i} - T_\sigma^{\widetilde{\mu}_j}\|_\sigma\right| \\ &+ \left|\|T_\sigma^{\widetilde{\mu}_i} - T_\sigma^{\widetilde{\mu}_j}\|_\sigma - \|T_\sigma^{\widehat{\mu}_i} - T_\sigma^{\widehat{\mu}_j}\|_\sigma\right| \\ &+ \left|\|T_\sigma^{\widehat{\mu}_i} - T_\sigma^{\widehat{\mu}_j}\|_\sigma - \widehat{W}^{\mathrm{LOT}}_{2,\sigma}(\widehat{\mu}_i,\widehat{\mu}_j)\right|,\end{aligned}$$

where $\widetilde{\mu}_i$ is defined as in the theorem statement and $\widehat{\mu}_i$ denotes the empirical measure of $\mu_i$. For the following consider the truncation radius of $\widetilde{\mu}$ to be $AR$ rather than $R$ as was the case in Lemma 5.2 and Lemma 5.3. We assume that $\operatorname{supp}(\widehat{\mu}_i) \subseteq B(0, AR)$; thus, let us assume that we sample from $\mu_i$ conditioned that we restrict to $B(0, AR)$. Since $\widetilde{\mu}_i|_{B(0,AR)} = \mu_i|_{B(0,AR)}$, we see that $\widehat{\mu}_i$ can equivalently be thought of as being sampled from $\widetilde{\mu}_i$ rather than $\mu_i$ conditioned that $\widehat{\mu}_i \subseteq B(0, AR)$. This means that the same bounds as before hold for most of the terms, while additionally,

$$\left|\|T_\sigma^{\mu_i} - T_\sigma^{\mu_j}\|_\sigma - \|T_\sigma^{\widetilde{\mu}_i} - T_\sigma^{\widetilde{\mu}_j}\|_\sigma\right| \leq \underbrace{\|T_\sigma^{\mu_i} - T_\sigma^{\widetilde{\mu}_i}\|_\sigma}_{\leq\eta} + \underbrace{\|T_\sigma^{\mu_j} - T_\sigma^{\widetilde{\mu}_i}\|_\sigma}_{\leq\eta} \leq 2\eta.$$

The rest of the terms are bounded the same exact way as before, and the result follows. □

In this section, we have shown that results for the case when the $\mu_i$ are compactly supported can be extended to non-compactly supported $\mu_i$ as long as their densities decay fast enough and the reference distribution $\sigma$ has a compact and convex support.

## 6. Conditions on $\mathscr{H}$ and $\mu$ (compact case)

In this section, we derive conditions on $\mathscr{H}$ and $\mu$ so that the assumptions of the theorems above are satisfied for $\mu_i \sim \mathscr{H}_\sharp \mu$. In particular, we can break down our requirements on $\mathscr{H}$ and $\mu$ by noting the necessary conditions on $\mu_i$ for the barycentric map estimator and entropic map estimator separately. For simplicity, we will assume that $\mathscr{H}$ is exactly compatible with respect to $\sigma$ and $\mu$.

In Theorem 6.1, we first describe the conditions to have Theorem 4.2 hold when $\mathscr{H}$ is *not* made of exactly-compatible transformations. Next, we consider the easier case of exactly-compatible transformations in Theorem 6.2 and Theorem 6.3 to ensure that Theorem 4.2 and Theorem 4.8 works.

**Theorem 6.1** *(Barycentric Map Case (compact, non-compatible)). Let Assumption 4.1 hold true. If $\mu$ has simply connected compact support with density such that $0 < c \leq f_\mu \leq C < \infty$, $h_i \in \mathscr{H}$ is continuously differentiable with $L_1 I \preceq J_{h_i}(x)$ for some $L_1 > 0$, and $(h_i)_\sharp \mu$ has convex support, then $T_\sigma^{(h_i)_\sharp \mu}$ is Lipschitz. Hence condition (i) of Theorem 4.2 holds.*

**Theorem 6.2** *(Barycentric Map Case (compact, compatible)). Along with Assumption 4.1 (with $\varepsilon = 0$ so that every $h \in \mathscr{H}$ is exactly compatible with $\sigma$ and $\mu$), assume that $h \in \mathscr{H}$ is continuous, $\mu_i \sim \mathscr{H}_\sharp \mu$ i.i.d., and that*

*(i) $\mu$ has and has simply connected compact supported;*
*(ii) $\sigma$ is chosen such that $T_\sigma^\mu$ is Lipschitz.*

*Then $\mu_i$ satisfies the condition (i) of Theorem 4.2 i.e., $T_\sigma^{\mu_i}$ is Lipschitz.*

For the entropic case, the assumptions on $\mu$ and $\sigma$ are the same, but we require an additional assumption regarding the Jacobian of elements of $\mathscr{H}$.

**Theorem 6.3** *(Entropic Map Case (compact)). Along with Assumption 4.1, assume that*

*(iii) **Conditions for $\sigma$ and $\mu$:** $\sigma$ and $\mu$ satisfy (A1)-(A3), and $\mu$ has simply connected compact support;*
*(iv) **Conditions for $\mathscr{H}$:** for each $h \in \mathscr{H}$, we have $h$ is continuous and $L_1 I \preceq J_h(x) I$ for some $L_1 > 0$; moreover, each $h \in \mathscr{H}$ is exactly compatible (i.e., $\varepsilon = 0$).*

*Then $\mu_i$ satisfies the condition (i) of Theorem 4.8.*

The proofs of Theorem 6.1, Theorem 6.2, and Theorem 6.3 are given in Appendix D.1.

## 7. Conditions on $\mathscr{H}$ and $\mu$ (non-compact case)

For the non-compactly supported cases, we need to add assumptions that $\mathscr{H}$ is closed under inversion as well as lower and upper boundedness of the density $f_\mu$. This gives us the following theorems.

**Theorem 7.1** *(Barycentric Map Case (non-compact)). Along with Assumption 4.1 (with $\varepsilon = 0$ so that every $h \in \mathscr{H}$ is exactly compatible with $\sigma$ and $\mu$), assume that $\mu_i \sim \mathscr{H}_\sharp \mu$ i.i.d. Assume further that*

*(i) for every $h \in \mathscr{H}$, there exists an inverse $h^{-1} \in \mathscr{H}$.*
*(ii) The density of $\mu$ is supported on all of $\mathbb{R}^n$ with $f_\mu(x) \leq C < \infty$ for all $x$, and $f_\mu(x) \geq c > 0$ for all $x \in B(0, AR)$. Moreover, $f_\mu$ has a decay rate as in Lemma 5.3 for $x \notin B(0, R)$.*
*(iii) every $h \in \mathscr{H}$, is bi-Lipschitz with $L_1|x-y| \leq |h(x) - h(y)| \leq L_2|x-y|$, where $a, A > 0$ is from Assumption 4.1.*

*Then $\mu_i$ satisfies the conditions of Theorem 5.4.*

**Theorem 7.2** *(Entropic Map Case (non-compact)). Assume that $\mu_i \sim \mathscr{H}_\sharp \mu$ i.i.d. and that $\mu$, $\mathscr{H}$, and $\sigma$ satisfy the conditions of Theorem 7.1. Then $\mu_i$ satisfies the conditions of Theorem 5.5.*

The proofs of both Theorems 7.1 and 7.2 are found in Appendix D.2.

## 8. Experiments

We demonstrate that Algorithm 2 does in fact attain correct embeddings given finite sampling and without explicitly computing the pairwise Wasserstein distances. We test both variants of our algorithm above using the linear program or entropic regularization

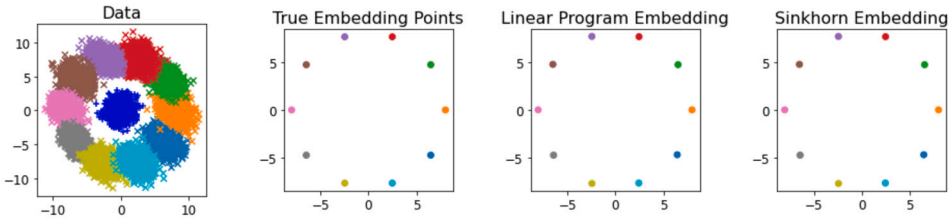



**Fig. 1.** 1-D Manifold of translations: **(Left)** reference measure $\sigma \sim \mathcal{N}(0, I)$ in blue and data measures $\mu_i$ which are Gaussians with the same covariance matrix and means $x_i$ uniformly sampled from the circle of radius 8. **(Left Middle)** Means $x_i$ of $\mu_i$ which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 2 using the linear program. **(Right)** Embedding attained with Algorithm 2 using the Sinkhorn distance with $\lambda = 1$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

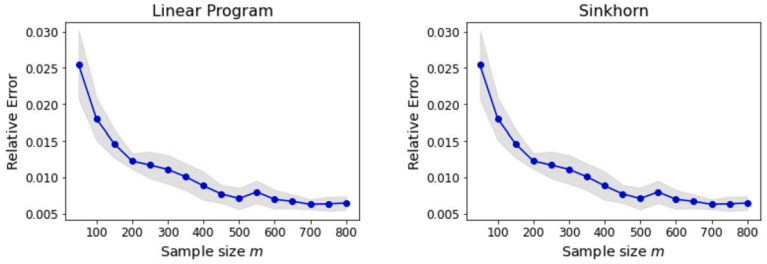



**Fig. 2.** Embedding error vs. $m$ (number of sample points from data and reference distributions for the 1-D translation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 1$ **(Right)**.

to compute the transport maps from the data to the reference measure, and illustrate the quality of embeddings as well as the relative embedding error

$$\min_{Q} \frac{\|Y - QX\|_{\mathrm{F}}}{\|Y\|_{\mathrm{F}}}$$

as a function of the sample size $m$ of the data and reference measures.

In all experiments, we generate $N$ data measures, $\mu_i$, which are Gaussians of various means and covariance, and a fixed reference measure $\sigma$ drawn from the standard normal distribution $\mathcal{N}(0, I)$. We randomly sample $m$ points from each measure to form the empirical measure, and random noise from a Wishart distribution is added to the covariance matrices of the data measures $\mu_i$. Additionally, in each experiment we compute the optimal rotation of the embeddings to properly align them with the true embedding and thus give an accurate error estimate for each trial.

For each experiment, we provide a figure for qualitative assessment of the embedding as well as a quantitative figure in which we compute the relative error as above for the embeddings as a function of $m$, the sample size used to generate the empirical data and reference measures. For the latter figures, we run 10 trials of the embedding and average the relative error; error bands showing one standard deviation are shown on each figure. A Jupyter notebook containing all of the experiments that generate the figures below can be found at https://github.com/varunkhuran/LOTWassMap.

### 8.1. Experiment 1: circle translation manifold

First, we consider a 1-dimensional manifold of translations as follows. We uniformly choose $N = 10$ points on the circle of radius 8, which we denote $x_i$, and each data measure $\mu_i$ is a Gaussian with mean $x_i$ and covariance matrix $\begin{bmatrix} 1 & -.5 \\ -.5 & 1 \end{bmatrix}$. Thus, our data set is a set of Gaussians translated around the circle. The Wishart noise added to the covariance matrix prior to sampling the $\mu_i$ is of the form $GG^{\top}$ where $G$ has i.i.d. $\mathcal{N}(0, 0.5)$ entries. We choose the standard normal distribution $\mathcal{N}(0, I)$ as our reference measure $\sigma$. We randomly sample $m = 1000$ points from each data measure and the reference measure independently. Fig. 1 shows the original sampled data and the reference measure (in blue), the true embedding points $x_i$, and the embeddings of Algorithm 2 when using the linear program and Sinkhorn with regularization parameter $\lambda = 1$.

One can easily see that the embeddings are qualitatively good as expected given the theory above and the results of [21] in similar experiments. Fig. 2 shows the relative error vs. sampling size $m$ of the measures, and one can see the good performance for modest sample sizes.

### 8.2. Experiment 2: rotation manifold

Next, we consider a 1-dimensional rotation manifold in which we generate $N = 10$ data measures of Gaussians whose means lie at uniform samples of the circle of radius 8, which we denote $(8\cos\theta_i, 8\sin\theta_i)$, and whose covariance matrices are rotations of $\begin{bmatrix} 2 & 0 \\ 0 & .5 \end{bmatrix}$

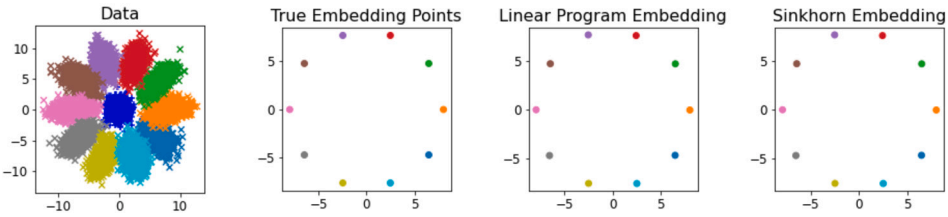



**Fig. 3.** 1-D Manifold of rotations: **(Left)** reference measure $\sigma \sim \mathcal{N}(0, I)$ in blue and data measures $\mu_i$ which are Gaussians with means lying on the circle of radius 8 and covariance matrices that are rotations of each other. **(Left Middle)** Means $x_i$ of $\mu_i$ which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 2 using the linear program. **(Right)** Embedding attained with Algorithm 2 using the Sinkhorn distance with $\lambda = 1$.

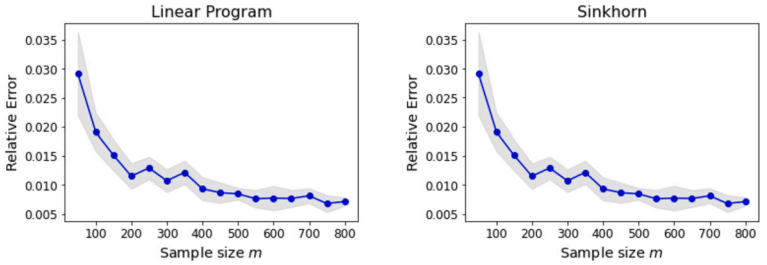



**Fig. 4.** Embedding error vs. $m$ (number of sample points from data and reference distributions for the 1-D rotation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 1$ **(Right)**.

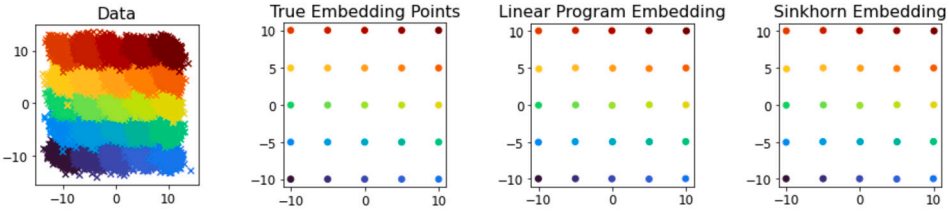



**Fig. 5.** 2-D Manifold of translations: **(Left)** data measures $\mu_i$ which are Gaussians with the same covariance matrix and means $x_i$ taken from a $5 \times 5$ uniform grid on $[-10, 10]^2$. **(Left Middle)** Means $x_i$ of $\mu_i$ which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 2 using the linear program. **(Right)** Embedding attained with Algorithm 2 using the Sinkhorn distance with $\lambda = 10$.

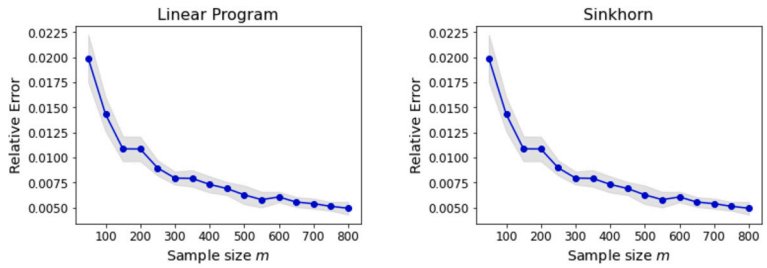



**Fig. 6.** Embedding error vs. $m$ (number of sample points from data and reference distributions for the 2-D translation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 10$ **(Right)**.

by the angles $\theta_i$. As in experiment 1, the noise level added is 0.5 and we sample $m = 1000$ points from each measure. Fig. 3 shows the data measures, true embedding, and embeddings from Algorithm 2 using both the linear program and Sinkhorn (with $\lambda = 1$) to compute the optimal transport maps. Fig. 4 shows the relative error vs. sample size.

### 8.3. Experiment 3: grid translation manifold

Here, we consider a 2-dimensional translation manifold in which we generate $N = 25$ data measures of Gaussians whose means lie on a $5 \times 5$ uniform grid on the cube $[-10, 10]^2$ and which have constant covariance matrix $\begin{bmatrix} 1 & -.5 \\ -.5 & 1 \end{bmatrix}$. We sample $m = 1000$ points from each measure and the noise level is again 0.5. In the Sinkhorn embedding, we use regularization $\lambda = 10$. Figs. 5 and 6 show the data, embeddings, and relative error vs. sample size.

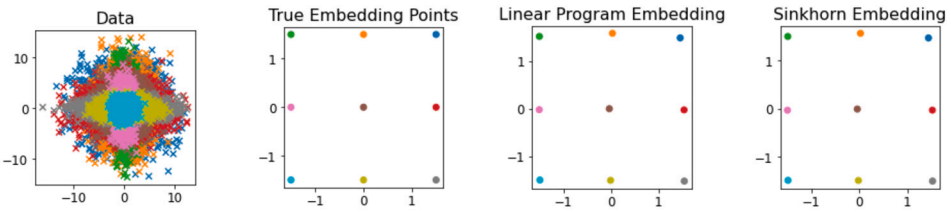



**Fig. 7.** 2-D Manifold of Anisotropic Dilations: **(Left)** data measures $\mu_i$ which are Gaussians with mean 0 and anisotropically dilated covariance matrices where dilations are taken from a $3 \times 3$ uniform grid on $[1,4]^2$. **(Left Middle)** Dilation factors $(x_i, y_i)$ of $\mu_i$ which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 2 using the linear program. **(Right)** Embedding attained with Algorithm 2 using the Sinkhorn distance with $\lambda = 100$.

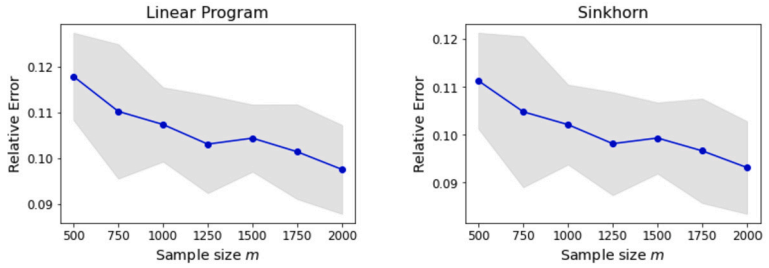



**Fig. 8.** Embedding error vs. $m$ (number of sample points from data and reference distributions for the 2-D translation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 10$ **(Right)**.

### 8.4. Experiment 4: dilation manifold

Here, we consider a 2-dimensional anisotropic dilation manifold in which we generate $N = 9$ data measures of Gaussians with mean 0 and anisotropically scaled covariance matrices of the form $\text{diag}(\alpha_i^2, \beta_i^2)$ for $(\alpha_i, \beta_i)$ taken from a uniform $3 \times 3$ grid on $[1,4]^2$. We sample $m = 1000$ points from the reference measure and $n = 2500$ points from the data measures and the noise level added to the covariance matrices is 0.5 as before. In the Sinkhorn embedding, we use regularization $\lambda = 100$. Fig. 7 show the data measures, true embedding parameters, and embeddings from Algorithm 2. Note that the true embedding parameters are centered to allow them to be comparable to the output of Algorithm 2 which are naturally centered.

Fig. 8 shows the relative error vs. $m$, and for this experiment we choose $n = m$ so that the sampling order of the data and reference measure are the same. For this case, we see that the relative error of the embedding decays much more slowly than the previous experiments. One possible reason for this is that there is significant overlap in the distributions for the dilated measures, and to overcome this issue one may have to sample many more points in forming the empirical distribution so that the tails of the data measures are sampled more frequently.

### 8.5. Experiment 5: time comparison

Here, we repeat Experiment 3 in which data measures are centered on a uniform grid and are translations of a fixed Gaussian measure. We plot the time it takes to compute the embedding via Algorithm 2 using the Linear Program or Sinkhorn with $\lambda = 1$ and the Wassmap algorithm of [21] which requires computing the entire square Wasserstein distance matrix $[W_2(\mu_i, \mu_j)]_{i,j=1}^N$ and the SVD of its centered version as in Algorithm 1 (Fig. 9). For this experiment, we always choose $n = m$ so that the reference measure and data measure sampling rates are the same. One can easily see that a substantial gain in timing is achieved by LOT Wassmap, while previous experiments show that the quality of the embedding does not degrade significantly when LOT is used.

Finally, we plot the timing for the same experiment for the Linear Program and Sinkhorn with $\lambda = 1$ and $\lambda = 10$ for larger sample sizes to illustrate the character of these choices (Fig. 10). As expected, larger regularization parameter yields faster computation time, though the difference is relatively small even for modestly large sample size.

## Acknowledgments

K.H. acknowledges support from the UTA Research Enhancement Program from the College of Science at the University of Texas at Arlington, The Fields Institute for Research in Mathematical Sciences and the Army Research Office under Grant Number W911NF-23-1-0213. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. C.M. is supported by NSF awards DMS-2306064, DMS-2410140 and by a seed grant from the School of Data Science and Society at the University of North Carolina. A.C. is partially supported by NSF awards DMS-2012266, CISE-2403452 and a gift from Intel research.

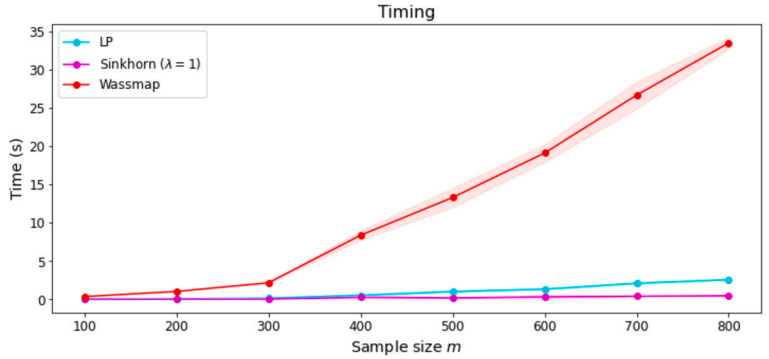



**Fig. 9.** Timing vs. sample size $m$ of the reference distribution and data measures. The data set consists of $N = 25$ measures translated on a $5 \times 5$ uniform grid on $[-10, 10]^2$ as in Experiment 3. Shown are the computation times to compute the Wassmap embedding and the embeddings of Algorithm 2 using the Linear Program (LP) and Sinkhorn with regularization parameter $\lambda = 1$.

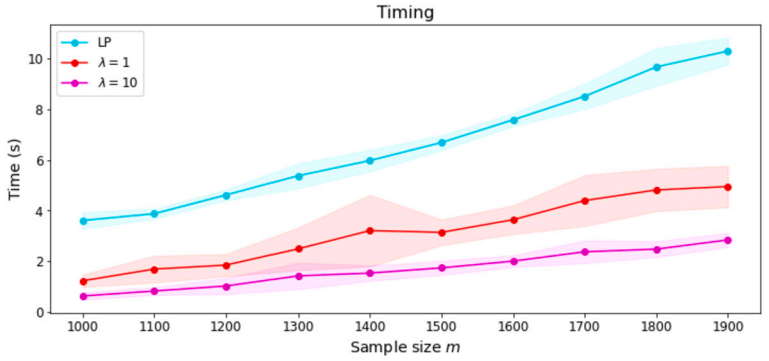



**Fig. 10.** Timing vs. sample size $m$ of the reference distribution and data measures. The data set consists of $N = 25$ measures translated on a $5 \times 5$ uniform grid on $[-10, 10]^2$ as in Experiment 3. Shown are the computation times to compute the embeddings of Algorithm 2 using the Linear Program (LP) and Sinkhorn with regularization parameters $\lambda = 1$ and $\lambda = 10$.

K.H. and A.C. thank the Fields Institute and participants of the Focus Program on Data Science, Approximation Theory, and Harmonic Analysis for their hospitality, which facilitated the initial discussions of this research.

The authors thank the reviewer for their detailed feedback that improved the manuscript.

## Appendix A. Helper theorems and lemmas

We use the following lemma to extend Corollary 3.2 to get our main theorem (Theorem 3.3). The proof follows standard arguments, e.g., as in [27]; the proof is included for completeness.

**Lemma A.1** *([27, Theorem 14.2.1], for example). Consider a matrix $V$ whose columns are centered vectors $v_1, \dots, v_n$ such that $\sum_{j=1}^n v_j = 0$. Let $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ be the centering matrix from MDS (Algorithm 1), $G = V^\top V$ be the Gram matrix for $V$, and $D$ be the squared distance matrix $D_{ij} = \|v_i - v_j\|^2$. Then $G = -\frac{1}{2} J D J$.*

**Proof.** Note first that

$$(JDJ)_{ij} = D_{ij} + \frac{1}{n^2}\sum_{k,\ell=1}^{n} D_{k\ell} - \frac{1}{n}\sum_{k=1}^{n}(D_{ik} + D_{kj}).$$

Moreover, because $D_{ij} = v_i^\top v_i + v_j^\top v_j - 2v_i^\top v_j$, we get that

$$(JDJ)_{ij} = v_i^\top v_i + v_j^\top v_j - 2v_i^\top v_j + \frac{1}{n^2}\left(2n\sum_{k=1}^{n} v_k^\top v_k - 2\mathbf{1}^\top V^\top V\mathbf{1}\right)$$
$$- \frac{1}{n}\left(nv_i^\top v_i + nv_j^\top v_j + 2\sum_{k=1}^{n} v_k^\top v_k - 2\mathbf{1}^\top V^\top v_j - 2v_i^\top V\mathbf{1}\right).$$

Note here that $V\mathbf{1} = 0$ since $\sum_{j=1}^n v_j = 0$. After canceling terms, we get

$$(JDJ)_{ij} = -2v_i^\top v_j = -2G_{ij}.$$

So our result is immediate. □

The next results are used to recount the $\varepsilon$-compatibility as well as its effects on LOT. First, we show that every $\varepsilon$-compatible map has a compatible map (with $\varepsilon = 0$) nearby whose LOT distance from the $\varepsilon$-compatible map is small.

**Lemma A.2.** *Assume that*

*(i) $\sigma$ is supported on a compact convex set $\Omega \subset \mathbb{R}^n$ with probability density $f_\sigma$ bounded above and below by positive constants.*
*(ii) $\mu$ has finite $p$-th moment with bound $M_p$ with $p > d$ and $p \geq 4$.*
*(iii) There exist $A > 0$ such that every $h \in \mathscr{H}$ satisfies $\|h(x)\| \leq A\|x\|$ for every $x \in \Omega$.*

*Let $\mathscr{H}$ be $\varepsilon$-compatible with respect to $\sigma$ and $\mu$. Then for every $h \in \mathscr{H}$ there exists a compatible $g$ such that*

$$\left\| T_\sigma^{g\sharp\mu} - T_\sigma^{h\sharp\mu} \right\|_\sigma \leq C_{n,p,\Omega,A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}$$
$$\|h \circ T_\sigma^\mu - T_\sigma^{h\sharp\mu}\|_\sigma < \varepsilon + C_{n,p,\Omega,A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}.$$

**Proof.** Let $h \in \mathscr{H}$, then there exists an exactly compatible transformation $g$ such that $g \circ T_\sigma^\mu = T_\sigma^{g\sharp\mu}$ with $\|h - g\|_\mu < \varepsilon$ by definition of $\varepsilon$-compatibility. Then notice that

$$\begin{aligned}\left\| h \circ T_\sigma^\mu - T_\sigma^{h\sharp\mu} \right\|_\sigma &= \left\| h \circ T_\sigma^\mu - g \circ T_\sigma^\mu + T_\sigma^{g\sharp\mu} - T_\sigma^{h\sharp\mu} \right\|_\sigma \\ &\leq \|h - g\|_\mu + \left\| T_\sigma^{g\sharp\mu} - T_\sigma^{h\sharp\mu} \right\|_\sigma.\end{aligned}$$

By assumption, we know that $\|h - g\|_\mu < \varepsilon$. Since $h \in \mathscr{H}$ satisfies (iii), we have

$$\int_\Omega \|x\|^p f_{h\sharp\mu}(x)dx = \int_\Omega \|x\|^p d(h\sharp\mu)(x) = \int_\Omega \underbrace{\|h(x)\|^p}_{\leq A^p\|x\|^p} d\mu(x) \leq A^p M_p$$

Similarly, note that $g \in \mathscr{H}$ because $\mathscr{H}$ consists of all $\varepsilon$-compatible pushforwards with respect to $\sigma$ and $\mu$ (i.e. there exists compatible $g'$ such that $\|g - g'\|_\mu < \varepsilon$ and $g' \circ T_\sigma^\mu = T_\sigma^{(g')\sharp\mu}$), and we can see that $g$ is itself the compatible transformation. This implies that we have the same moment bound for $g$. Now using Theorem 5.1 and equation 9 of [3], we get that

$$\begin{aligned}\left\| T_\sigma^{g\sharp\mu} - T_\sigma^{h\sharp\mu} \right\|_\sigma &\leq C_{n,p,\Omega,A^pM_p} W_1(g\sharp\mu, h\sharp\mu)^{\frac{p}{6p+16n}} \\ &\leq C_{n,p,\Omega,A^pM_p} W_2(g\sharp\mu, h\sharp\mu)^{\frac{p}{6p+16n}} \\ &\leq C_{n,p,\Omega,A^pM_p} \|h - g\|_\mu^{\frac{p}{6p+16n}} \\ &\leq C_{n,p,\Omega,A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}.\end{aligned}$$

This implies that

$$\|h \circ T_\sigma^\mu - T_\sigma^{h\sharp\mu}\|_\sigma < \varepsilon + C_{n,p,\Omega,A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}. \quad \square$$

Now we can show that the LOT embedding between exactly compatible transformations is isometric with the Wasserstein manifold.

**Lemma A.3.** *Let $g_1$ and $g_2$ be exactly compatible transformations, i.e. $g_1 \circ T_\sigma^\mu = T_\sigma^{(g_1)\sharp\mu}$ and $g_2 \circ T_\sigma^\mu = T_\sigma^{(g_2)\sharp\mu}$, then*

$$\left\| T_\sigma^{(g_1)\sharp\mu} - T_\sigma^{(g_2)\sharp\mu} \right\|_\sigma = W_2\Big((g_1)\sharp\mu, (g_2)\sharp\mu\Big).$$

**Proof.** First notice that since everything is absolutely continuous, we can use a change of variables formula to get

$$\left\| T_\sigma^{(g_1)\sharp\mu} - T_\sigma^{(g_2)\sharp\mu} \right\|_\sigma = \left\| I - T_\sigma^{(g_2)\sharp\mu} \circ T_{(g_1)\sharp\mu}^\sigma \right\|_{(g_1)\sharp\mu}.$$

Because $T_{(g_1)\sharp\mu}^{(g_2)\sharp\mu}$ is the minimizer of the optimal transport problem and the triangle inequality, we get

$$W_2\Big((g_1)\sharp\mu, (g_2)\sharp\mu\Big) = \left\| I - T_{(g_1)\sharp\mu}^{(g_2)\sharp\mu} \right\|_{(g_1)\sharp\mu} \leq \left\| I - T_\sigma^{(g_2)\sharp\mu} \circ T_{(g_1)\sharp\mu}^\sigma \right\|_{(g_1)\sharp\mu}$$

$$\leq \left\| I - T^{(g_2)_\sharp\mu}_{(g_1)_\sharp\mu} \right\|_{(g_1)_\sharp\mu} + \left\| T^{(g_2)_\sharp\mu}_{(g_1)_\sharp\mu} - T^{(g_2)_\sharp\mu}_{\sigma} \circ T^{\sigma}_{(g_1)_\sharp\mu} \right\|_{(g_1)_\sharp\mu}.$$

Note that Theorem 24 of [25] implies that given an exactly compatible transformation $g$, $J_g(T^\mu_\sigma(x))$ must share the same eigenspaces as $J_{T^\mu_\sigma}(x)$. By Corollary 4 of [25], we know that exactly compatible transformations are optimal transport maps themselves. This means that $T^{g_\sharp\mu}_\mu = g$ for exactly compatible transport maps. Moreover, for an exactly compatible $h' \in \mathscr{H}$, this means that $T^{(g')_\sharp\mu}_{g_\sharp\mu} = g' \circ g^{-1}$ because $g' \circ g^{-1}$ is a gradient of a convex function (since the Jacobian of $g$ and $g'$ share the same eigenspaces) that pushes $g_\sharp\mu$ to $(g')_\sharp\mu$. In the context of $g_1$ and $g_2$, this gives us that

$$T^{(g_2)_\sharp\mu}_{(g_1)_\sharp\mu} = g_2 \circ g_1^{-1} = g_2 \circ T^\mu_\sigma \circ T^\sigma_\mu \circ g_1^{-1} = T^{(g_2)_\sharp\mu}_\sigma \circ T^\sigma_{(g_1)_\sharp\mu}.$$

In particular, we get that

$$\left\| T^{(g_1)_\sharp\mu}_\sigma - T^{(g_2)_\sharp\mu}_\sigma \right\|_\sigma = W_2\Big((g_1)_\sharp\mu, (g_2)_\sharp\mu\Big). \quad \square$$

Finally, we show that $\varepsilon$-compatible transformations have LOT embeddings that are "$\varepsilon^{\frac{p}{6p+16n}}$-isometric" in the sense of the following theorem.

**Theorem A.4.** *Assume that*

*(i)* $\sigma$ *is supported on a compact convex set* $\Omega \subset \mathbb{R}^n$ *with probability density* $f_\sigma$ *bounded above and below by positive constants.*
*(ii)* $\mu$ *has finite* $p$*-th moment with bound* $M_p$ *with* $p > n$ *and* $p \geq 4$.
*(iii) There exist constants* $a, A > 0$ *such that every* $h \in \mathscr{H}$ *satisfies* $a\|x\| \leq \|h(x)\| \leq A\|x\|$.

*Let* $\mathscr{H}$ *be* $\varepsilon$*-compatible with respect to absolutely continuous measures* $\sigma$ *and* $\mu$ *and assume that* $h_\sharp\mu$ *is absolutely continuous. Then for* $h_1, h_2 \in \mathscr{H}$,

$$\left| W_2\Big((h_1)_\sharp\mu, (h_2)_\sharp\mu\Big) - \left\| T^{(h_1)_\sharp\mu}_\sigma - T^{(h_2)_\sharp\mu}_\sigma \right\|_\sigma \right| < 2\left(\varepsilon + C_{n,p,\Omega,a^{-1}A^p M_p} \cdot \varepsilon^{\frac{p}{6p+16n}}\right) < C\varepsilon^{\frac{p}{6p+16n}}$$

**Proof.** By definition, we know that there exist $g_1$ and $g_2$ such that $\|g_1 - h_1\|_\mu < \varepsilon$ and $\|g_2 - h_2\|_\mu < \varepsilon$. First, note that

$$\left\| T^{(h_1)_\sharp\mu}_\sigma - T^{(h_2)_\sharp\mu}_\sigma \right\|_\sigma \leq \left\| T^{(h_1)_\sharp\mu}_\sigma - T^{(g_1)_\sharp\mu}_\sigma \right\|_\sigma + \left\| T^{(g_1)_\sharp\mu}_\sigma - T^{(g_2)_\sharp\mu}_\sigma \right\|_\sigma + \left\| T^{(g_2)_\sharp\mu}_\sigma - T^{(h_2)_\sharp\mu}_\sigma \right\|_\sigma.$$

By Lemma A.3, we know that

$$\left\| T^{(g_1)_\sharp\mu}_\sigma - T^{(g_2)_\sharp\mu}_\sigma \right\|_\sigma = W_2\Big((g_1)_\sharp\mu, (g_2)_\sharp\mu\Big).$$

However, by equation 2.1 of [3] and the triangle inequality, we have

$$\begin{aligned} W_2\Big((g_1)_\sharp\mu, (g_2)_\sharp\mu\Big) &\leq \underbrace{W_2\Big((g_1)_\sharp\mu, (h_1)_\sharp\mu\Big)}_{\leq \|g_1 - h_1\|_\mu < \varepsilon} + W_2\Big((h_1)_\sharp\mu, (h_2)_\sharp\mu\Big) + \underbrace{W_2\Big((h_2)_\sharp\mu, (g_2)_\sharp\mu\Big)}_{\leq \|h_2 - g_2\|_\mu < \varepsilon} \\ &\leq W_2\Big((h_1)_\sharp\mu, (h_2)_\sharp\mu\Big) + 2\varepsilon. \end{aligned}$$

Moreover, by Lemma A.2, for $i = 1, 2$, we know that

$$\left\| T^{(g_i)_\sharp\mu}_\sigma - T^{(h_i)_\sharp\mu}_\sigma \right\|_\sigma \leq C_{n,p,\Omega,a^{-1}A^p M_p} \cdot \varepsilon^{\frac{p}{6p+16n}}.$$

This implies that

$$\begin{aligned} W_2\Big((h_1)_\sharp\mu, (h_2)_\sharp\mu\Big) &\leq \left\| T^{(h_1)_\sharp\mu}_\sigma - T^{(h_2)_\sharp\mu}_\sigma \right\|_\sigma \\ &\leq W_2\Big((h_1)_\sharp\mu, (h_2)_\sharp\mu\Big) + 2\left(\varepsilon + C_{n,p,\Omega,a^{-1}A^p M_p} \varepsilon^{\frac{p}{6p+16n}}\right), \end{aligned}$$

and the proof is complete. $\square$

## Appendix B. Plug-in estimator approximation results

In this section, we provide some auxiliary results that are used along the way to prove the theorems of Section 4.

### B.1. Using the linear program to compute transport maps

Recall that for a random variable $X_m$, we say that $X_m = O_p(a_m)$ if for every $\varepsilon > 0$ there exists $M > 0$ and $N > 0$ such that

$$\mathbb{P}\Big( |X_m/a_m| > M \Big) < \varepsilon \quad \forall m \geq N.$$

The following theorem from [18] is used in the proofs of our main results, including Theorem 4.2.

**Theorem B.1** *([18, Theorem 2.2]). Suppose that $T_\sigma^\mu$ is L-Lipschitz, and $\mu$ is compactly supported and $\mathbb{E}_\sigma[\exp(t\|x\|^\alpha)] < \infty$ for some $t > 0, \alpha > 0$. Assume we draw $k$ i.i.d. samples from $\mu$ and consider the estimator $\widehat{\mu}$. Then*

$$\sup_{\gamma \in \Gamma_{\min}} \int \|T_\sigma^{\widehat{\mu}}(x;\gamma_{LP}) - T_\sigma^\mu(x)\|^2 d\sigma(x) \leq O_p(r_n^{(k)} \log(1+k)^{t_{n,\alpha}}),$$

*where*

$$r_n^{(k)} = \begin{cases} 2k^{-1/2} & n = 2,3 \\ 2k^{-1/2}\log(1+k) & n = 4 \\ 2k^{-2/n} & n \geq 5 \end{cases}, \quad t_{n,\alpha} = \begin{cases} (4\alpha)^{-1}(4 + ((2\alpha + 2n\alpha - n) \vee 0)) & n < 4 \\ (\alpha^{-1} \vee 7/2) - 1 & n = 4 \\ 2(1+n^{-1}) & n > 4 \end{cases},$$

*so that $r_n^{(k)}$ and $t_{n,\alpha}$ are on the order of $k^{-1/n}$ and $2(1+n^{-1})$, respectively.*

We utilize the above theorem for the case that $\sigma$ is compactly supported in the proof of Theorem 4.2. Consequently, the expectation bound holds for all $t, \alpha > 0$. We see that for $n < 4$, one can choose $\alpha$ large enough so that $t_{n,\alpha}$ is arbitrarily close to 1 if $n = 1$, $\frac{3}{2}$ if $n = 2$, and 2 if $n = 3$. Similarly, for $n = 4$ we may choose $\alpha$ large enough so that $t_{n,\alpha} = \frac{5}{2}$. This simplifies the statement in Theorem 4.2.

**Remark B.2.** We note that Theorem B.1 is the "semi-discrete" version described in [18]. The paper also provides equivalent bounds in the instance that $\sigma$ is similarly estimated. However, the bounds only guarantee that the transport maps agree when integrated against $\widehat{\sigma}$, whereas we need the bound for $\sigma$ itself.

### B.2. Approximating with finite samples from the reference distribution

Some of the norms from Theorem 4.2 and Theorem 4.8 are assumed to be integrated against the true $\sigma$. However, we need to consider the discretized $\sigma$ for each norm, and establish that we can estimate these norms with high probability. For these bounds, we use McDiarmid's inequality on the function

$$f(X_1, ..., X_m) = \frac{1}{m}\sum_{j=1}^m \left| T_\sigma^{\widehat{\mu_1}}(X_j;\gamma_{\widehat{\mu_1}}) - T_\sigma^{\widehat{\mu_2}}(X_j;\gamma_{\widehat{\mu_2}}) \right|^2 = \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2,$$

where $X_j \sim \sigma$, $\gamma_{\widehat{\mu_j}}$ is a transport plan between $\sigma$ and $\widehat{\mu}_j$ for $j = 1, 2$, and $\gamma \in \{\gamma_{LP}, \gamma_\beta\}$ denotes the optimization method used to get $\gamma_{\widehat{\mu_j}}$. If $\mu_i$ are supported in a ball of radius $R$, then McDiarmid's inequality implies

$$\mathbb{P}\left( \left| \frac{1}{m}\sum_{j=1}^m |T_\sigma^{\widehat{\mu_1}}(X_j;\gamma_{\widehat{\mu_1}}) - T_\sigma^{\widehat{\mu_2}}(X_j;\gamma_{\widehat{\mu_2}})|^2 - \|T_\sigma^{\widehat{\mu_1}}(\cdot;\gamma_{\widehat{\mu_1}}) - T_\sigma^{\widehat{\mu_2}}(\cdot;\gamma_{\widehat{\mu_2}})\|_\sigma^2 \right| > t \right) \leq 2e^{-m\frac{t^2}{32R^4}}.$$

Note that since $f = \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2$, we get

$$\mathbb{P}\left( \left| \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2 - W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2 \right| > t \right) \leq 2e^{-m\frac{t^2}{32R^4}}. \tag{18}$$

**Theorem B.3.** *Consider $\mu_i, \sigma \in W_2(\mathbb{R}^n)$. Assume $supp(\mu_i) \subset B(0,R)$ for $i = 1,2$. Let $\delta > 0$. Then with probability at least $1 - \delta$,*

$$\left| W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma) - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma) \right| \leq R\sqrt{\frac{2\log(2/\delta)}{m}},$$

*where $m$ is the number of samples used to estimate $\sigma$. Here, if $\sigma$ is not absolutely continuous, then the transport maps are constructed from barycentric projections of transport plans.*

**Proof.** Define

$$a = W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma), \quad b = \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma).$$

Then both $a \leq 2R$ and $b \leq 2R$. Now, since $a^2 - b^2 = (a+b)(a-b)$, we get that

$$|a-b| \geq \frac{1}{4R}|a^2-b^2|.$$

This, together with (18), implies that

$$\mathbb{P}\left(\left|\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma) - W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma)\right| > t\right) \leq 2e^{-m\frac{t^2}{2R^2}}.$$

Solving $\delta = 2e^{-m\frac{t^2}{2R^2}}$ for $t$ yields the conclusion. □

The following corollary is geared towards showing Theorem 4.8 as we use the estimated optimal transport map generated from using the Sinkhorn transport plan solution.

**Corollary B.4.** *Under the assumptions of Theorem B.3, suppose $X_j \sim \sigma$ i.i.d. for $(j = 1, \dots, m)$ and let $\widehat{\sigma} = \frac{1}{m}\sum_{j=1}^m \delta_{X_j}$. Then with probability at least $1-\delta$,*

$$\left|\|T_{\widehat{\sigma}}^{\widehat{\mu_1}}(\cdot\,;\gamma_{\widehat{\mu_1},\beta}) - T_{\widehat{\sigma}}^{\widehat{\mu_2}}(\cdot\,;\gamma_{\widehat{\mu_2},\beta})\|_\sigma^2 - \widehat{W}_{2,\widehat{\sigma}}^{LOT}(\widehat{\mu}_i,\widehat{\mu}_j)\right| \leq R\sqrt{\frac{2\log(2/\delta)}{m}}.$$

**Proof.** Use the sampling $X_j \sim \sigma$ that generates $\widehat{\sigma}$, we can use McDiarmid's inequality to get

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{j=1}^m |T_{\widehat{\sigma}}^{\widehat{\mu_1}}(X_j;\gamma_{\widehat{\mu_1},\beta}) - T_{\widehat{\sigma}}^{\widehat{\mu_2}}(X_j;\gamma_{\widehat{\mu_2},\beta})|^2 - \|T_{\widehat{\sigma}}^{\widehat{\mu_1}}(\cdot\,;\gamma_{\widehat{\mu_1},\beta}) - T_{\widehat{\sigma}}^{\widehat{\mu_2}}(\cdot\,;\gamma_{\widehat{\mu_2},\beta})\|_\sigma^2\right| > t\right) \leq 2e^{-m\frac{t^2}{32R^4}}.$$

We use the same sampling for $\widehat{\sigma}$ as we do for the concentration for McDiarmid's inequality. Notice, however, that

$$\underbrace{\frac{1}{m}\sum_{j=1}^m |T_{\widehat{\sigma}}^{\widehat{\mu_1}}(X_j;\gamma_{\widehat{\mu_1},\beta}) - T_{\widehat{\sigma}}^{\widehat{\mu_2}}(X_j;\gamma_{\widehat{\mu_2},\beta})|^2}_{\widehat{W}_{2,\widehat{\sigma}}^{LOT}(\widehat{\mu}_i,\widehat{\mu}_j)} = \underbrace{\|T_{\widehat{\sigma}}^{\widehat{\mu_1}}(\cdot\,;\gamma_{\widehat{\mu_1},\beta}) - T_{\widehat{\sigma}}^{\widehat{\mu_2}}(\cdot\,;\gamma_{\widehat{\mu_2},\beta})\|_{\widehat{\sigma}}^2}_{W_{2,\widehat{\sigma}}^{LOT}(\widehat{\mu}_i,\widehat{\mu}_j)}.$$

This yields the result. □

## Appendix C. Non-compactly supported measures proofs and results

Here, we give the proofs of the lemmas preceding Theorems 5.4 and 5.5.

**Proof of Lemma 5.2.** We will construct the measure $\widetilde{\mu}$ by constructing a transport map that sends $\mu$ to a compactly supported absolutely continuous measure. In particular, for some $0 < \rho \ll 1$, consider the map

$$S_{R,\rho}(x) = \begin{cases} x & x \in B(0,R) \\ R\frac{x}{\|x\|} + \min\{\|x\| - R, \rho\}\frac{x}{1+\|x\|} & x \notin B(0,R) \end{cases}.$$

Let $\widetilde{\mu} = (S_{R,\rho})_\sharp \mu$, then the compact set that $\widetilde{\mu}$ will be supported on is $\overline{B(0,R+\rho)}$ since for $x \in \mathbb{R}^n$ with $\|x\| \gg R$ we have $\|S_{R,\rho}(x)\| < R+\rho$. Now note that

$$\begin{aligned} W_1(\mu,\widetilde{\mu}) &= \min_{S:S_\sharp\mu=\widetilde{\mu}} \int_{\mathbb{R}^n} \|S(x)-x\| d\mu(x) \leq \int_{\mathbb{R}^n} \|S_{R,\rho}(x)-x\| d\mu(x) \\ &= \int_{B(0,R)} \underbrace{\|x-x\|}_{=0} d\mu(x) + \int_{\mathbb{R}^n\setminus B(0,R)} \left\|\left(1 - \frac{R}{\|x\|} - \frac{\min\{\|x\|-R,\rho\}}{1+\|x\|}\right)x\right\| d\mu(x) \\ &\leq \int_{\mathbb{R}^n\setminus B(0,R)} \|x\| + \underbrace{R}_{\leq\|x\|} + \underbrace{\frac{\|x\|\min\{\|x\|-R,\rho\}}{1+\|x\|}}_{\leq\rho\leq 1\leq\|x\|} d\mu(x) \leq \int_{\mathbb{R}^n\setminus B(0,R)} 3\|x\| d\mu(x). \end{aligned}$$

To bound this integral, recall that

$$f_\mu(x) < \left(\frac{\eta}{\widetilde{C}_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{\|x\|^{n+2}}$$

with

$$\widetilde{C}_{n,p,\Omega,M_p} = C_{n,p,\Omega,M_p}(3C)^{\frac{p}{6p+16n}}$$

where $C$ denotes a constant of integration over concentric $n$-spheres. Recall that $d\mu(x) = f_\mu(x)dx$; thus,

$$\begin{aligned}
\int_{\mathbb{R}^n\setminus B(0,R)} 3\|x\| d\mu(x) &= \int_{\mathbb{R}^n\setminus B(0,R)} 3\|x\| f_\mu(x)dx \\
&\le \int_{\mathbb{R}^n\setminus B(0,R)} \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{C\|x\|^{n+1}}dx \\
&\le \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \underbrace{\int_{r\ge R} \frac{r^{n-1}}{r^{n+1}}dr}_{\le 1} \\
&= \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}},
\end{aligned}$$

where $C$ is a constant from integrating over concentric $n$-spheres. Invoking Theorem 5.1, this means that

$$\|T_\sigma^\mu - T_\sigma^{\widetilde{\mu}}\|_\sigma \le C_{n,p,\Omega,M_p} W_1(\mu,\widetilde{\mu})^{\frac{p}{6p+16n}} \le C_{n,p,\Omega,M_p}\frac{\eta}{C_{n,p,\Omega,M_p}} = \eta.$$

To see that $\widetilde{\mu}$ is compactly supported, notice that for $x \in \mathbb{R}^n \setminus B(0,R)$, we have

$$\|S_{R,\rho}(x)\| = \left\| R\frac{x}{\|x\|} + \min\{\|x\| - R, \rho\}\frac{x}{1+\|x\|}\right\| \le R + \rho\underbrace{\frac{\|x\|}{1+\|x\|}}_{\le 1} \le R + \rho.$$

The case for when $x \in B(0,R)$ is trivial since $S_{R,\rho}$ is the identity map on $B(0,R)$. Moreover, to see that $\widetilde{\mu}$ is absolutely continuous with respect to the Lebesgue measure, we will take a generic set $A$ and break it up into components and analyze each component. We first notice that $S_{R,\rho}$ is continuous. Indeed, for $x$ such that $\|x\| = R$, we see that

$$\underbrace{R\frac{x}{\|x\|}}_{x} + \underbrace{\min\{\|x\| - R, \rho\}}_{=\|x\|-R=0}\frac{x}{1+\|x\|} = x.$$

Now, let $A \in \mathbb{R}^n$ such that $\lambda(A) = 0$ for the Lebesgue measure $\lambda$, then

$$\begin{aligned}
&A = (A\cap B(0,R)) \oplus (A\setminus \overline{B(0,R)}) \oplus (A\cap \partial B(0,R)) \\
\implies &(S_{R,\rho})_\sharp\mu(A) = (S_{R,\rho})_\sharp\mu(A\cap B(0,R)) + (S_{R,\rho})_\sharp\mu(A\setminus\overline{B(0,R)}) + (S_{R,\rho})_\sharp\mu(A\cap\partial B(0,R)) \\
&\quad = \mu({S_{R,\rho}}^{-1}(A\cap B(0,R))) + \mu({S_{R,\rho}}^{-1}(A\setminus\overline{B(0,R)})) + \mu({S_{R,\rho}}^{-1}(A\cap\partial B(0,R))) \\
&\quad = \mu(A\cap B(0,R)) + \underbrace{\mu(A\cap\partial B(0,R))}_{\le\mu(\partial B(0,R))=0} + \mu({S_{R,\rho}}^{-1}(A\setminus\overline{B(0,R)})),
\end{aligned}$$

where we use the additivity of measures over disjoint sets, the form of $S_{R,\rho}$ on $B(0,R)$, and the absolutely continuity of $\mu$ so that $\mu(\partial B(0,R)) \le \lambda(\partial B(0,R)) = 0$. Moreover, note that $\mu(A\cap B(0,R)) \le \mu(A) \le \lambda(A) = 0$. The only term left is $A\setminus\overline{B(0,R)}$. Since $S_{R,\rho}$ is smooth on $\mathbb{R}^n\setminus B(0,R)$, there exists a density $g$ for $(S_{R,\rho})_\sharp\mu$ with respect to $\mu$ for sets in $\mathbb{R}^n\setminus B(0,R)$. This means $(S_{R,\rho})_\sharp\mu \ll \mu$ on $\mathbb{R}^n\setminus\overline{B(0,R)}$. Since $\mu \ll \lambda$, we have

$$\lambda(A) = 0 \implies \mu(A) = 0 \implies \mu(A\setminus\overline{B(0,R)}) = 0 \implies (S_{R,\rho})_\sharp\mu(A\setminus\overline{B(0,R)}) = 0.$$

This shows that $(S_{R,\rho})_\sharp\mu$ is absolutely continuous with respect to $\lambda$, so the proof is complete. □

**Proof of Lemma 5.3.** Rather than constructing a transport map, we will construct a density $f_{\widetilde{\mu}}$ and will argue that the transport map from $\mu$ to $\widetilde{\mu}$ (the measure with density $f_{\widetilde{\mu}}$) behaves nicely. To do this, consider the following density

$$f_{\widetilde{\mu},a,R}(x) = \begin{cases} f_\mu(x) & x \in B(0,R) \\ f_\mu\left(R\frac{x}{\|x\|}\right) + \alpha\left(\frac{\|x\|}{R} - 1\right) & x \in B(0,a)\setminus B(0,R) \\ 0 & \text{otherwise} \end{cases},$$

for some $\alpha > 0$. Notice that $a$ is not specified at the moment, but it depends on $R$ and $\alpha$. Since we want $\widetilde{\mu}$ to be a probability measure, we note that

$$\widetilde{\mu}(\mathbb{R}^d) = \underbrace{\int_{B(0,R)} f_\mu(x)dx}_{\mu(B(0,R))} + \underbrace{\int_R^a r^{d-1}C(r)\left(f_\mu\left(R\frac{x}{\|x\|}\right) + \alpha\left(\frac{\|x\|}{R} - 1\right)\right)dr}_{I(a)},$$

where $C(r)$ is the integral over the sphere at radius $r$. Notice that $I(a)$ has an integrand that is increasing as a function of $r$ so that $I(a)$ itself is increasing as a function of $a$ (i.e. $\lim_{a\to\infty} I(a) = \infty$). Moreover, because $I(R) = 0$, we know from the intermediate value theorem that there exists some $a^*$ such that $I(a^*) = \mu(\mathbb{R}^d \setminus B(0,R))$. Note that from this construction, $\widetilde{\mu}$ is compactly supported, absolutely continuous with respect to the Lebesgue measure, and $0 < c \le b \le f_{\widetilde{\mu}} \le B < \infty$ for some constants $b$ and $B$.

Now, we would like to bound $W_1(\mu, \widetilde{\mu})$. Because we assume that $\mu$ has a density, there exists pushforwards that push $\mu$ to $\widetilde{\mu}$, but we will consider $S$ such that $S_\sharp \mu = \widetilde{\mu}$ and $S(x) = x$ if $x \in B(0,R)$. The set of such maps $S$ is non-empty, because we can consider the optimal transport problem between the restricted measures $\mu|_{\mathbb{R}^d\setminus B(0,R)}$ to $\widetilde{\mu}|_{\mathbb{R}^d\setminus B(0,R)}$ with the same total mass. There certainly exist pushforwards, say $S_|$ between the restricted measures because the measures have densities. To form a map $S$, let $S$ be the identity on $B(0,R)$ and $S_|$ on $\mathbb{R}^d \setminus B(0,R)$. Note that $S(x) \in B(0,a)$ for $x \in B(0,a) \setminus B(0,R)$; thus, there exists $\widetilde{C}$ such that $\|S(x)\| \le \widetilde{C}\|x\|$ (if $a < 2R$, then $\widetilde{C} \le 2$). For the following calculation, we assume that

$$\begin{aligned} f_\mu(x) &\le \left(\frac{\eta}{\widetilde{C}'_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{\|x\|^{n+2}} \\ &:= \left(\frac{\eta}{C_{n,p,\Omega,M_p}\left((\widetilde{C}+1)C_{\text{sphere}}\right)^{\frac{p}{6p+16n}}}\right)^{\frac{6p+16n}{p}} \frac{1}{\|x\|^{n+2}} \\ &= \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{(\widetilde{C}+1)C_{\text{sphere}}\|x\|^{n+2}}, \end{aligned}$$

where $C_{\text{sphere}}$ denotes a constant from integrating over concentric $n$-spheres and $C_{n,p,\Omega,M_p}$ denotes the constant from Theorem 5.1. Now note that

$$\begin{aligned} W_1(\mu,\widetilde{\mu}) &\le \int_{\mathbb{R}^d} \|S(x) - x\| d\mu(x) = \int_{B(0,R)} \underbrace{\|x - x\|}_{=0} d\mu(x) + \int_{\mathbb{R}^d\setminus B(0,R)} \|S(x) - x\| d\mu(x) \\ &\le \int_{\mathbb{R}^d\setminus B(0,R)} \|S(x)\| + \|x\| d\mu(x) \le \int_{\mathbb{R}^d\setminus B(0,R)} (\widetilde{C}+1)\|x\| f_\mu(x)dx \\ &\le \int_{\mathbb{R}^d\setminus B(0,R)} (\widetilde{C}+1)\left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{(\widetilde{C}+1)C_{\text{sphere}}\|x\|^{n+1}} dx \\ &\le \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \underbrace{\int_{r\ge R} \frac{r^{n-1}}{r^{n+1}} dr}_{\le 1} \le \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}}. \end{aligned}$$

Invoking Theorem 5.1, this means that

$$\|T_\sigma^\mu - T_\sigma^{\widetilde{\mu}}\|_\sigma \le C_{n,p,\Omega,M_p} W_1(\mu,\widetilde{\mu})^{\frac{p}{6p+16n}} \le C_{n,p,\Omega,M_p} \frac{\eta}{C_{n,p,\Omega,M_p}} = \eta.$$

Thus, we have the desired result. □

## Appendix D. Proofs and results for conditions on $\mathscr{K}$ and $\mu$

This section provides the proofs of the results in Sections 6 and 7.

### D.1. Compact case proofs and results

Here we prove the results of Section 6 which provide conditions on $\sigma$, $\mu$, and $\mathscr{K}$ which guarantee that $\mu_i \sim \mathscr{K}_\sharp \mu$ satisfy the conditions of the theorems from Section 4.

**Proof of Theorem 6.1.** Caffarelli's regularity theorem implies that $T_\sigma^{(h_i)_\sharp\mu}$ is continuous, hence Lipschitz (since supp($\mu$) is compact). To show that Caffarelli's theorem applies, $(h_i)_\sharp\mu$ needs convex support and its density needs to be bounded away from 0 and $\infty$. One of the assumptions in this theorem is that $(h_i)_\sharp\mu$ has convex support; thus we must show the density is bounded away from 0 and $\infty$. If $h_i$ is continuously differentiable with $0I \prec L_1 I \preceq J_{h_i}(x)$, then the minimum eigenvalue of $J_{h_i}$ is bounded away from 0. Noticing that $h_i$ is a proper map, we can use Hadamard's global inverse function theorem to see that $h_i^{-1}$ exists. Recalling $\mu$ has density $f_\mu$, we can use the change of variables density formula

$$f_{(h_i)_\sharp\mu}(x) = f_\mu(h_i^{-1}(x))|J_{h_i^{-1}}(x)|.$$

Since $\mu$ has a density such that $0 < c \le f_\mu \le C < \infty$, we can see that $f_{(h_i)_\sharp\mu}$ is bounded away from 0 and $\infty$ by using the change of variables density formula. In particular, note that $f_{(h_i)_\sharp\mu} \ge cL_1 > 0$. For the upper-bound, we need that $h_i$ is Lipschitz. This is immediate since $h_i$ is a continuously differentiable (hence continuous) map over a compact set; thus, $h_i$ is Lipschitz with some Lipschitz constant $L_2$. This means that $f_{(h_i)_\sharp\mu}(x) \le CL_2 < \infty$. □

**Proof of Theorem 6.2.** For the barycentric map estimator, we already showed that the $\mu_i$'s are compactly supported in a ball of radius $AR$ in the proof of Theorem 4.2. We now show that $T_\sigma^{\mu_i}$ is Lipschitz. To make sure that each $T_\sigma^{\mu_i}$ is Lipschitz, we will only need that $h_i$ is continuous because continuous maps over compact sets are Lipschitz. In particular, note that $\mu_i = (h_i)_\sharp\mu$ for some $h_i \in \mathscr{H}$. Moreover, notice that we only care about how $h_i$ acts on supp($\mu$), which is compact; thus, $h_i$ is Lipschitz on the set of interest. Now, by compatibility, we know that $T_\sigma^{\mu_i} = h_i \circ T_\sigma^{\mu}$, which implies that if $h_i$ is Lipschitz and $T_\sigma^{\mu}$ is Lipschitz, then $T_\sigma^{\mu_i}$ is Lipschitz. □

**Proof of Theorem 6.3.** For the entropic map estimator, the $\mu_i$'s need to again be compactly supported, $T_\sigma^{\mu_i}$ needs to be Lipschitz, and $\sigma$ and $\mu_i$ together satisfy assumptions $(A1) - (A3)$. It will turn out, that we will only need that there exist constants $L_1, L_2 > 0$ such that

$$L_1 I \preceq J_h(x) \preceq L_2 I.$$

This occurs if $h$ is continuous, compatible, and has lower bound on Jacobian as $L_1 I \preceq J_h(x)$. To see the upper bound, we just see that $h$ being a continuous map on a compact set (support of $\mu$) gives that on the support of $\mu$, we have $J_h(x) \preceq L_2 I$ for some $L_2$. Now similar to Remark 4.3, we can use Hadamard's global inverse function theorem to see that $h^{-1}$ exists with $L_2^{-1} I \preceq J_{h^{-1}}(x) \preceq L_1^{-1} I$. Since we only sample a finite number of measure-valued data points $\mu_i$, we know that there must be a maximum $L_2$ that applies for all $\mu_i$.

That $\mu_i$ is compactly supported and each $T_\sigma^{\mu_i}$ are Lipschitz follow from the same analysis as in the proof of Theorem 6.2.

- **Ensuring that $\mu_i$ satisfy $(A1)$:** Recall that the change of variables formula for the density of a pushforward measure $\widetilde{\mu} = h_\sharp\mu$ is given by

  $$f_{\widetilde{\mu}}(x) = f_\mu(h^{-1}(x))|J_{h^{-1}}(x)|,$$

  where $|J_{h^{-1}}(x)|$ denotes the determinant of the Jacobian of $h^{-1}$.
  From the discussion above, we get that $|J_{h^{-1}}| > 0$ for all $x$. In particular, since the determinant of a matrix is the product of its eigenvalues, we have that

  $$L_2^{-n} \le |J_{h^{-1}}(x)| = \prod_{j=1}^{n} \lambda_j(J_{h^{-1}}(x)) \le L_1^{-n}.$$

  Finally, since $\mu$ itself adheres to (A1), this implies that

  $$\frac{b}{L_2^n} \le f_\mu(x)|J_{h^{-1}}(x)| \le \frac{B}{L_1^n}.$$

  So $(A1)$ holds for $\widetilde{\mu}$ if there are constants $L_1, L_2 > 0$ such that

  $$L_1 I \preceq J_h(x) \preceq L_2 I.$$

- **Ensuring that $\mu_i$ satisfy $(A2)$:** From [22, Corollary 4.2.10], we can ensure that $(A2)$ is satisfied if $(A3)$ is satisfied, which is proved below.
- **Ensuring that $\mu_i$ satisfy $(A3)$:** First, notice that by compatibility of $h$, we have that $T_\sigma^{h_\sharp\mu} = h \circ T_\sigma^{\mu}$; thus, a direct corollary of [25, Theorem 24] gives that

  $$(mL_1)I \preceq J_{T_\sigma^{h_\sharp\mu}}(x) \preceq (L_2 L)I$$

  for all $x$, where $m$ and $L$ come from assuming $\sigma$ and $\mu$ satisfy (A3) whilst $L_1$ and $L_2$ come from above. So $(A3)$ holds for $\sigma$ and $\widetilde{\mu}$. □

The result above essentially states that the entropic estimator works if every $h \in \mathscr{H}$ is (exactly) compatible and is uniformly positive definite.

### D.2. Non-compact case proofs and results

Here we prove the results of Section 7 which provide conditions on $\sigma$, $\mu$, and $\mathscr{H}$ which guarantee that $\mu_i \sim \mathscr{H}\sharp\mu$ satisfy the conditions of the theorems from Section 5.

**Proof of Theorem 7.1.** Assume that $\widetilde{\mu}$ is the truncated measure approximating $h_\sharp\mu$ for $h \in \mathscr{H}$. Given the assumptions of Lemma 5.3, the truncated measure $\widetilde{\mu}$ is compactly supported, upper and lower bounded, and absolutely continuous. If we can ensure that the truncated measure $\widetilde{\mu}$ also has uniformly convex support, we will fulfill the conditions of Caffarelli's regularity theorem, which guarantees that the optimal transport map is Lipschitz continuous.

- **Decay rate condition:** Assuming that $\mu$ has the necessary decay rate $f_\mu(x) \leq C < \infty$ and $0 < c \leq f_\mu(x)$ on a large enough ball where the decay rate is active, we need that $h_\sharp\mu = \overline{\mu}$ also has the same decay rate up to a constant. For what follows, we must show that $h \in \mathscr{H}$ has an inverse $h^{-1}$. Indeed, because of the bi-Lipschitz assumption (iii) of Theorem 7.1, we know that

  $$L_1 I \preceq J_h(x) \preceq L_2 I.$$

  Because $J_h(x)$ is invertible for all $x$, we can Hadamard's inverse function theorem to conclude that $h^{-1}$ exists. Moreover, this implies that

  $$L_2^{-1} I \preceq J_{h^{-1}}(x) \preceq L_1^{-1} I.$$

  Since $\mathscr{H}$ satisfies Assumption 4.1 (iii) (i.e.

  $$a\|x\| \leq \|h(x)\| \leq A\|x\|$$

  for some $a, A > 0$), then we know that

  $$A^{-1}\|x\| \leq \|h^{-1}(x)\| \leq a^{-1}\|x\|,$$

  or equivalently,

  $$\frac{A^{-1}}{\|h^{-1}(x)\|} \leq \frac{1}{\|x\|} \leq \frac{a^{-1}}{\|h^{-1}(x)\|}.$$

  Thus, for $\|x\| \geq AR$ (so that $\|h^{-1}(x)\| \geq R$) and the bounds above, we find that

  $$\begin{aligned} f_{\overline{\mu}}(x) &= f_\mu(h^{-1}(x)) \underbrace{|J_{h^{-1}}(x)|}_{\leq L_1^{-n}} \\ &\leq \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{C'\|h^{-1}(x)\|^{n+2}} L_1^{-n} \\ &\leq \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{C'\|x\|^{n+2}} L_1^{-n} A^{n+2}. \end{aligned}$$

  The constants $L_1$ and $A$ can be absorbed into the other decay rate constants; thus, Assumption 4.1 (iii) and our bi-Lipschitz assumption (iii) gives us the decay rate we want. Noting that the form of the density $f_{\overline{\mu}}$ also implies that $cL_1^{-n} \leq f_{\overline{\mu}}(x)$ on some large enough ball. In particular, we get that the truncated measure $\widetilde{\mu}$ has a density $0 < b \leq f_{\widetilde{\mu}}(x) \leq B < \infty$ from Lemma 5.3.
- **Uniformly convex support:** If $\mu$ is supported on all of $\mathbb{R}^n$, we would want $h \in \mathscr{H}$ such that $\overline{\mu} = h_\sharp\mu$ is also supported on all of $\mathbb{R}^n$. Recall that the resulting density of $\overline{\mu}$ is given by

  $$f_{\overline{\mu}}(x) = f_\mu(h^{-1}(x)) \underbrace{|J_{h^{-1}}(x)|}_{\leq L_1^{-n}}$$

  Note that $\overline{\mu}$ is supported on all of $\mathbb{R}^n$ if $\|h^{-1}(x)\| \to \infty$ as $\|x\| \to \infty$. Indeed, if we assume Assumption 4.1 (iv), then $A^{-1}\|x\| \leq \|h^{-1}(x)\|$, which implies that $\overline{\mu}$ is supported on all of $\mathbb{R}^n$. This would imply that the truncated measure $\widetilde{\mu}$ will be supported on a ball of some radius. This implies that the support of $\widetilde{\mu}$ is uniformly convex and compact.

From the decay rate condition and the uniformly convex support condition, we get that the truncated measure $\widetilde{\mu}$ will satisfy the assumptions of Caffarelli's regularity theorem. This implies that $T_\sigma^{\widetilde{\mu}}$ will be a $C^2$ and Lipschitz function (since $T_\sigma^{\widetilde{\mu}}$ pushes forward a compact support to a compact support). The other assumptions of the theorem are trivially satisfied. □

**Proof of Theorem 7.2.** From the proof of Theorem 7.1 above, we easily see that if Assumption 4.1 is fulfilled and $\mu$ fulfills the conditions of Lemma 5.3 and is supported on all of $\mathbb{R}^n$, then $T_\sigma^{\widetilde{\mu}}$ will be Lipschitz. We need, however, that $\widetilde{\mu}$ also satisfies $(A1)$-$(A3)$ from 4.6. We get $(A1)$ for free since the density $f_{\widetilde{\mu}}$ is lower bounded from the proof of Lemma 5.3. We also get $(A2)$ since $T_\sigma^{\widetilde{\mu}}$ is differentiable from Caffarelli's regularity theorem [11–13] and if (A3) is satisfied, which comes from [22, Corollary 4.2.10].

Now we only need to ensure that $(A3)$ holds. Indeed, since Caffarelli's regularity theorem holds, we know that the potential $\phi$ such that $T_\sigma^{\widetilde{\mu}} = \nabla\phi$ is strictly convex, which implies that $\nabla^2\phi(x)$ is positive definite. Moreover, the minimum eigenvalue of $\nabla^2\phi(x)$ is a continuous function of $x$. Since $x \in \text{supp}(\sigma)$, which is compact, we know that $0 < \lambda_{\min}(\sigma) = \min_{x\in\text{supp}(\sigma)} \lambda_{\min}(\nabla^2\phi(x))$, which implies that $J_{T_\sigma^{\widetilde{\mu}}}(x) \succeq \lambda_{\min}(\sigma) I$. This guarantees that $(A3)$ is satisfied for $\sigma$ and $\widetilde{\mu}$. □

## Data availability

No data was used for the research described in the article.

## References


[1] Akram Aldroubi, Shiying Li, Gustavo K. Rohde, Partitioning signal classes using transport transforms for data analysis and machine learning, Sampl. Theory Signal Process. Data Anal. 19 (1) (2021) 1–25.
[2] Jason Altschuler, Jonathan Weed, Philippe Rigollet, Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration, Adv. Neural Inf. Process. Syst. 2017-December (2017) 1965–1975.
[3] Luigi Ambrosio, Nicola Gigli, A user's guide to optimal transport, in: Modelling and Optimisation of Flows on Networks, Springer, 2013, pp. 1–155.
[4] Luigi Ambrosio, Nicola Gigli, Giuseppe Savaré, Gradient Flows: In Metric Spaces and in the Space of Probability Measures, Springer Science & Business Media, 2008.
[5] Ery Arias-Castro, Adel Javanmard, Bruno Pelletier, Perturbation bounds for Procrustes, classical scaling, and trilateration, with applications to manifold learning, J. Mach. Learn. Res. 21 (2020).
[6] Martin Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein generative adversarial networks, in: Doina Precup, Yee Whye Teh (Eds.), Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 214–223.
[7] Saurav Basu, Soheil Kolouri, Gustavo K. Rohde, Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry, Proc. Natl. Acad. Sci. 111 (9) (2014) 3448–3453.
[8] Mikhail Belkin, Partha Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[9] R.J. Berman, Convergence rates for discretized Monge–Ampère equations and quantitative stability of optimal transport, Found. Comput. Math. 21 (2021) 1099–1140.
[10] Yann Brenier, Polar factorization and monotone rearrangement of vector-valued functions, Commun. Pure Appl. Math. 44 (4) (1991) 375–417.
[11] Luis A. Caffarelli, Boundary regularity of maps with convex potentials, Commun. Pure Appl. Math. 45 (9) (1992) 1141–1151.
[12] Luis A. Caffarelli, The regularity of mappings with a convex potential, J. Am. Math. Soc. 5 (1) (1992) 99–104.
[13] Luis A. Caffarelli, Boundary regularity of maps with convex potentials–II, Ann. Math. 144 (3) (1996) 453–496.
[14] Yongxin Chen, Filemon Dela Cruz, Romeil Sandhu, Andrew L. Kung, Prabhjot Mundi, Joseph O. Deasy, Allen Tannenbaum, Pediatric sarcoma data forms a unique cluster measured via the earth mover's distance, Sci. Rep. 7 (1) (2017) 7035.
[15] Ronald R. Coifman, Stéphane Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30.
[16] Michael AA Cox, Trevor F. Cox, Multidimensional scaling, in: Handbook of Data Visualization, Springer, 2008, pp. 315–347.
[17] Marco Cuturi, Sinkhorn distances: lightspeed computation of optimal transport, in: NIPS, vol. 2, 2013, p. 4.
[18] Nabarun Deb, Promit Ghosal, Bodhisattva Sen, Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections, Adv. Neural Inf. Process. Syst. 34 (2021) 29736–29753.
[19] Alex Delalande, Quentin Mérigot, Quantitative stability of optimal transport maps under variations of the target measure, arXiv preprint, arXiv:2103.05934, 2021.
[20] Nicola Gigli, On Hölder continuity-in-time of the optimal transport map towards measures along a curve, Proc. Edinb. Math. Soc. 54 (2) (2011) 401–409.
[21] Keaton Hamm, Nick Henscheid, Shujie Kang, Wassmap: Wasserstein isometric mapping for image manifold learning, arXiv preprint, arXiv:2204.06645, 2022.
[22] Jean-Baptiste Hiriart-Urruty, Claude Lemaréchal, Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods, Springer Verlag, 1996.
[23] James M. Joyce, Kullback-Leibler divergence, in: International Encyclopedia of Statistical Science, Springer, 2011, pp. 720–722.
[24] Marc Khoury, Yifan Hu, Shankar Krishnan, Carlos Scheidegger, Drawing large graphs by low-rank stress majorization, in: Computer Graphics Forum, vol. 31, Wiley Online Library, 2012, pp. 975–984.
[25] Varun Khurana, Harish Kannan, Alexander Cloninger, Caroline Moosmüller, Supervised learning of sheared distributions using linearized optimal transport, Sampl. Theory Signal Process. Data Anal. 21 (1) (2023).
[26] Laurens van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.
[27] Kantilal Varichand Mardia, Multivariate analysis, Technical report, 1979.
[28] James Mathews, Maryam Pouryahya, Caroline Moosmüller, Ioannis G. Kevrekidis, Joseph O. Deasy, Allen Tannenbaum, Molecular phenotyping using networks, diffusion, and topology: soft-tissue sarcoma, Sci. Rep. 9 (2019) 13982.
[29] Quentin Mérigot, Alex Delalande, Frédéric Chazal, Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space, in: Silvia Chiappa, Roberto Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 26–28 Aug 2020, in: Proceedings of Machine Learning Research, vol. 108, PMLR, 2020, pp. 3186–3196.
[30] Jacob Miller, Vahan Huroyan, Stephen Kobourov, Spherical graph drawing by multi-dimensional scaling, arXiv preprint, arXiv:2209.00191, 2022.
[31] Gal Mishne, Ronen Talmon, Ron Meir, Jackie Schiller, Maria Lavzin, Uri Dubin, Ronald R. Coifman, Hierarchical coupled-geometry analysis for neuronal structure and activity pattern discovery, IEEE J. Sel. Top. Signal Process. 10 (7) (2016) 1238–1253.
[32] Caroline Moosmüller, Alexander Cloninger, Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations, Inf. Inference 12 (1) (2023) 363–389.
[33] Marshall Mueller, Shuchin Aeron, James M. Murphy, Abiy Tasissa, Geometric sparse coding in Wasserstein space, arXiv preprint, arXiv:2210.12135, 2022.
[34] Se Rim Park, Soheil Kolouri, Shinjini Kundu, Gustavo K. Rohde, The cumulative distribution transform and linear pattern classification, Appl. Comput. Harmon. Anal. 45 (3) (2018) 616–641.
[35] Aram-Alexandre Pooladian, Jonathan Niles-Weed, Entropic estimation of optimal transport maps, arXiv:2109.12004, 2021.
[36] Yossi Rubner, Carlo Tomasi, Leonidas J. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. 40 (2) (2000) 99–121.

[37] Justin Solomon, Raif Rustamov, Leonidas Guibas, Adrian Butscher, Wasserstein propagation for semi-supervised learning, in: International Conference on Machine Learning, 2014, pp. 306–314.
[38] Joshua B. Tenenbaum, Vin De Silva, John C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[39] Cédric Villani, Optimal Transport: Old and New, vol. 338, Springer Science & Business Media, 2008.
[40] Wei Wang, John A. Ozolek, Dejan Slepčev, Ann B. Lee, Cheng Chen, Gustavo K. Rohde, An optimal transportation approach for nuclear structure-based pathology, IEEE Trans. Med. Imaging 30 (3) (2010) 621–631.
[41] Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, Gustavo K. Rohde, A linear optimal transportation framework for quantifying and visualizing variations in sets of images, Int. J. Comput. Vis. 101 (2013) 254–269.
[42] M.E. Werenski, R. Jiang, A. Tasissa, S. Aeron, J.M. Murphy, Measure estimation in the barycentric coding model, in: Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022, pp. 23781–23803.
[43] Gale Young, Aiston S. Householder, Discussion of a set of points in terms of their mutual distances, Psychometrika 3 (1) (1938) 19–22.
[44] Nathan Zelesko, Amit Moscovich, Joe Kileel, Amit Singer, Earthmover-based manifold learning for analyzing molecular conformation spaces, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1715–1719.
[45] Yin Zhang, Rong Jin, Zhi-Hua Zhou, Understanding bag-of-words model: a statistical framework, Int. J. Mach. Learn. Cybern. 1 (1–4) (2010) 43–52.