UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning on the Space of Probability Measures

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics

by

Varun Khurana

Committee in charge:

Professor Alex Cloninger, Chair
Professor Mikhail Belkin
Professor Jelena Bradic
Professor Gal Mishne
Professor Ruth J. Williams

2024

The dissertation of Varun Khurana is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my mother, father, brother, sister, and grandmother.

EPIGRAPH

A mind that is full of conclusions is a dead mind, it is not a living mind.

A living mind is a free mind, learning, never concluding.

—Jiddu Krishnamurti

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks to Mesa Rim Climbing Centers, the beach, park, and my lovely friends and family for providing the opportunity to clear my head when I felt stuck in my research and for bringing balance to my life. Thanks also to all the coffee shops in San Diego.

Chapter 2, in full, is a reprint of *Supervised learning of sheared distributions using linearized optimal transport* as it appears in Sampl. Theory Signal Process. Data Anal. 21, 1 (2023). `https://doi.org/10.1007/s43670-022-00038-2` by Khurana, V., Kannan, H., Cloninger, A. et al. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of *Linearized Wasserstein dimensionality reduction with approximation guarantees* by Cloninger, A., Hamm, K., Khurana, V., and Moosmüller, C. February 14, 2023. as it appears in `https://arxiv.org/abs/2302.07373`. The paper is currently under revision for the Applied Computational Harmonic Analysis (ACHA) Journal. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of *Lattice approximations in wasserstein space* (2023) by K. Hamm and V. Khurana as it appears in `https://arxiv.org/abs/2310.09149`. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of *A Neural Network Two-Sample Test* (2024) by V. Khurana, A. Cloninger, and X. Cheng as it is due to appear. The dissertation author was the primary investigator and author of this paper.

VITA

2019	B. E. in Mechanical Engineering, University of California Berkeley
2019	B. S. in Mathematics *Honors*, University of California Berkeley
2021	M.A. in Mathematics, University of California San Diego
2024	Ph. D. in Mathematics, University of California San Diego

PUBLICATIONS

Khurana, V., Kannan, H., Cloninger, A. et al. Supervised learning of sheared distributions using linearized optimal transport. Sampl. Theory Signal Process. Data Anal. 21, 1 (2023). https://doi.org/10.1007/s43670-022-00038-2

Cloninger, A., Hamm, K., Khurana, V., and Moosmüller, C. February 14, 2023. Linearized Wasserstein dimensionality reduction with approximation guarantees. https://arxiv.org/abs/2302.07373

K. Hamm and V. Khurana. Lattice approximations in wasserstein space, 2023. https://arxiv.org/abs/2310.09149

ABSTRACT OF THE DISSERTATION

Learning on the Space of Probability Measures

by

Varun Khurana

Doctor of Philosophy in Mathematics

University of California San Diego, 2024

Professor Alex Cloninger, Chair

This thesis studies computationally feasible machine learning methods, based on optimal transport and neural network theory, applied to measure-valued data. We first analyze linearized optimal transport (LOT), which essentially embeds measure-valued data into an L^2 space, where out-of-the-box machine learning techniques are available. We analyze the situations when LOT provides an isometric embedding with respect to the Wasserstein-2 distance and provide necessary bounds when we can achieve a pre-specified linear separation level in the LOT embedding space. Second, we produce a computationally feasible algorithm to recover low-dimensional structures in measure-valued data by using the LOT embedding along with dimensionality reduction techniques. Using computational methods for solving optimal transport problems such as the

Sinkhorn algorithm or linear programming, we provide approximation guarantees in terms of the sampling rates. Third, we study structured approximations of measures in Wasserstein space by a scaled Voronoi partition of $\mathbb{R}^d$ generated from a full rank lattice. We show that these structured approximations match rates of optimal quantizers and empirical measure approximation in most instances. We then extend these results for noncompactly supported measures that decay fast enough. Finally, we study methods for comparing probability measures by analyzing a neural network two-sample test. In particular, we perform time-analysis on a related neural tangent kernel (NTK) two-sample test and extend the analysis to the neural network two-sample test with a small-time training regime. We also show the amount of time needed before the two-sample test detects a deviation $\varepsilon > 0$ in the case the probability measures considered are different versus when they are the same.

Chapter 1

Introduction

This thesis considers studying measure-valued data and showcases a variety of techniques for studying measures or probability distributions. In this chapter, we discuss motivation for such theory, a preliminary background for some of the ideas used in the rest of the thesis, and general overview of the rest of the thesis.

1.1 Motivation

In practice, the structure of data for analysis depends on the domain studied and range anywhere from cases where each data point is an n-dimensional vector to cases where each data point is a time-series of mathematical objects. For our case, we study measure-valued data. In particular, each observed data point can be thought of as a measure or a finite sample from a measure. This type of data practically arises as point clouds of information so that each data point is a point cloud of possibly different sizes. Situations encountering this type of data have become increasingly frequent in practical applications.

For example, biologists often consider a bulk samples of cells from a specimen as one data point so that each data point is indeed thought of as a point cloud [19, 29, 101]. In image analysis, the pixels in each image can be normalized so that each image yields a probability measure for each of the RGB values on the pixel grid [59, 78, 71, 93]; thus, each image can be thought of as a probability density. In neural network training and analysis, the hidden representations of data in neural networks can be thought of as point clouds approximating some measure and evolution of those point clouds is essentially a time-series of measures. Viewing neural network training as a flow of measures in the hidden representation space provides a gamut of tools at the disposal of the data scientist.

At the core of all the data analysis examples mentioned is the fundamental problem of comparing point clouds, or more generally, probability measures. The traditional methods of comparing probability measures such as the popular Kullback-Leibler (KL) divergence [60] as well as Maximum Mean Discrepancy (MMD) [47] come with their fair share of difficulties. In particular, KL-divergence blows up to $+\infty$ when the measures compared have non-overlapping support. On the other hand, since kernel MMD is the norm of the difference of mean embeddings (with approximately norm 1), comparing measures with disjoint support leads to a saturation level of about $\sqrt{2}$. Moreover, kernel MMD depends on the associated reproducing kernel Hilbert space and lacks some amount of interpretability useful for analysis. Optimal transport (OT) has risen as a particularly powerful method for comparing probability measures and providing some strong interpretability since the distance metric using optimal transport naturally gives rise to a geometry for the space of probability measures. Once a method for comparing measures is established, we can focus on how to solve run-of-the-mill machine learning and statistics problems such as classification, regression, unsupervised dimensionality reduction, and two-sample testing on the space of probability measures. The next section will discuss some fundamentals of OT and two-sample testing tools and provide the necessary framework to discuss the overview of the rest of the thesis.

1.2 Background

1.2.1 Optimal Transport

Gaining significant importance in recent years, optimal transport (OT) arises as the most natural methodology for computing distance between measures [91]. The central problem that OT solves is finding methods to transport the mass of one probability measure to another probability measure. This can be done with a transport map when the associated measures do not need to be split or with a transport plan which allows for mass splitting. With this in mind, the optimal transport distance between two probability measures is calculated by finding the transport map or transport plan that minimizes a transportation cost, which is usually the distance metric on the underlying space.

More rigorously, let $\mathcal{P}_2(X)$ denote the set of all probability measures on a metric space X with finite second moment. Given $\mu, \nu \in \mathcal{P}_2(X)$, a transport plan that transports μ to ν is given by a product measure γ with marginals μ and ν. We denote the set of all transport plans for μ and ν by $\Gamma(\mu,\nu)$. The Wasserstein-p distance between measures μ and ν is now given by

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_X d(x,y)^p d\gamma(x,y) \right)^{1/p}.$$

Under regularity assumptions, solving the optimal transport problem yields an optimal transport (OT) map T_μ^ν so that the minimizing product measure is of the form $(\mu, T_{\mu\,\sharp}^\nu \mu)$ where $T_{\mu\,\sharp}^\nu \mu(A) = \mu({T_\mu^\nu}^{-1}(A)) = \nu(A)$ [17]. We will equivalently denote the space of measures with the W_p-distance metric as the Wasserstein-p space and just say “Wasserstein” space when $p = 2$. Building classifiers and regressors for measure-valued data with finite samples will become simple after a transformation of measures, called *Linearized Optimal Transport (LOT)* or *Cumulative Distribution Transform (CDT)* [1], to a space that is already amenable to machine learning since the space of measures is not conducive to standard machine learning techniques and algorithms due to its nonlinearity.

1.2.2 Two-Sample Tests

A more simple question when comparing two distributions is essentially whether they are the same or not, and a traditional method to solve this problem is a two-sample hypothesis test. We will consider a neural network two-sample test in Chapter 5, but to give a brief overview, let us cover what the two-sample test exactly is. Assume that you are given two datasets, say P and Q, and you want to test whether or not these datasets came from the same probability measure or not. In particular, we want to assess whether to accept the null hypothesis H_0 or reject it for the alternative hypothesis H_1, where

$$H_0 : p = q, \qquad\qquad H_1 : p \neq q,$$

with p being the density generating samples for P and q being the density generating samples for Q. Given an estimator f that trains on training datasets with labels, say 1 and -1 respectively, a two-sample test can be constructed as

$$\mu_P - \mu_Q = \big(\mathbb{E}_{x\sim P} - \mathbb{E}_{x\sim Q}\big) f(x).$$

Given a threshold $\tau > 0$, we reject the null for the alternative if $|\mu_P - \mu_Q| > \tau$. Initially, [62] showed properties and analyzed performance of the so-called Classifier Two-Sample Test (C2ST) and specifically showcased theoretically what the statistical power of such two-sample tests. Soon, [47] developed two-sample tests corresponding to kernels, which was further expanded to neural tangent kernels by [30].

1.3 Overview

The thesis will be split into four chapters, not including this one, covering the following material.

1. **Chapter 2**: This chapter is based on the paper [57], where the author of this dissertation is the main author. Given a reference measure $\sigma \in \mathcal{P}_2(\mathbb{R}^d)$ and a target measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the author essentially studies the "Linearized Optimal Transport" (LOT) embedding, where the optimal transport map T_σ^μ, discussed in Section 1.2, embeds probability measures into an L^2-space. Using the regular L^2-distance in this space, the author discusses when the L^2-distance equals the W_2-distance on probability measures so that the LOT embedding is an isometry. Additionally, the author shows necessary conditions on a dataset of measures that will ensure linear separability in the LOT embedding space with a pre-specified separation level. Finally, the author uses multiple reference distributions to produce better separation guarantees.

2. **Chapter 3**: This chapter is based on the paper [32], where the author of this dissertation is the main author. Using the framework from Chapter 2, the author applies the LOT embedding to introduce LOT Wassmap, a computationally feasible algorithm to approximate low-dimensional structures in the Wasserstein space using manifold learning algorithms, the Sinkhorn algorithm, and LOT embeddings. This algorithm avoids computing a pairwise distance matrix and significantly reduces computational cost. Moreover, the PI provided 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. These approximations are guaranteed by showing that pairwise distances of estimated optimal transport maps converge to the true optimal transport map with rates depending on the sample sizes of the measures involved.

3. **Chapter 4**: This chapter is based on (insert reference here), where the author of this

dissertation is a co-author. This chapter considers structured approximation of measures in Wasserstein space $W_p(\mathbb{R}^d)$ for $p \in [1,\infty)$ by discrete and piecewise constant measures based on a scaled Voronoi partition of $\mathbb{R}^d$. We show that if a full rank lattice Λ is scaled by a factor of $h \in (0,1]$, then approximation of a measure based on the Voronoi partition of $h\Lambda$ is $O(h)$ regardless of d or p. We then use a covering argument to show that N-term approximations of compactly supported measures is $O(N^{-\frac{1}{d}})$ which matches known rates for optimal quantizers and empirical measure approximation in most instances. Finally, we extend these results to noncompactly supported measures with sufficient decay.

4. **Chapter 5**: This chapter is based on (insert reference here), where the author of this dissertation is the main author of the paper. This chapter constructs a neural network two-sample test and analyzes the behavior of the test in terms of training time. In particular, we approximate the finite-sample neural network dynamics and population-level neural network dynamics with the zero-time neural tangent kernel (NTK) population-level dynamics. The zero-time NTK two-sample test grows as a function of the zero-time kernel eigenvalues and projection of an approximated function on the associated eigenfunctions. We relate these dynamics to the finite-sample neural network two-sample test as well as the populationlevel neural network two-sample test.

Chapter 2

Linearized Optimal Transport

Joint work with Harish Kannan, Caroline Moosmüller, Alex Cloninger

In this chapter, we study supervised learning tasks on the space of probability measures. We approach this problem by embedding the space of probability measures into L^2 spaces using the optimal transport framework. In the embedding spaces, regular machine learning techniques are used to achieve linear separability. This idea has proved successful in applications and when the classes to be separated are generated by shifts and scalings of a fixed measure. This paper extends the class of elementary transformations suitable for the framework to families of shearings, describing conditions under which two classes of sheared distributions can be linearly separated. We furthermore give necessary bounds on the transformations to achieve a pre-specified separation level, and show how multiple embeddings can be used to allow for larger families of transformations. We demonstrate our results on image classification tasks.

2.1 Introduction

We consider the problem of classifying probability measures μ_i on $\mathbb{R}^n$ based on a finite set of pre-classified training data $\{(\mu_i, y_i)\}_{i=1}^N$, where y_i denote the labels. The aim is to use the given training data to build a function f that assigns a probability measure to its correct label, i.e. we study supervised learning techniques on the space of probability measures.

The problem of classifying probability measures rather than points in $\mathbb{R}^n$ has a number of applications, a few examples being classification of population groups [33], and classification of flow cytometry and other measurements of cell or gene populations per person [19, 29, 101]. Note that for application purposes, we need to consider samples of probability measures μ_i, hence the task requires one to meaningfully compare and classify point clouds.

The largest issue associated with this classification problem is the generation of features of μ_i that can be used to build a classifier f. Many methods use an embedding idea to transform the set of probability measures into a Hilbert space in which regular machine learning techniques can be applied for the classification task, e.g. embeddings through moments or kernels [73, 76].

In this paper we are interested in such embeddings based on the optimal transport framework [91]. Optimal transport gives rise to a natural distance on the space of probability measures via the Wasserstein distance, which quantifies the minimal work necessary to move one distribution into another using an optimal transport plan. Optimal transport has gained high interest in the machine learning community in recent years, for example for generative models, semi-supervised learning or imaging applications [10, 82, 85].

We use the optimal transport plan or map to build an embedding of probability measures into an L^2-space known as "Linear Optimal Transportation" (LOT) [94, 78, 1, 71, 45] or "Monge embedding" [68]. LOT is a set of transformations based on optimal transport maps, which map a distribution μ to the optimal transport map that takes a fixed reference distribution σ to μ:

$$\mu \mapsto T_\sigma^\mu, \qquad \text{where } T_\sigma^\mu := \arg \min_{T \in \Pi_\sigma^\mu} \int \|T(x) - x\|_2^2 d\sigma(x), \tag{2.1}$$

where Π_σ^μ denotes the set of measure preserving maps from σ to μ. Through the embedding (2.1), the optimal transport map to a fixed reference σ is used as a feature of μ.

Note that LOT takes the manifold of probability measures into a Hilbert space of L^2 functions. This makes LOT particularly interesting as a feature space. Indeed, it has been demonstrated in various applications that within the LOT embedding space, classes of probability measures can be well separated with linear machine learning tools. The main applications concern signal and image classification tasks [59, 78, 71], such as distinguishing facial expressions, separating healthy from cancerous tissue classes [93], and visualizing phenotypic differences between types of cells [12].

While the LOT embedding space is well studied in 1-dimension [78], since LOT can be thought of as a generalized CDF, many questions remain open in higher dimensions. This has to do with the fact that in higher dimensions, there is a large family of potential group actions that can be applied to a distribution μ_i (e.g., shifts, scalings, shearings, rotations), and Π_σ^μ contains a large number of measure preserving maps.

It has been shown that shifts and scalings behave well with respect to the LOT embedding [1, 71, 78], meaning that two classes of probability measures obtained from scaling or shifting of a fixed measure can be linearly separated in the LOT embedding space. The reason lies in a property we refer to as the “compatibility condition”, which is satisfied by shifts and scalings [1, 71]. This property describes an interplay between LOT and the pushforward operator, or in terms of Riemannian geometry, the invertability of the exponential map [45]. Similarly, small perturbations of the distributions in these classes can still be linearly separated under certain minimal separation conditions [71].

The contributions of this paper are threefold. We first describe conditions under which families of shearings satisfy the compatibility condition, enlarging the space of functions for which linear classification results hold in the LOT embedding space (Section 2.3). The second contribution concerns binary classification results with pre-specified level of separation (Section 2.4). We give necessary bounds on the classes of probability measures to achieve linear separation in the embedding space with given separation level. The bounds are in terms of the parameters associated with the set of elementary transformations that are used to create the two classes. In the third part (Section 2.5), we study embeddings using multiple references. Based on the set of elementary transformations, we quantify the number references needed to achieve a desired separation level in the embedding space. The paper closes with classification experiments on sheared distributions.

2.2 Tools from optimal transport

This paper deals with probability measures on $\mathbb{R}^n$, i.e. with elements of the space $\mathcal{P}(\mathbb{R}^n)$. We mostly deal with probability measures that have bounded second moment, and denote the respective space by $\mathcal{P}_2(\mathbb{R}^n)$. The Lebesgue measure is denoted by λ.

To any probability measure σ, we assign the function space $L^2(\mathbb{R}^n, \sigma)$, which is equipped with the L^2-norm with respect to σ:

$$\|f\|_\sigma^2 = \int \|f(x)\|_2^2 \, d\sigma(x).$$

If a measure σ is absolutely continuous with respect to λ, written as $\sigma \ll \lambda$, then there exists a density $f_\sigma : \mathbb{R}^n \to \mathbb{R}$ such that

$$\sigma(A) = \int_A f_\sigma(x) d\lambda(x),$$

with $A \subseteq \mathbb{R}^n$ measurable. For the most part, the probability measures we consider are absolutely continuous with respect to λ.

A function $S : \mathbb{R}^n \to \mathbb{R}^n$ gives rise to the pushforward measure of σ:

$$S_\sharp \sigma(A) = \sigma(S^{-1}(A)). \tag{2.2}$$

where $A \subset \mathbb{R}^n$ measurable. Throughout this paper, we denote the Jacobian of a function S by J_S.

Given two measures, σ and μ there may exist many maps S such that $S_\sharp \sigma = \mu$. In order to find a unique map that pushes σ into μ, the theory of optimal transport [91] imposes an "optimality condition" on the map S. It has to minimize the overall cost of pushing σ into μ, where cost is measured by a metric in the underlying space (here we use the Euclidean distance in $\mathbb{R}^n$):

$$\int \|S(x) - x\|_2^2 d\sigma(x). \tag{2.3}$$

If such a cost minimizing function exists, then

$$W_2(\sigma, \mu)^2 = \min_{S : S_\sharp \sigma = \mu} \int \|S(x) - x\|_2^2 d\sigma(x). \tag{2.4}$$

is called the *Wasserstein-2* distance between σ and μ. Note that the Wasserstein problem can also be considered for different norms (like p-norm) and on Riemannian manifolds [17, 91, 67, 5].

Brenier's theorem [17] states that under the assumption of $\sigma \ll \lambda$, a unique map exists that pushes σ into μ and minimizes (2.3). We call this map "the optimal transport from σ to μ" and denote it by T_σ^μ.

We furthermore make use of the following result:

Theorem 2.1 (Brenier's theorem [17])**.** *If $\sigma \ll \lambda$, the optimal transport map T_σ^μ is uniquely defined as the gradient of a convex function φ, i.e. $T_\sigma^\mu(x) = \nabla\varphi(x)$, where φ is the unique convex function that satisfies $(\nabla\varphi)_\sharp \sigma = \mu$. Uniqueness of φ is up to an additive constant.*

2.2.1 Linear optimal transport embeddings

In this section, we introduce linear optimal transport embeddings, as proposed by [94, 78, 45]. A fixed *reference measure* σ gives rise to an embedding of $\mathcal{P}_2(\mathbb{R}^n)$ into $L^2(\mathbb{R}^n,\sigma)$ via the map

$$\mu \mapsto T_\sigma^\mu. \tag{2.5}$$

We denote this map by F_σ, and call it "LOT" or "LOT embedding" (sometimes F_σ is called a *Monge map* as well [68]). The LOT embedding can be very useful as a feature space to use linear machine learning techniques to classify subsets of $\mathcal{P}_2(\mathbb{R}^n)$ [71, 78]. Other fields of application include the approximation of the Wasserstein distance with a linear L^2-distance [71, 68], and fast barycenter computation and clustering [68].

From a theoretical point of view, the regularity of (2.5) has been studied in [68, 45]. Indeed, the Hölder regularity of (2.5) is not better than $1/2$. We also mention the results of [15], where a map related to LOT is analyzed, namely $\sigma \mapsto T_\sigma^\mu$.

A central property in the study of LOT is the so-called *compatibility condition* [71, 1]. It describes an interplay between LOT and the pushforward operator (4.1).

Definition 2.2. *Fix $\sigma,\mu \in \mathcal{P}_2(\mathbb{R}^n)$ with $\sigma \ll \lambda$. The LOT embedding F_σ is called* compatible *with the μ-pushforward of a function $S \in L^2(\mathbb{R}^n,\mu)$ if*

$$F_\sigma(S_\sharp\mu) = S \circ F_\sigma(\mu).$$

Note that the compatibility condition of Definition 4.2 can also be written as

$$T_\sigma^{S_\sharp\mu} = S \circ T_\sigma^\mu.$$

Considering the manifold $(\mathcal{P}_2(\mathbb{R}^n), W_2)$ with exponential map (the pushforward operator: $\exp_\sigma(S) = S_\sharp\sigma$) [45], LOT can be viewed as the exponential map's right-inverse (i.e. $\exp_\sigma \circ F_\sigma = \mathrm{id}$).

For $\sigma = \mu$, the compatibility condition forces LOT to be a left-inverse as well (i.e. $F_\sigma \circ \exp_\sigma = \mathrm{id}$).

Under the assumption of the compatibility condition, a series of interesting results can be derived. First, the Wasserstein-2 distance can be computed from the linear L^2-distance,

$$W_{2,\sigma}^{\mathrm{LOT}}(\mu_1,\mu_2) := \|F_\sigma(\mu_1) - F_\sigma(\mu_2)\|_\sigma \tag{2.6}$$

if μ_1, μ_2 have been obtained from a fixed *template* μ via pushforwards of two functions S_1, S_2 for which the compatibility condition holds [71], i.e. in this case

$$W_2(S_{1\sharp}\mu, S_{2\sharp}\mu) = W_{2,\sigma}^{\mathrm{LOT}}(S_{1\sharp}\mu, S_{2\sharp}\mu). \tag{2.7}$$

This is of particular interest when trying to compute the pairwise distance between many measures $\{\mu_i\}_{i=1,\dots,N}$, when each μ_i is obtained from a fixed template μ via the process $\mu_i = S_{i\sharp}\mu$ with compatible functions $\{S_i\}_{i=1,\dots,N}$ ([1] calls such a process an "algebraic generative model"). In this setting, one can compute the N transport maps $T_\sigma^{\mu_i}$, and then compute $\binom{N}{2}$ linear distances via (2.6), which is computationally much cheaper (especially for large N), than computing $\binom{N}{2}$ transport maps (Wasserstein-2 distances). These results also generalize to when the compatibility condition is only satisfied up to an error $\varepsilon > 0$ [71]. Then the linear distance (2.6) approximates W_2 up to an error of order $\varepsilon^{1/2}$. Other approximation results (that do not need the compatibility condition) can be found in [68].

Second, under the assumption of the compatibility condition, convexity is preserved under LOT [1, 71]. In particular, if $\mathcal{H} \subseteq L^2(\mathbb{R}^n, \sigma)$ is a set of convex and compatible functions with respect to σ and μ, then $F_\sigma(\mathcal{H} \star \mu)$ is also convex, where $\mathcal{H} \star \mu = \{h_\sharp\mu : h \in \mathcal{H}\}$ (a similar results holds for almost convex sets [71]). The preservation of convexity is crucial to deduce linear separability results in the embedding space through the Hahn-Banach theorem (e.g. to apply LOT in supervised learning). Indeed it has been shown that under the assumption of the compatibility condition, binary classification of sets of probability measures can be achieved in the LOT embedding space with linear methods, i.e. in the embedding space, a separating hyperplane can be found [71, 78].

Yet the compatibility condition (Definition 4.2) is very restrictive, and cannot be expected to hold for all S. As of now, it is known that shifts and scalings, i.e. functions of the form $S(x) = cx + b$ with $c > 0$ and $b \in \mathbb{R}^n$, satisfy Definition 4.2 for all choices of σ, μ [78, 1, 71]. For fixed σ, [1] also shows that for the compatibility condition to hold for all μ, S has to be a shift/scaling.

It is our aim to extend the set of compatible functions S beyond shifts and scalings to make LOT applicable to a broader range of applications. In particular we study (generalized) affine transformations. Note that because of the result in [1], to increase the set of compatible functions, the reference σ and the template μ can no longer be chosen independently. In the next section we establish necessary relationships between σ, μ and S for Definition 4.2 to hold.

2.3 Compatibility condition for affine transformations

In this section we study the conditions under which affine transformations $S(x) = Ax + b$ (and generalizations of such transformations) satisfy the compatibility condition (Definition 4.2). Our results show that fixing the reference σ and template μ generates necessary conditions for maps S to satisfy the compatibility conditions with respect to σ and μ. Conversely, fixing the template μ and the transformations S generates necessary conditions that references σ must satisfy in order for the compatibility condition to hold. These results strongly depend on the following theorem.

Theorem 2.3 (Informal Statement of Theorem 2.30)**.** *Let $\sigma, \mu \in \mathcal{P}_2(\mathbb{R}^n)$ and let $\sigma \ll \lambda$. Let $S \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ such that $S = \nabla\varphi$ for some twice differentiable function φ. We also assume that S satisfies the compatibility condition (Definition 4.2). Then the Jacobian of S, J_S, is symmetric* *positive definite and shares the same eigenspaces as the Jacobians of T_σ^μ and $T_\sigma^{S_\sharp\mu}$.*

Proof. The proof can be found in Section 2.7. □

We get the following corollary.

Corollary 2.4. *Let $\sigma,\mu \in \mathcal{P}_2(\mathbb{R}^n)$ and let $\sigma \ll \lambda$. If $S \in C^1(\mathbb{R}^n,\mathbb{R}^n)$ such that $S = \nabla\varphi$ for some twice differentiable φ and S satisfies the compatibility condition for σ and μ, then S is an optimal transport map.*

Proof. In particular, note that Theorem 2.3 states that if $S = \nabla\varphi$ for some φ; and if the compatibility condition holds, then $\nabla^2\varphi$ is positive definite. Thus, φ must have been convex. In light of Brenier's theorem Theorem 2.1, S must be an optimal transport map. Informally, Theorem 2.3 above states that this optimal transport map S must be transporting mass in the same directions (eigenspaces) as T_σ^μ. □

We use Theorem 2.3 above to extend a form of LOT isometry to the case when S is an affine transformation. The only caveat for our extension is that the orthonormal basis on which we shear must be constant. The relevant function class for this setting is given in the following definition.

Definition 2.5. *Given an orthogonal matrix $P \in \mathbb{R}^{n\times n}$, define the* constant orthonormal basis shears *as the class of maps*

$$\mathcal{F}(P) = \left\{ x \mapsto \tilde{P}^\top \begin{bmatrix} f_1((\tilde{P}x)_1) \\ f_2((\tilde{P}x)_2) \\ \vdots \\ f_n((\tilde{P}x)_n)) \end{bmatrix} + b : \begin{smallmatrix} f_j\text{:}\mathbb{R}\to\mathbb{R} \textit{ is monotonically} \\ \textit{increasing and differentiable} \\ \textit{and } b \in \mathbb{R}^n \end{smallmatrix} \right\},$$

where $\tilde{P}$ is a row-permutation of the orthogonal matrix P.

Note that affine transformations $S(x) = Ax + b$ with $A = P^T DP$ and $d_i > 0, i = 1, \dots, n$ (i.e. symmetric positive definite matrices diagonalizable by P), are elements of $\mathcal{F}(P)$. Indeed, choose $f_i(y) = d_i y, i = 1, \dots, n$.

Given a fixed template distribution μ, we show that demanding that the compatibility condition holds (under suitable conditions), if we fix either the reference distribution σ or the set of transformations, then the other (either the reference or transformations) can be fully characterized.

Fixed Reference and Template: Assume we fix the template distribution μ and reference distribution σ. If the Jacobian of $T_\sigma^\mu(x)$ has spectral decomposition $P^\top D(x)P$ for a constant orthogonal matrix P, then the set of compatible transformations can be fully characterized:

Theorem 2.6 (Conditions on transformations). *Let $\sigma, \mu \in \mathcal{P}_2(\mathbb{R}^n)$ with $\sigma \ll \lambda$. If the Jacobian of T_σ^μ has a constant orthonormal basis given by an orthogonal matrix P (i.e. $J_{T_\sigma^\mu}(x) = P^\top D(x)P$), then $\mathcal{F}(P)$ is the set of transformations for which the compatibility condition (Definition 4.2) holds.*

Proof. The proof of the theorem can be found in Section 2.7. □

Example 2.7 (Gaussians). *To illustrate Theorem 2.6, we provide a simple example with Gaussians. If both σ and μ are Gaussian distributions, for example $\mathcal{N}(m_1, I)$ and $\mathcal{N}(m_2, \Sigma_2)$, then*

$$T_\sigma^\mu(x) = m_2 + \Sigma_2^{1/2}(x - m_1),$$

and $J_{T_\sigma^\mu}(x) = \Sigma_2^{1/2}$. If Σ_2 is positive definite, then it can be decomposed as $P^T DP$. Therefore, Theorem 2.6 allows all generalized shears in Definition 2.5 that point in the same direction as Σ_2.

Fixed Shear and Template: Now we fix the transformation to be a type of generalized shear and the template distribution μ, and characterize the set of reference distributions such that compatibility condition holds.

Theorem 2.8 (Conditions on reference distribution)**.** *Let P be an orthogonal matrix, let $S(x) = P^\top g(Px) + b$ for $g(z) = \begin{bmatrix} g_1(z_1) & \dots & g_n(z_n) \end{bmatrix} : \mathbb{R}^n \to \mathbb{R}^n$ where $g_j : \mathbb{R} \to \mathbb{R}$ is differentiable and $b \in \mathbb{R}^n$, and let $\mu \in \mathcal{P}_2(\mathbb{R}^n)$ be a fixed template distribution with $\mu \ll \lambda$. Then $\Sigma = \{f_\sharp\mu : f \in \mathcal{F}(P)\}$ is the set of reference distributions such that the compatibility condition (Definition 4.2) holds.*

Proof. The proof can be found in Section 2.7. □

In Theorem 2.8, note that the reference distributions in Σ end up being absolutely continuous since they are the smooth pushforward of an absolutely continuous measure. Additionally, we get the following corollary.

Corollary 2.9. *Given the family of transformations of the form $S(x)$ from Theorem 2.8 above, i.e. for S in the set*

$$\left\{ S(x) = P^\top g(Px) + b : b \in \mathbb{R}^n, g_j : \mathbb{R} \to \mathbb{R} \text{ differentiable} \right\},$$

the set of reference distributions such that the compatibility condition holds for all of the transformations simultaneously is $\Sigma = \{f_\sharp\mu : f \in \mathcal{F}(P)\}$.

Proof. Inspecting the proof of Theorem 2.8, we see that the set of reference distributions Σ does not depend on the choice of functions $g_j : \mathbb{R} \to \mathbb{R}$ but rather only on P. □

Remark 2.10. *In Theorem 2.8, if we let $g_i(z_i) = d_i z_i$ for fixed d_i, then $S(x) = Ax + b$, where $A = P^\top DP$. Thus, for a template distribution $\mu \in \mathcal{P}_2(\mathbb{R}^n)$ with $\mu \ll \lambda$, $\Sigma = \{f_\sharp\mu : f \in \mathcal{F}(P)\}$ is again the set of reference distributions such that the compatibility condition holds for constant shears such as $Ax + b$.*

A corollary of the theorems above is when the transformations used are constant shears.

Corollary 2.11. *Consider an affine transformation $S(x) = Ax + b$, where A is symmetric positive definite with orthonormal basis given by an orthogonal matrix P. For a template distribution*

$\mu \in \mathcal{P}_2(\mathbb{R}^n)$ with $\mu \ll \lambda$, $\Sigma = \{f_\sharp\mu : f \in \mathcal{F}(P)\}$ is the set of reference distributions such that the compatibility condition holds.

Example 2.12 (Gaussians with fixed shear)**.** *To illustrate Theorem 2.8, we provide a simple example again with Gaussians. Let $\mu = \mathcal{N}(m_1, I_n)$. Consider a symmetric positive definite matrix A with spectral decomposition $A = P^\top \Lambda P$ and a corresponding fixed shear $S(x) = Ax + b$ for some $b \in \mathbb{R}^n$, which yields the pushforward $S_\sharp\mu = \mathcal{N}(Am_1 + b, AA^\top)$. For simplicity, we will check that the subset of compatible affine transformations*

$$\begin{aligned}\mathcal{F}_{affine}(P) &= \{f(x) = Cx + d : f \in \mathcal{F}(P)\}\\ &= \{P^\top DPx + d : D_{ij} = 0\; \forall\, i \neq j, D_{ii} > 0, d \in \mathbb{R}^n\}\end{aligned}$$

yields reference distributions $\sigma \in \{f_\sharp\mu : f \in \mathcal{F}_{affine}(P)\}$ so that the compatibility condition hold. In particular note that for $f(x) = Cx + d = P^\top DPx + d$, our reference distributions have the form

$$\sigma = \mathcal{N}(Cm_1 + d, CC^\top) = \mathcal{N}(Cm_1 + d, P^\top D^2 P).$$

Since the optimal transport map between two general Gaussians $\mathcal{N}(\tilde{m}_1, \Sigma_1) \to \mathcal{N}(\tilde{m}_2, \Sigma_2)$ is given by

$$\tilde{m}_2 + \Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}\Big(x - \tilde{m}_1\Big),$$

see [87], we know that

$$\begin{aligned}T_\sigma^\mu &= m_1 + \underbrace{(CC^\top)^{-\frac{1}{2}}(CC^\top)^{\frac{1}{2}}(CC^\top)^{-\frac{1}{2}}}_{(CC^\top)^{-1/2} = (C^2)^{-1/2} = C^{-1}}\Big(x - Cm_1 - d\Big)\\ &= m_1 + C^{-1}(x - Cm_1 - d) = C^{-1}(x - d).\end{aligned}$$

So we have that

$$S \circ T_\sigma^\mu(x) = AC^{-1}(x-d)+b.$$

On the other hand because $C = P^\top DP = C^\top$ and $A = P^\top \Lambda P = A^\top$ (so that $AC = CA$), we have that

$$\begin{aligned}
T_\sigma^{S_\sharp\mu} &= Am_1 + b + M(x-(Cm_1+d))\\
M &= (CC^\top)^{-1/2}((CC^\top)^{1/2}AA^\top(CC^\top)^{1/2})^{1/2}(CC^\top)^{-1/2}\\
&= C^{-1}(CA^2C)^{1/2}C^{-1} = C^{-1}(C^2A^2)^{1/2}C^{-1} = AC^{-1}\\
\implies T_\sigma^{S_\sharp\mu}(x) &= \underbrace{Am_1 - Am_1}_{0} + b + AC^{-1}(x-d) = S\circ T_\sigma^\mu(x).
\end{aligned}$$

So we actually get compatibility here and in section 2.12 we present a numerical validation of this fact.

Shears are Not Compatible in General: Another consequence of Theorem 2.3 is that non-trivial orthogonal transformations cannot be transformations that satisfy the compatibility condition.

Theorem 2.13. *Let $\sigma \ll \lambda, \mu \in \mathcal{P}_2(\mathbb{R}^n)$, and let $S(x) = Ax+b$ be a compatible transformation (i.e $S \circ T_\sigma^\mu = T_\sigma^{S_\sharp\mu}$) such that $b \in \mathbb{R}^n$ is a shift and $A \in \mathbb{R}^{n\times n}$ is an orthogonal matrix. Then A must be the identity.*

Proof. The proof can be found in Section 2.7. □

2.4 Binary classification with pre-specified separation

The main application of LOT isometries is to embed a subset of $\mathcal{P}_2(\mathbb{R}^n)$ into a linear space where binary classification is easily accomplished via linear separability. We show that data generated from a suitably bounded set of transformations still allows one to have LOT linear separability in a suitable supervised learning paradigm. We focus on classifying two classes. For the multi-class classification problem, one can use these results to build an ensemble of one-v-one classifiers.

Consider the following data-generating process:

Definition 2.14 (Elementary Transformation Generated Process)**.** *Consider a class of functions $\mathcal{H} \subseteq \{h : \mathbb{R}^n \to \mathbb{R}^n\}$. Let μ_1 or μ_2 be two probability measures. Then we call $\mathcal{H} \star \mu_1 = \{h_\sharp \mu_1 : h \in \mathcal{H}\}$ and $\mathcal{H} \star \mu_2 = \{h_\sharp \mu_2 : h \in \mathcal{H}\}$ the measures generated from elementary transformation $\mathcal{H}$ and μ_1 and $\mathcal{H}$ and μ_2, respectively. Moreover, assume that $\mathcal{H} \star \mu_1$ have label $y = 1$ and $\mathcal{H} \star \mu_2$ have label $y = -1$.*

Given a reference σ and a set of measures $\mathcal{Q}$, let $F_\sigma(\mathcal{Q})$ be the embedding of $\mathcal{Q}$ into the LOT space $L^2(\mathbb{R}^n, \sigma)$. Given the data generating process above, our goal is to show that the linear separability of $F_\sigma(\mathcal{H} \star \mu_1)$ and $F_\sigma(\mathcal{H} \star \mu_2)$ is well characterizable with respect to $\mathcal{H}$ and the distance between μ_1 and μ_2. We summarize the main result in the theorem below with proof given in Section 2.8:

Theorem 2.15. *Consider distributions $\mu_1, \mu_2, \sigma \in \mathcal{P}_2(\mathbb{R}^n)$, where μ_1 and μ_2 have bounded support, Wasserstein-2 distance $W_2(\mu_1, \mu_2) > 0$, and $\sigma \ll \lambda$. Pick a separation level δ such that $W_2(\mu_1, \mu_2) > \delta > 0$ and an error level $\varepsilon > 0$. Define $L \leq \frac{W_2(\mu_1, \mu_2) - \delta}{2} - \varepsilon$. Let*

$$\mathcal{H} \subseteq \{h : \mathbb{R}^n \to \mathbb{R}^n | h = \nabla \phi \textit{ for convex } \phi, ||h - I||_{\mu_i} \leq L, i \in \{1, 2\}\}$$

be some convex set of transformations such that $\mathcal{H}$ is compatible for σ and μ_1 as well as σ and μ_2. Furthermore, define the ε-tube of this set of transformations

$$\mathcal{H}_\varepsilon = \{\tilde{h} : \mathbb{R}^n \to \mathbb{R}^n | \; ||h - \tilde{h}||_{\mu_i} < \varepsilon; i \in \{1, 2\}, h \in \mathcal{H}\}.$$

Then, the sets $F_\sigma(\mathcal{H}_\varepsilon \star \mu_1)$ and $F_\sigma(\mathcal{H}_\varepsilon \star \mu_2)$ are linearly separable with separation at least δ.

Remark 2.16. *In Theorem 2.15, it should be emphasized that either σ needs to be chosen to be compatible with $\mathcal{H}$ and μ_1 and μ_2 or $\mathcal{H}$ needs to be chosen so that σ is compatible with μ_1 as well as μ_2 with respect to $\mathcal{H}$. This can occur, for example, if we choose $\sigma = N(0,I)$ to be an isotropic Gaussian and let $\mu_1 = N(0,\Sigma_1)$ and $\mu_2 = N(0,\Sigma_2)$ be such that $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$ (i.e. their covariances have the same orthonormal eigenbasis, say P), then $\mathcal{H} = \mathcal{F}(P)$ from Definition 2.5 works for compatibility according to Theorem 2.6. Another scenario would be to consider absolutely continuous target measures μ_0 and μ_1 with a constant speed geodesic $(\mu_t)_{t\in[0,1]}$, then Lemma 7.2.1 of [7] implies that*

$$T_{\mu_t}^{\mu_1} = T_{\mu_t}^{(T_{\mu_0}^{\mu_1})_\sharp \mu_0} = T_{\mu_0}^{\mu_1} \circ T_{\mu_t}^{\mu_0},$$

which is essentially the compatibility condition with $S = T_{\mu_0}^{\mu_1}$. In light of the proof for Theorem 2.30, we note that the Jacobian $J_{T_{\mu_t}^{\mu_0}}(x)$ must have the same eigenspaces as $J_{T_{\mu_0}^{\mu_1}}(T_{\mu_t}^{\mu_0}(x))$. Now, let $J_{T_{\mu_0}^{\mu_1}}(x) = P(x)^\top \operatorname{diag}(d(x))P(x)$ for some orthogonal matrix-valued and vector-valued function defined by $P(x)$ and $D(x)$, respectively, with $d(x) > 0$. If there exists a map S such that

$$J_S(x) = P(x)^\top \operatorname{diag}(\tilde{d}(x))P(x)$$

for some other vector-valued function $\tilde{d}(x) > 0$, then S should also be compatible. To see this, notice that $(S \circ T_{\mu_t}^{\mu_0})_\sharp \mu_t = S_\sharp \mu_0 = (T_{\mu_t}^{S_\sharp \mu_0})_\sharp \mu_t$, and since $J_{S\circ T_{\mu_t}^{\mu_0}}$ is symmetric positive-definite by construction, we know that $S \circ T_{\mu_t}^{\mu_0} = T_{\mu_t}^{S_\sharp \mu_0}$. In this case,

$$\{S : \mathbb{R}^n \to \mathbb{R}^n | J_S(x) = P(x)^\top \operatorname{diag}(\tilde{d}(x))P(x), \tilde{d}(x) > 0\}$$

is a candidate for $\mathcal{H}$.

Remark 2.17. *The bounds on the function class $\mathcal{H}$ ensure that $\mathcal{H} \star \mu_1$ and $\mathcal{H} \star \mu_2$ are disjoint.*

However, note that there can still exist function classes $\mathcal{H}$ without a bound on it, where $\mathcal{H} \star \mu_1$ and $\mathcal{H} \star \mu_2$ are still disjoint. For example, one can consider the case when $\mathcal{H}$ is the set of all shifts and when μ_1 and μ_2 are a uniform distribution on the unit square and an isotropic Gaussian. In this case, the sets $\mathcal{H} \star \mu_1$ and $\mathcal{H} \star \mu_2$ are disjoint.

Remark 2.18. *Notice that the functions $\mathcal{F}(P)$ from Definition 2.5 satisfy the conditions of $\mathcal{H}$ being the gradient of a convex function in Theorem 2.15 above. In particular, every $S = \tilde{P}^\top f(Px) \in \mathcal{F}(P)$ can be written as $S = \nabla\phi$ for some convex ϕ. To see this, let p_{ij} denote the (i, j)th entry of $\tilde{P}^\top$, then we have that*

$$\begin{aligned}\phi(x) &= \int_{\mathbb{R}} \Big((S(x))_j\Big) dx_j = \int_{\mathbb{R}} \sum_{k=1}^{n} p_{jk} f_k\Big(\sum_{i=1}^{n} p_{ki} x_i\Big) dx_j \\ &= \sum_{k=1}^{n} p_{jk} \int_{\mathbb{R}} f_k\left(\sum_{i=1}^{n} p_{ki} x_i\right) dx_j.\end{aligned}$$

Note that $J_{S(x)}(s) = P^\top J_{f(Px)}(x)$ and

$$(J_{f(Px)}(x))_{ij} = f_i'\left(\sum_{\ell=1}^{n} P_{i\ell} x_\ell\right) p_{ij} \implies J_{f(Px)}(x) = \operatorname{diag}(f'(Px))P$$

using the chain rule, where $(f'(Px))_j = f_j'((Px)_j)$. This tells us that $J_{S(x)}(x) = P^\top \operatorname{diag}(f'(Px))P$ so J_S is symmetric. Since the f_j's are increasing and differentiable, it is immediate that J_S is positive definite. This implies that ϕ is convex.

When we assume that $\mathcal{H}$ is compatible with respect to μ_1 and μ_2 and use either of these templates as the reference distribution, we actually gain better results than the general separation theorem above. The proof for the theorem below is in Section 2.8:

Theorem 2.19. *Fix $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^n)$ with finite support and $\mu_1, \mu_2 \ll \lambda$. Let $\mathcal{H}$ be a convex set of transformations that are compatible with μ_1 and μ_2 (this includes shifts and scalings). Let $\mathcal{H}_\varepsilon = \{h_\varepsilon : \|h - h_\varepsilon\|_{\mu_j} < \varepsilon, j = 1, 2\}$.*

1. *(Linear separability) If $\mathcal{H} \star \mu_1$ and $\mathcal{H} \star \mu_2$ are disjoint, then $F_{\mu_1}(\mathcal{H} \star \mu_1)$ and $F_{\mu_1}(\mathcal{H} \star \mu_2)$ are linearly separable.*

2. *(Linear separability of ε-tube functions) If the minimal separation between $\mathcal{H} \star \mu_1$ and $\mathcal{H} \star \mu_2$ is greater than 2ε, then $F_{\mu_1}(\mathcal{H}_\varepsilon \star \mu_1)$ and $F_{\mu_1}(\mathcal{H}_\varepsilon \star \mu_2)$ are linearly separable.*

3. *(Sufficient conditions for separation) If we assume:*

   (a) For every $h \in \mathcal{H}$ and every $x \in \mathbb{R}^n$ that $\|h(x)\|_2 \geq \sqrt{2}\|x - x_0\|_2$ where x_0 is the mean of the normalized measure $|\mu_1 - \mu_2|$

   $$x_0 = \frac{1}{|\mu_1 - \mu_2|(\mathbb{R}^n)} \int_{\mathbb{R}^n} z d|\mu_1 - \mu_2|(z),$$

   (b) $\sup_{h,\tilde{h} \in \mathcal{H}} \|h - \tilde{h}\|_{\mu_1} \leq W_2(\mu_1, \mu_2) - \delta - 2\varepsilon$ for $\delta > 0$,

   then $F_{\mu_1}(\mathcal{H}_\varepsilon \star \mu_1)$ and $F_{\mu_1}(\mathcal{H}_\varepsilon \star \mu_2)$ are separated by at least $\delta > 0$.

Remark 2.20. *Notice that if we choose $\mathcal{H}$ to be shifts and scalings, the first statement of Theorem 2.19 is the direct generalization of corollary 4.3 of [71] since shifts and scalings are compatible with every probability measure.*

Remark 2.21. *Notice that in Theorem 2.19, the condition $W_2(\mu_1, \mu_2) - \delta \geq \sup_{h,\tilde{h} \in \mathcal{H}} \|\tilde{h} - h\|_{\mu_1}$ in the third statement is essentially the same condition the one in Theorem 2.15 because by rewriting the condition in Theorem 2.15, we get $\sup_{h \in \mathcal{H}} \|h - I\|_{\mu_1} \leq \frac{W_2(\mu_1, \mu_2) - \delta}{2}$. This comes from the fact that*

$$2 \sup_{h \in \mathcal{H}} \|h - I\|_{\mu_1} \geq \sup_{h,\tilde{h} \in \mathcal{H}} \|\tilde{h} - h\|_{\mu_1} \geq \sup_{h \in \mathcal{H}} \|h - I\|_{\mu_1} - \inf_{\tilde{h} \in \mathcal{H}} \|\tilde{h} - I\|_{\mu_1}.$$

If the problem setting allows $I \in \mathcal{H}$, then the right hand side is just $\sup_{h \in \mathcal{H}} \|h - I\|_{\mu_1}$. Thus, in this case, Theorem 2.19 is stronger than Theorem 2.15 since our function class has the larger bound $\sup_{h \in \mathcal{H}} \|h - I\|_{\mu_1} \leq W_2(\mu_1, \mu_2) - \delta$.

Theorem 2.15 above acts as a blueprint for controlling the degree of separation in the LOT embedding via the bounds of the function class $\mathcal{H}$. For the specific setting of shears,

$$\mathcal{H}_{\gamma,M,M_b} = \left\{ Ax+b : \substack{A \text{ is symmetric positive definite with} \\ \lambda_{\min}(A) > \gamma \text{ and } \lambda_{\max}(A) < M, \text{ and } \|b\|_2 \leq M_b} \right\}, \tag{2.8}$$

we can choose γ, M, and M_b in a way that guarantees that $F_\sigma(\mathcal{H}_{\gamma,M,M_b} \star \mu_1)$ and $F_\sigma(\mathcal{H}_{\gamma,M,M_b} \star \mu_2)$ are δ-separated.

Corollary 2.22. *Consider two distributions μ_1 and μ_2 with Wasserstein-2 distance $W_2(\mu_1,\mu_2)$. Let us denote $R_1 = \max_{x \in supp(\mu_1)} \|x\|_2$ and $R_2 = \max_{x \in supp(\mu_2)} \|x\|_2$. For the function class of shears $\mathcal{H}_{\gamma,M,0}$ and $\sigma \ll \lambda$, we can ensure that $F_\sigma(\mathcal{H}_{\gamma,M,0} \star \mu_1)$ and $F_\sigma(\mathcal{H}_{\gamma,M,0} \star \mu_2)$ are δ-separated if*

Case 1: *assuming that $W_2(\mu_1,\mu_2) > (R_1+R_2)+\delta$, then M is chosen such that*

$$2 < M \leq \frac{W_2(\mu_1,\mu_2) - \delta + (R_1+R_2)}{R_1+R_2},$$

and

Case 2: *assuming that $\delta < W_2(\mu_1,\mu_2) < (R_1+R_2+\delta)$, then either M is chosen such that*

$$1 < M \leq \frac{W_2(\mu_1,\mu_2) - \delta + (R_1+R_2)}{R_1+R_2}$$

or γ is chosen such that

$$\gamma \geq \frac{\delta - W_2(\mu_1,\mu_2) + R_1 + R_2}{R_1+R_2}.$$

Proof. This comes straight from Corollary 2.38 provided that $M_b = 0$ and $\varepsilon = 0$. □

2.5 Binary Classification with Multiple References

It is possible to achieve better separation with a larger function class than the class of bounded shears described in Section 2.4. The cost of this better separation, however, is to use multiple LOT spaces. Note that once a set of two measures $\mathcal{H}\star\mu_1$ and $\mathcal{H}\star\mu_2$ are separable in LOT space with respect to one reference (from Theorem 2.15), then $\mathcal{H}\star\mu_1$ and $\mathcal{H}\star\mu_2$ must be separable in LOT space with respect to multiple references.

First we must provide a couple of definitions to extend our framework to multiple references.

Definition 2.23. *Given a family of functions $\mathcal{H}$ and a family of N reference measures $\sigma_1,\dots,\sigma_N$, the multiple reference LOT embedding of μ, denoted $F_N(\mu)$, is defined as*

$$F_N(\mu) = F_{\sigma_1}(\mathcal{H}\star\mu)\times\ldots\times F_{\sigma_N}(\mathcal{H}\star\mu).$$

Definition 2.24. *Given $\delta^* > 0$ and $(T_{1,\mu},\dots,T_{N,\mu}) \in F_N(\mu)$ and $(S_{1,\gamma},\dots,S_{N,\gamma}) \in F_N(\gamma)$, the families are called* δ^*-separable *if the product metric on $F_{\sigma_1}(\mathcal{P}_2)\times\ldots\times F_{\sigma_N}(\mathcal{P}_2)$ satisfies*

$$\left\| \left(d_{\sigma_1}(T_{1,\mu},T_{1,\gamma}),\dots,d_{\sigma_N}(T_{N,\mu},T_{N,\gamma}) \right) \right\|_2 > \delta^*,$$

where d_{σ_j} is the metric corresponding to $F_{\sigma_j}(\mathcal{P}_2)$ and $\|\cdot\|_2$ is the regular ℓ_2-norm that we are applying to the Euclidean point

$$\left(d_{\sigma_1}(T_{1,\mu},T_{1,\gamma}),\dots,d_{\sigma_N}(T_{N,\mu},T_{N,\gamma}) \right).$$

Lemma 2.25. *Let $\mu,\gamma \in \mathcal{P}_2$ with bounded support, $\varepsilon > 0$, and*

$$\mathcal{H} = \{h:\mathbb{R}^n\to\mathbb{R}^n | h = \nabla\phi \textit{ for convex } \phi, \|h-I\|_\mu \le L, \|h-I\|_\gamma \le L\},$$

where $2(L+\varepsilon) < W_2(\mu,\gamma)$. *Consider a desired separation level* δ^*. *If we have absolutely continuous (with respect to the Lebesgue measure) reference measures* $\sigma_1,\ldots,\sigma_N$ *such that* $\mathcal{H}$ *is compatible for* σ_j *and* μ *as well as* σ_j *and* γ *for K of the reference measures, where* $K \geq \left(\frac{\delta^\star}{W_2(\mu,\gamma)-2(L+\varepsilon)}\right)^2$*, then* $F_N(\mu)$ *and* $F_N(\gamma)$ *are* δ^**-separable with respect to* $\mathcal{H}_\varepsilon$ *and the given family of reference measures.*

Notice that the Lemma 2.25 allows one to pick a larger function class $\mathcal{H}$ and a small separation level δ^* than with just one reference measure; however, the number of LOT spaces that you must embed into is the cost of this better performance.

A basic (well-known) exercise in linear algebra shows that in any finite dimensional vector space V, for any $0 < r < p$, and for $x \in V$, we have

$$\|x\|_p \leq \|x\|_r \leq n^{\frac{1}{r}-\frac{1}{p}}\|x\|_p.$$

Even though $F_{\sigma_1}(\mathcal{P}_2) \times \ldots \times F_{\sigma_N}(\mathcal{P}_2)$ is an infinite-dimensional space, the product metric on this product space is actually acting on $\mathbb{R}_{>0} \times \ldots \times \mathbb{R}_{>0}$. This means that the ℓ_p and ℓ_r norm inequalities above hold for our product space when endowed with the product metric. This essentially signals "stronger" linear separability.

To see this with two spaces, assume that $F_{\sigma_1}(\mathcal{H}\star\mu)$ and $F_{\sigma_1}(\mathcal{H}\star\gamma)$ are δ_1-separated in $F_{\sigma_1}(\mathcal{P}_2)$ and that $F_{\sigma_2}(\mathcal{H}\star\mu)$ and $F_{\sigma_2}(\mathcal{H}\star\gamma)$ are δ_2-separated in $F_{\sigma_2}(\mathcal{P}_2)$, then in the product space, we have

$$\begin{aligned}
\max(\delta_1,\delta_2) &= \left\| \begin{pmatrix} d_{\sigma_1}(F_{\sigma_1}(\mathcal{H}\star\mu),F_{\sigma_1}(\mathcal{H}\star\gamma)) \\ d_{\sigma_2}(F_{\sigma_2}(\mathcal{H}\star\mu),F_{\sigma_2}(\mathcal{H}\star\gamma)) \end{pmatrix} \right\|_\infty \\
&\leq \left\| \begin{pmatrix} d_{\sigma_1}(F_{\sigma_1}(\mathcal{H}\star\mu),F_{\sigma_1}(\mathcal{H}\star\gamma)) \\ d_{\sigma_2}(F_{\sigma_2}(\mathcal{H}\star\mu),F_{\sigma_2}(\mathcal{H}\star\gamma)) \end{pmatrix} \right\|_2 \\
&\leq \sqrt{2}\left\| \begin{pmatrix} d_{\sigma_1}(F_{\sigma_1}(\mathcal{H}\star\mu),F_{\sigma_1}(\mathcal{H}\star\gamma)) \\ d_{\sigma_2}(F_{\sigma_2}(\mathcal{H}\star\mu),F_{\sigma_2}(\mathcal{H}\star\gamma)) \end{pmatrix} \right\|_\infty \\
&= \sqrt{2}\max(\delta_1,\delta_2).
\end{aligned}$$

We are more interested, however, in providing lower bounds for the product ℓ_2-norm. To investigate this, let's assume that $\mathcal{H}$ is fixed and that we have N templates distributions $\sigma_1,\ldots,\sigma_N$. Now if μ is a generic distribution, let

$$\begin{aligned}
F_N(\mu) &= F_{\sigma_1}(\mathcal{H}\star\mu)\times F_{\sigma_2}(\mathcal{H}\star\mu)\times\ldots\times F_{\sigma_N}(\mathcal{H}\star\mu) \\
&\subseteq F_{\sigma_1}(\mathcal{P}_2)\times\ldots\times F_{\sigma_N}(\mathcal{P}_2)
\end{aligned}$$

denote the embedding of $\mathcal{H}\star\mu$ into the product LOT space defined by $\sigma_1,\ldots,\sigma_N$. We will now prove the result.

Proof of Lemma 2.25. From Theorem 2.15, we know that for every j, $F_{\sigma_j}(\mathcal{H}\star\mu)$ and $F_{\sigma_j}(\mathcal{H}\star\mu)$ can be δ_j-separated for some $\delta_j < W_2(\mu,\gamma)-2(L+\varepsilon)$. Now notice that the degree of separation in the product space is

$$\left\| \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_N \end{pmatrix} \right\|_2 = \sqrt{\sum_{j=1}^N \delta_j^2} < \sqrt{\sum_{j=1}^N (W_2(\mu,\gamma)-2(L+\varepsilon))^2} = \sqrt{N}(W_2(\mu,\gamma)-2(L+\varepsilon)).$$

Thus, if we want to be at least δ^*-separated in the product space, then we must have

$$\delta^* \leq \left\| \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_N \end{pmatrix} \right\|_2 < \sqrt{N}(W_2(\mu,\gamma) - 2(L+\varepsilon))$$

$$\implies N > \left(\frac{\delta^*}{W_2(\mu,\gamma) - 2(L+\varepsilon)} \right)^2$$

So we're done. □

Example 2.26. *To show the tradeoff of Lemma 2.25, let's try a multiple LOT embedding example with Gaussians. Using the previous examples, assume that we have two template distributions* $\mu_1 = \mathcal{N}(0,\Sigma_1)$ *and* $\mu_2 = \mathcal{N}(0,\Sigma_2)$. *We know that* $W_2(\mu_1,\mu_2)^2 = \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{1/2})$. *Recalling that the optimal transport map from* μ_1 *to* μ_2 *is given by*

$$\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})\Sigma_1^{-1/2}x =: A_{\mu_1\to\mu_2}x,$$

we consider the set of shears

$$\mathcal{H} = \{Ax : A = A^\top \in \mathbb{R}^{n\times n}, AA_{\mu_1\to\mu_2} = A_{\mu_1\to\mu_2}A, MI_n \succeq A \succeq mI_n \succ 0\}$$

as our set of transformations, where the commuting property of A with $A_{\mu_1\to\mu_2}$ *ensures that* $\mathcal{H}$ *compatible with* μ_1 *and* μ_2. *To ensure separation, we use* $L \leq \frac{W_2(\mu_1,\mu_2)-\delta}{2}$, *which is shown in Section 2.9 to imply that*

$$\max\left(|M-1|, |1-m|\right) \leq \frac{W_2(\mu_1,\mu_2) - \delta}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}.$$

Now let us define our reference distributions to be of the form $\sigma_1 = (h_1)_\sharp\mu_1$ *and* $\sigma_2 = (h_2)_\sharp\mu_2$ *for*

$h_1(x) = A_1x$ and $h_2(x) = A_2x$, where A_1, A_2 are chosen so that $h_1, h_2 \in \mathcal{H}$, and so

$$\sigma_1 = (h_1)_\sharp\mu_1 = \mathcal{N}(0, A_1\Sigma_1A_1^\top), \quad \sigma_2 = (h_2)_\sharp\mu_2 = \mathcal{N}(0, A_2\Sigma_2A_2^\top).$$

Notice that the bounds on M and m imply that there are infinite choices of reference distributions to choose from. Moreover, we show in Section 2.9 that

$$\frac{M^2}{m}W_2(\mu_1,\mu_2) \geq \|T_{\sigma_j}^{h_\sharp\mu_1} - T_{\sigma_j}^{\tilde{h}_\sharp\mu_2}\|_{\sigma_j} \geq \frac{m^2}{M}W_2(\mu_1,\mu_2)$$

for our choices of reference distributions and any $h, \tilde{h} \in \mathcal{H}$. Now choosing N reference distributions, our multiple LOT embedding has minimal separation bounded below by

$$\sqrt{\sum_{j=1}^{N} \|T_{\sigma_j}^{(h_1)_\sharp\mu_1} - T_{\sigma_j}^{(h_2)_\sharp\mu_2}\|_{\sigma_j}^2} \geq \sqrt{\sum_{j=1}^{N} \frac{m^4}{M^2}W_2(\mu_1,\mu_2)^2} = \sqrt{N}\frac{m^2}{M}W_2(\mu_1,\mu_2).$$

These choices of σ_1 and σ_2 ensure that each reference is compatible with μ_1 and $\mathcal{H}$ as well as μ_2 and $\mathcal{H}$. Notice that as δ becomes closer to $W_2(\mu_1,\mu_2)$, we find that both m and M become closer to 1, which means that our set of shears become closer to the identity. Using multiple LOT embeddings; however, we can actually use the maximal function class of shears $\mathcal{H}$ when $M = 1 + \frac{W_2(\mu_1,\mu_2)}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}$ and $m = 1 - \frac{W_2(\mu_1,\mu_2)}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}$. To get the same separation with the largest possible function class as when we have $\delta > 0$, we need

$$\sqrt{N}\left(\frac{\left(1 - \frac{W_2(\mu_1,\mu_2)}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}\right)^2}{1 + \frac{W_2(\mu_1,\mu_2)}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}}\right) \geq \left(\frac{\left(1 - \frac{W_2(\mu_1,\mu_2)-\delta}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}\right)^2}{1 + \frac{W_2(\mu_1,\mu_2)-\delta}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}}\right).$$

Rearranging the inequality and squaring both sides, we get the following bound for N

$$N \geq \left(\frac{2 \max\limits_{j=1,2} \|\Sigma_j^{1/2}\|_F + W_2(\mu_1,\mu_2)}{2 \max\limits_{j=1,2} \|\Sigma_j^{1/2}\|_F + W_2(\mu_1,\mu_2) - \delta} \right)^2 \left(\frac{2 \max\limits_{j=1,2} \|\Sigma_j^{1/2}\|_F - W_2(\mu_1,\mu_2) + \delta}{2 \max\limits_{j=1,2} \|\Sigma_j^{1/2}\|_F - W_2(\mu_1,\mu_2)} \right)^4 .$$

Thus, if needed, we can allow δ to stay small (or even become zero), which would allow us to use the maximal function class of shears $\mathcal{H}$; however, the cost of this larger function class and separation level is increasing the number of reference distributions.

2.6 Numerical experiments

2.6.1 Binary classication of MNIST Images

In this section we present pairwise binary classification results on sheared MNIST images which are motivated by the linear separability result presented in Corollary 2.22 and also illustrate the benefit of using multiple references as indicated by lemma 2.25.

The LOT embedding pipeline for an image [1]

1. Obtain the image represented as a $n \times n$ matrix of pixel values.
2. Assuming that the image is supported on a $n \times n$ grid on the unit square, obtain the point cloud which forms the support of the pixel values corresponding to the image.
3. Obtain the discrete measure μ induced by the image on the unit square. Each point in the support of the image has a pixel value which (after normalization) will be the mass associated with μ .

[1] Code for our LOT classification experiments on MNIST images can be found at https://github.com/srjrhkannan/LOTpython

4. Let σ denote a discrete reference measure [2]. Compute the discrete transport coupling matrix P_{σ}^{μ} [3]. For each point x in the support of the reference σ, choose $T_{\sigma}^{\mu}(x)$ as the point in the support of μ such that $T_{\sigma}^{\mu}(x) = \text{argmax}_{y \in supp(\mu)} P_{\sigma}^{\mu}(x,y)$. Here $P(x,y)$ denotes the amount of mass transported from $x \in supp(\sigma)$ to $y \in supp(\mu)$. This is done to extract an approximate Monge map from the coupling matrix [71].

5. The LOT embedding of the image corresponding to the reference σ is chosen to be T_{μ}^{σ}. Note that $T_{\mu}^{\sigma} \in \mathbb{R}^{2m}$, where m denotes the size of the size of the support σ, i.e. $m := |supp(\sigma)|$. Henceforth this $\mathbb{R}^{2m}$ vector will be referred to as the *LOT feature* corresponding to the particular image that is being embedded.

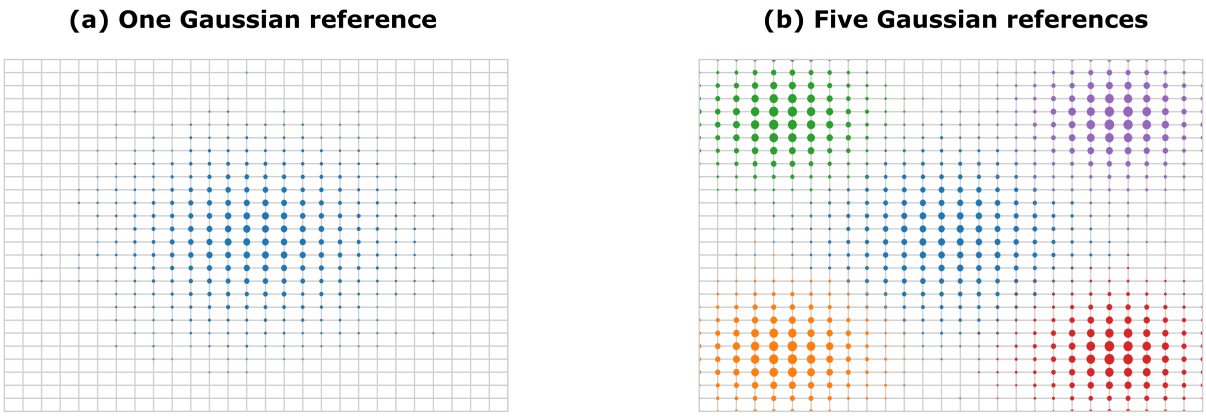


Figure 2.1: a) A Gaussian reference distribution approximated on a 28×28 grid. b) Five different Gaussian distributions approximated on a 28×28 grid to be used as multiple reference for LOT embedding.

2.6.2 Experimental settings

The MNIST images are sheared using the transformation described in Section 2.10 and the values for each of the parameters $\lambda_1, \lambda_2, \theta, b$ are drawn randomly from a pre-fixed range for

[2] In case the desired reference is an absolutely continuous measure on the unit square, then we work with the discrete measure it induces on the $n \times n$ grid on the unit square (See Figure 2.1).

[3] https://pythonot.github.io/ [43]

each image. We perform classification experiments for the MNIST images under two different shearing conditions (See Figure 2.2). For one set of shearing conditions, termed as *mild shearing* , the parameters of shearing for each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 1.5]$, θ is randomly chosen in the interval $[0, 360]$ degrees and the shifts b are randomly chosen in the interval $[-5, 5]$. For the other set of shearing conditions termed as *severe shearing*, the parameters of shearing for each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 2.5]$, θ is randomly chosen in the interval $[0, 360]$ degrees and the shifts b are randomly chosen in the interval $[-5, 5]$. Then the *LOT feature* corresponding to each of the sheared images are computed using the embedding pipeline described in subsection 2.6.1 and then classification experiments are performed using Linear Discriminant Analysis (LDA) [50] [4].

To test the performance of LDA (Linear Discriminant Analysis) classification of two distinct classes of MNIST [40] digits using LOT features, we study the test error of the LDA classifier as a function of the number of training images chosen for each digit. For each fixed number, N_{train}, of training images, we train the LDA classifier using a randomly chosen set of N_{train} images from each digit class and test the classification results on a randomly chosen set of 1000 test images from each digit class. We then repeat this experiment for each fixed N_{train} using 20 different randomly chosen set of training images (N_{train} images from each digit class) and 1000 test images from each digit class.

2.6.3 Observations

In Figure 2.3 we report the mean test error for classification of MNIST ones and twos and in Figure 2.4 we report the mean test error for classification of MNIST sevens and nines for various choices of reference distributions and under different shearing conditions. Therein for comparison, we also report the results obtained using the semi-discrete optimal transport [68]

[4] https://scikit-learn.org/

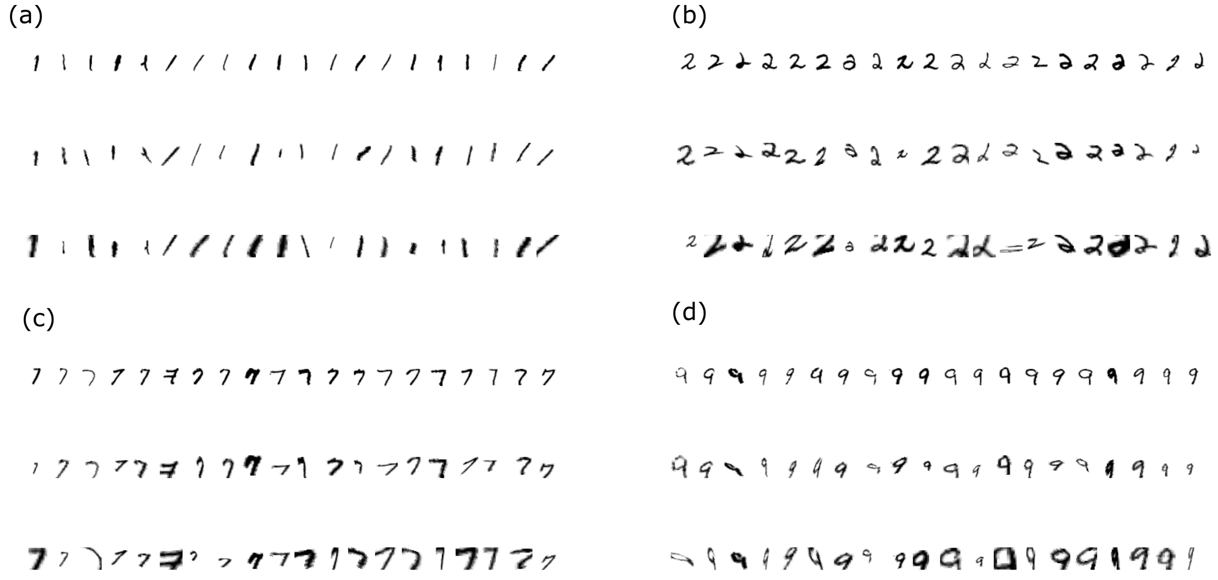


Figure 2.2: In each figure, the first row shows the true unsheared MNIST image. The second row shows the corresponding mildly sheared MNIST image. The parameters (Section 2.10) of shearing for each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 1.5]$, θ is randomly chosen in the interval $[0, 360]$ degrees and the shifts b are randomly chosen in the interval $[-5, 5]$. The third row shows the corresponding severely sheared MNIST image. The parameters (Section 2.10) of shearing for each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 2.5]$, θ is randomly chosen in the interval $[0, 360]$ degrees and the shifts b are randomly chosen in the interval $[-5, 5]$

framework which uses a uniform reference measure. The corresponding standard deviations are reported in Figures 2.8 and 2.9. We observe that the LOT framework is able to achieve low test errors with a relatively low number of training images. Moreover we see that using multiple references does indeed lead to a decrease in the classification error. Interestingly, we observe that using multiple references also helps reduce over-fitting (See Figure 2.5). The trade-off observed is that using multiple references increases the length of the feature vector while on the other hand it leads to a decrease in the test error.

In Figure 2.6 we illustrate as a heat-map, the mean test errors for binary classification of all pairs of MNIST digits using 50 training images per class and for different choices of references. Also, in Table 2.1 we report the range of test errors and standard deviations observed across all the classification experiments corresponding to Figure 2.6. Further in Figure 2.11, for comparison, we report the classification results for sheared MNIST 7s and 9s using convolutional neural networks with 1586 training parameters (labelled small CNN) and 3650 training parameters (labelled large CNN) under identical training and testing conditions as that of the discrete LOT classifier.

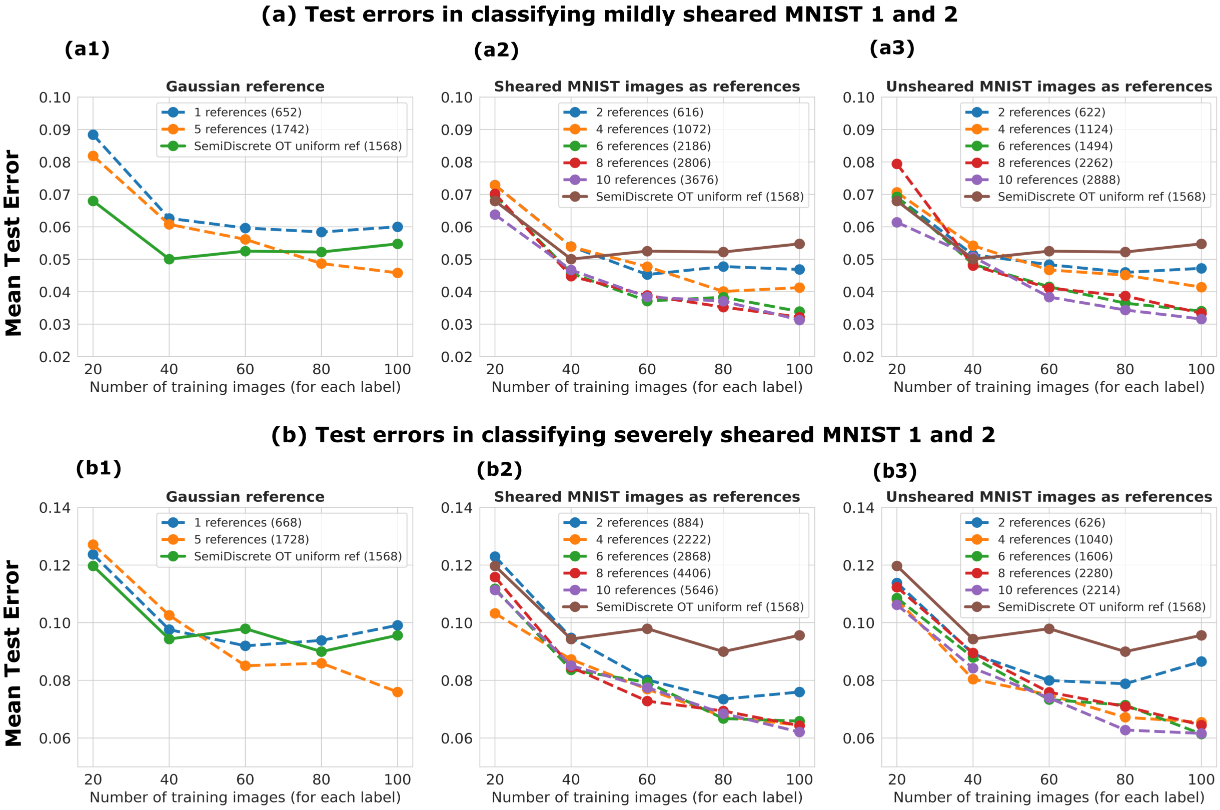


Figure 2.3: (a) Test errors for binary classification of mildly sheared MNIST 1s and 2s using (a1) Gaussian references (a2) sheared MNIST 1s and 2s as references (a3) unsheared MNIST 1s and 2s as references. (b) Test errors for binary classification of severely sheared MNIST 1s and 2s using (b1) Gaussian references (b2) sheared MNIST 1s and 2s as references (b3) unsheared MNIST 1s and 2s as references. In the cases where MNIST images are used as references, the results are reported for the cases where the number of references used is $2i$ for $i = 1, \cdots 5$ wherein i images from each class are randomly drawn to be used as references from a pool of images that do not correspond to any of the training and testing images. For each fixed number of training images per class, N_{train}, the mean test classification error averaged across 20 random choices of N_{train} training images (per class) and 1000 test images (per class) is reported. The number inside the parenthesis in the legends of the images denote the length of the LOT feature vector corresponding to the particular choice of references. In all figures, for comparison, the results for classification using the semi discrete linear optimal transport framework [68] which uses the uniform measure as the reference is also reported. Standard deviations for each of the corresponding classification tests are reported in the Figure 2.8.

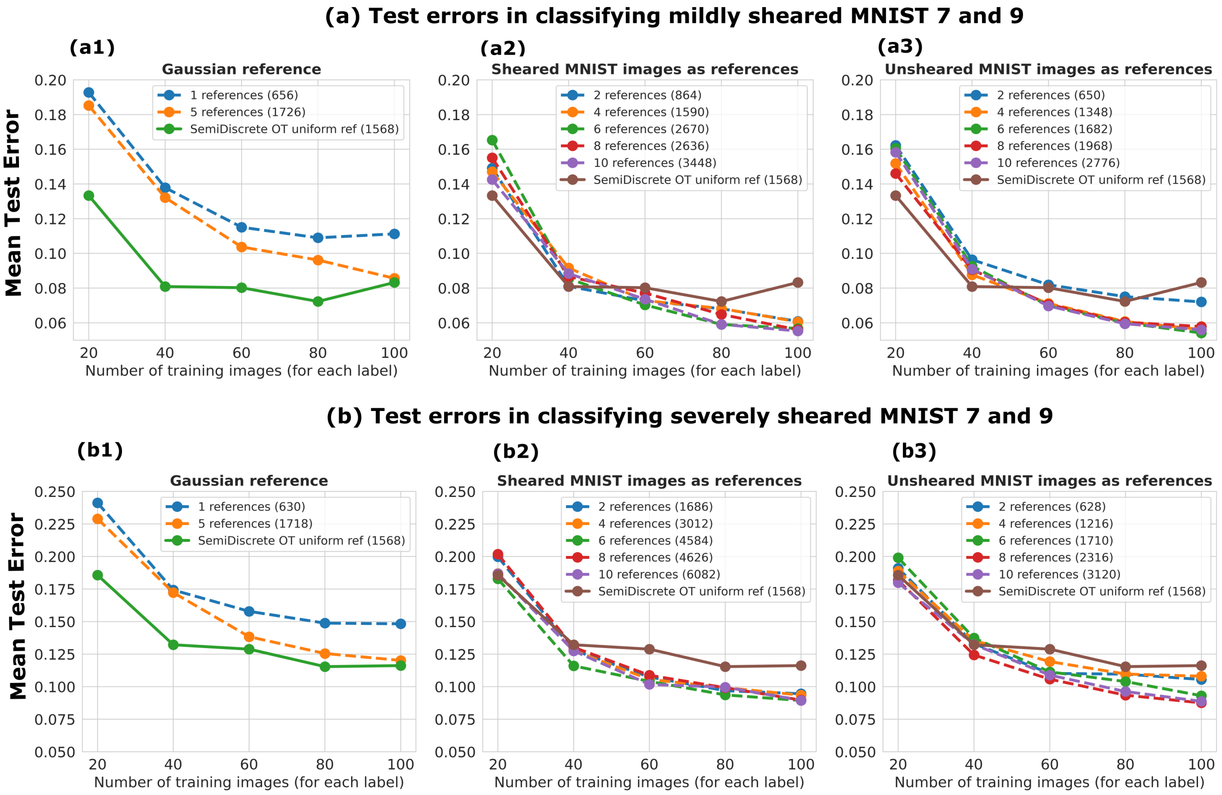


Figure 2.4: (a) Test errors for binary classification of mildly sheared MNIST 7s and 9s using (a1) Gaussian references (a2) sheared MNIST 7s and 9s as references (a3) unsheared MNIST 7s and 9s as references. (b) Test errors for binary classification of severely sheared MNIST 7s and 9s using (b1) Gaussian references (b2) sheared MNIST 7s and 9s as references (b3) unsheared MNIST 7s and 9s as references. In the cases where MNIST images are used as references, the results are reported for the cases where the number of references used is $2i$ for $i = 1, \cdots 5$ wherein i images from each class are randomly drawn to be used as references from a pool of images that do not correspond to any of the training and testing images. For each fixed number of training images per class, N_{train}, the mean test classification error averaged across 20 random choices of N_{train} training images (per class) and 1000 test images (per class) is reported. The number inside the parenthesis in the legends of the images denote the length of the LOT feature vector corresponding to the particular choice of references. In all figures, for comparison, the results for classification using the semi discrete linear optimal transport framework [68] which uses the uniform measure as the reference is also reported. Standard deviations for each of the corresponding classification tests are reported in the Figure 2.9.

2.7 Compatibility Condition Proofs

Lemma 2.27. *Suppose V is a finite-dimensional vector space, $\phi : V \to V$ is a diagonalizable linear map, and $U \subseteq V$ is a ϕ-invariant subspace. Then the restriction $\phi|_U : U \to U$ is diagonalizable.*

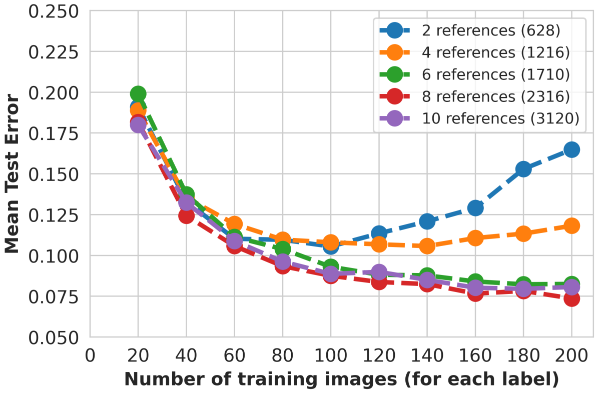


Figure 2.5: Illustration of the benefit of using multiple references to reduce overfitting in the classification of severely sheared MNIST 7s and 9s using true MNIST images as references under the same training and testing conditions of Figure 2.4 (b3).

Table 2.1: Range of mean value and standard deviations of test errors for pairwise classification of sheared MNIST images across all pairs of digits for various reference choices. The reported values are across 20 experiments involving different choices of 50 randomly drawn training images per class and 500 randomly drawn test images per class for each experiment.

Reference choice	**Mean**		**STD deviation**	
	Mild	Severe	Mild	Severe
1 Gaussian	$[0.0083, 0.1298]$	$[0.0198, 0.2132]$	$[0.0064, 0.0291]$	$[0.0108, 0.0382]$
2 MNIST	$[0.0078, 0.0880]$	$[0.0220, 0.1585]$	$[0.0056, 0.0244]$	$[0.0111, 0.0328]$

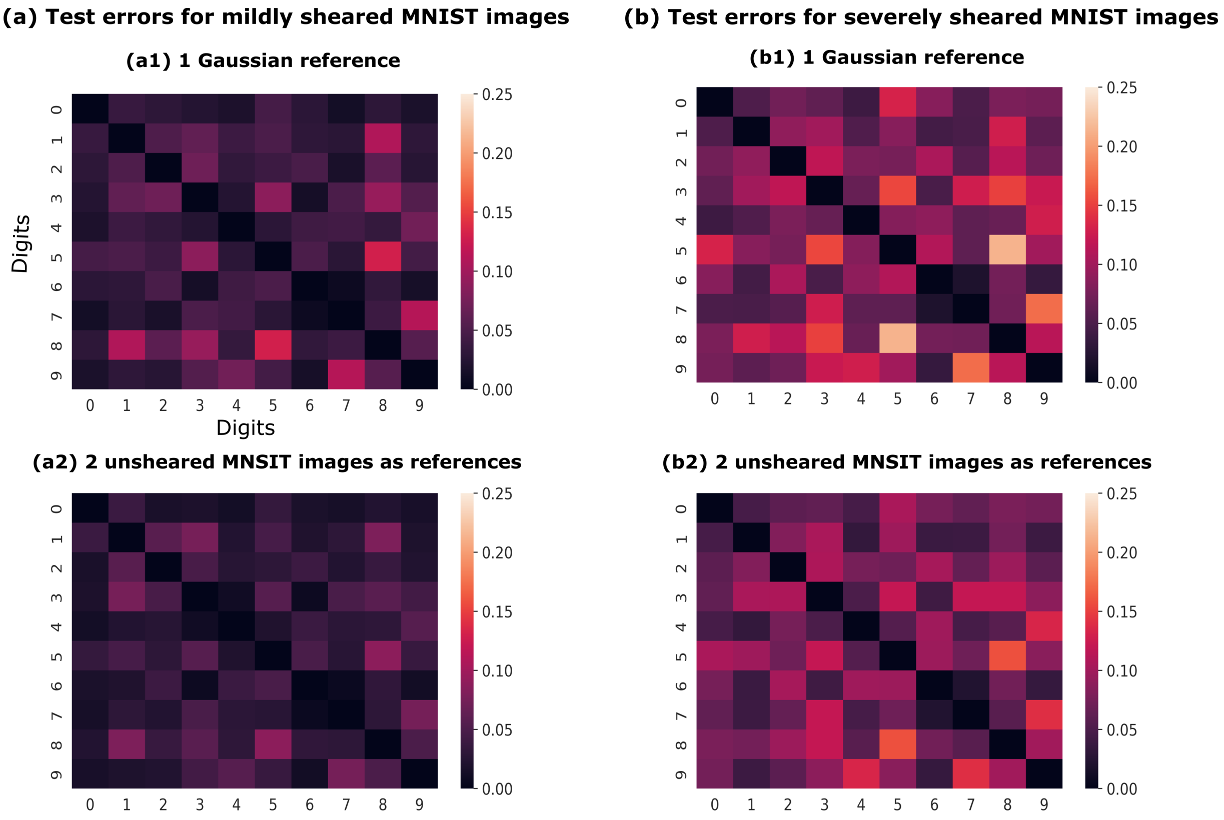
Proof. Let $\lambda_1, \dots, \lambda_k$ be distinct eigenvalues of ϕ. We will denote by $E(\lambda_k, \phi)$ the eigenspace of ϕ corresponding to eigenvalue λ_k. Since ϕ is diagonalizable over V, we can represent V as a direct sum

$$V = E(\lambda_1, \phi) \oplus \cdots \oplus E(\lambda_m, \phi).$$

This means exactly that any vector v is given by

$$v = w_1 + \cdots + w_m$$

where $w_i \in E(\lambda_i, \phi)$. As U is a finite dimensional vector space, we know that there exists a basis



(a) Test errors for mildly sheared MNIST images

(b) Test errors for severely sheared MNIST images

(a1) 1 Gaussian reference

(b1) 1 Gaussian reference

(a2) 2 unsheared MNSIT images as references

(b2) 2 unsheared MNSIT images as references

Figure 2.6: (a) Test errors for binary classification of all pairs of mildly sheared MNIST images using (a1) one Gaussian reference (a2) two unsheared MNIST images as references. (b) Test errors for binary classification of all pairs of mildly sheared MNIST images using (b1) one Gaussian reference (b2) two unsheared MNIST images as references. For each given pair of digits, in the case of MNIST images as references (a2),(b2), one image corresponding to each class is randomly drawn to serve as the references. The reported error is a mean value 20 experiments involving different choices of 50 randomly drawn training images per class and 500 randomly drawn test images per class for each experiment. The range of standard deviations for the test errors for each case is reported in Table 2.1.

for U given by $\{u_1,\ldots,u_k\}$. Let us consider the linear map

$$\Phi_i(u) = \prod_{\substack{j=1,\ldots,m \\ j\neq i}} (\lambda_j I - \phi|_U)u.$$

Note that this linear map is commutative in its order of composition. We now will take every basis vector u_i and represent it in terms of eigenvector. Note that because u_i is a vector in V, we

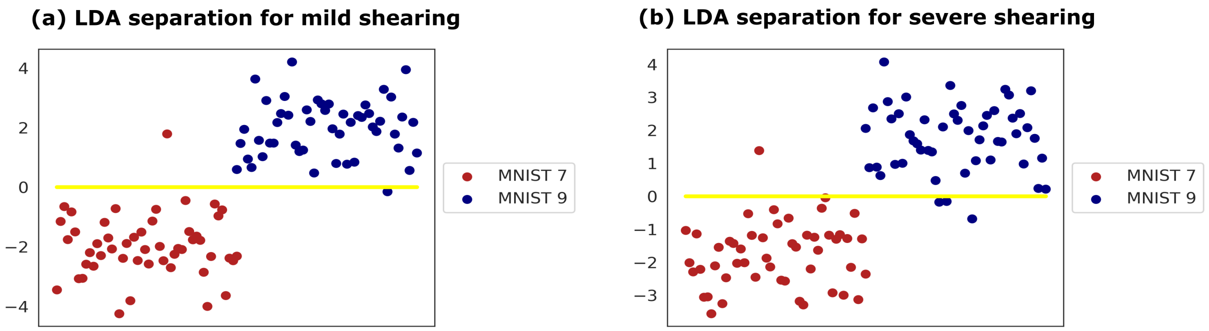


Figure 2.7: Visualization of separation in the LDA projection using LOT features with 2 unsheared MNIST images as references for 50 training images per class corresponding to (a) mildly sheared MNIST sevens and nines (b) severely sheared MNIST sevens and nines. The y-axis denotes the value of the projection onto the LDA separating line for the two classes, and the x-axis denotes an ordering of the MNIST images.

find that there exists eigenvectors $w_{1,1} \in E(\lambda_1,\phi),\ldots,w_{1,m} \in E(\lambda_m,\phi)$ such that

$$u_1 = w_{1,1} + w_{1,2} + \ldots, w_{1,m}.$$

Now let us create a set $\widehat{W}_1 = \{w_{1,1},\ldots,w_{1,2}\}$. Note that

$$\Phi_i(u_1) = \prod_{\substack{j=1,\ldots,m\\ j\neq i}} (\lambda_j - \lambda_i) w_{1,i} \implies w_{1,i} \in U,$$

since U is ϕ-invariant. Because this happens for arbitrary i, we know that $w_{1,i} \in U$ for all i. Note, that this set is linearly independent since each $w_{1,i}$ comes from a different eigenspace. We repeat this for u_j to obtain $\widehat{W}_j$, and note that $\widehat{W}_j \subseteq U$. Now, let us define $\bigcup\limits_{j=1}^{k} \widehat{W}_j = \widehat{W}$. Note that this is a spanning set of eigenvectors for U, and we can make this into a linearly independent set that still spans U by throwing away the linearly dependent vectors. Note that because of finite dimensionality, this process will stop, and will yield a linearly independent, spanning set of U, let's call it $\widehat{W}$, consisting of eigenvectors. So this means that $\phi|_U$ is diagonalizable since we found an eigenbasis for U. So we're done. □

The following theorem is a fundamental result from matrix analysis (see [53, Theorem

1.3.12]), but we provide a proof for convenience of the reader.

Theorem 2.28. *Let A and B be two $n \times n$ diagonalizable matrices that commute (i.e. $AB = BA$). Then there exists a basis of $\mathbb{R}^n$ consisting of simultaneous eigenvectors of A and B.*

Proof. We break this proof up into two parts. First we will show that given an eigenvector λ the eigenspace of A corresponding to λ (we denote this with $E(\lambda, A)$ is B-invariant. Consider $v \in E(\lambda, A)$, then notice that

$$ABv = BAv = B(\lambda v) = \lambda Bv.$$

This means that Bv is an eigenvector for A with eigenvalue λ, which means that $E(\lambda, A)$ is B-invariant since B maps elements of $E(\lambda, A)$ back into $E(\lambda, A)$. Now we show that there exists a basis for $\mathbb{R}^n$ consisting of simultaneous eigenvectors of A and B.

Note that because A is diagonalizable, we know that $\mathbb{R}^n$ can be represented as a direct sum given by

$$\mathbb{R}^n = \bigoplus_{i=1}^{k} E(\lambda_i, A),$$

where $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of A. Now to show that there exists a basis of $\mathbb{R}^n$ consisting of simultaneous eigenvectors of A and B, we only need to find a basis for each subspace $E(\lambda, A)$ because the concatenation of all these bases will yield a basis for $\mathbb{R}^n$. Now note that since $E(\lambda, A)$ is a B-invariant space by above and because B is diagonalizable, we know from Lemma 2.27 that the restriction of B to this eigenspace, $B|_{E(\lambda,\phi)}$, is diagonalizable, which means that there exists an eigenbasis of $E(\lambda, A)$ for the map B. Let us call this this eigenbasis $S_{\lambda,A} = \{w_1, \ldots, w_j\}$, where j is the dimension of $E(\lambda, A)$. Now, note that $S_{\lambda,A}$ consists of eigenvectors of both B and A. To see this, note that $S_{\lambda,A} \subseteq E(\lambda, A)$; thus, every w_i is an eigenvector of A. Moreover, $S_{\lambda,A}$ is an eigenbasis for $B|_{E(\lambda,A)}$ by construction (from Lemma 2.27).

This means that

$$S = \bigcup_{i=1}^{k} S_{\lambda_i,A}$$

forms a basis for $\mathbb{R}^n$ consisting of simultaneous eigenvectors of A and B. □

Lemma 2.29. *If two symmetric matrices A and B commute, then there exists spectral decompositions $A = Q^\top \Lambda Q$ and $B = P^\top DP$ such that the rows of Q are the same as the rows of P up to a permutation.*

Proof. We already know that if two diagonalizable matrices commute, then they share the same eigenvectors; thus, there exist an eigendecomposition for A and B with the same eigenvectors. By extension, this holds for symmetric matrices. If we assume that these eigendecompositions are given by $A = Q^\top \Lambda Q$ and $B = P^\top DP$, the eigenvectors of A are exactly the columns of $Q^\top$, and similarly, the eigenvectors of B are exactly the columns of $P^\top$. This implies that the columns of $Q^\top$ and $P^\top$ should be the same. The order of the columns can be permuted without loss of generality and still provide the same transformation A and B. Thus, we can assume that Q has the same rows as P. □

Theorem 2.30. *Let $S : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable map such that $S = \nabla\varphi$ for some φ. Let $\sigma, \mu \in \mathcal{P}_2(\mathbb{R}^n)$ with σ absolutely continuous with respect to the Lebesgue measure. Assume that the compatibility condition $S \circ T_\sigma^\mu = T_\sigma^{S_\sharp\mu}$ holds. Then $J_S(x)$ is a symmetric positive definite matrix for all x. Moreover, $J_S(T_\sigma^\mu(x))$, $J_{T_\sigma^\mu}(x)$, and $J_{T_\sigma^{S_\sharp\mu}}(x)$ share the same eigenspaces. Furthermore, the eigenvalues of $J_S(T_\sigma^\mu(x))$ are of the form $\frac{\lambda_{\sigma,\mu}}{\lambda_{\sigma,S_\sharp\mu}}$ where $\lambda_{\sigma,\mu}$ is an eigenvalue of $J_{T_\sigma^\mu}(x)$ and $\lambda_{\sigma,S_\sharp\mu}$ is an eigenvalue of $J_{T_\sigma^{S_\sharp\mu}}(x)$.*

Proof of Theorem 2.30. Recall that the main equation for us to study is

$$S \circ T_\sigma^\mu = T_\sigma^{S_\sharp\mu}.$$

By Theorem 2.1, there exist convex functions γ and ϕ such that $T_\sigma^\mu = \nabla\phi$ and $T_\sigma^{S_\sharp\mu} = \nabla\gamma$. By Clairaut's theorem (or the Schwarz theorem), $\nabla^2\gamma(x)$ and $\nabla^2\phi(x)$ are symmetric. Using the multivariate chain rule and the symmetry of $\nabla^2\gamma(x)$, we get that

$$\begin{aligned}
\nabla^2\gamma(x) &= J_S(\nabla\phi(x))\nabla^2\phi(x) \\
\nabla^2\gamma(x)^\top &= (\nabla^2\phi(x))^\top J_S(\nabla\phi(x))^\top \\
&= \nabla^2\phi(x)J_S(\nabla\phi(x))^\top.
\end{aligned}$$

Since $J_S = \nabla^2\varphi$ for some φ, then $J_S^\top(x) = J_S(x)$ for all $x \in \mathbb{R}^d$. Since $J_S(\nabla\phi(x))$ and $\nabla^2\phi(x)$ are symmetric matrices that commute, according to Lemma 2.29, there exists some orthogonal matrix P such that we can write the eigendecompositions of $\nabla^2\phi(x)$ and $J_S(\nabla\phi(x))$ as $\nabla^2\phi(x) = P^\top\Lambda_\phi(x)P$ and $J_S(\nabla\phi(x)) = P^\top\Lambda_S(\nabla\phi(x))P$ where the matrices Λ_ϕ and Λ_S are diagonal matrices with the eigenvalues of $\nabla^2\phi(x)$ and $J_S(\nabla\phi(x))$, respectively. Moreover, if Λ_γ denotes the diagonal matrix in the eigendecomposition for γ, then our matrix equations above can be written as

$$\begin{aligned}
\nabla^2\gamma(x) &= J_S(\nabla\phi(x))\nabla^2\phi(x) \\
P^\top\Lambda_\gamma(x)P &= P^\top\Lambda_S(\nabla\phi(x))PP^\top\Lambda_\phi(x)P \\
\Lambda_\gamma(x) &= \Lambda_S(\nabla\phi(x))\Lambda_\phi(x).
\end{aligned}$$

This immediately shows that every eigenvalue λ_S of $J_S(\nabla\phi(x))$ can be written as $\frac{\lambda_\gamma}{\lambda_\phi}$, where λ_γ is an eigenvalue of $\nabla^2\gamma(x)$ and λ_ϕ is an eigenvalue of $\nabla^2\phi(x)$. Since $\nabla^2\phi(x)$ and $\nabla^2\gamma(x)$ are Hessians of a convex function, they must be positive definite. This implies that all the eigenvalues of $J_S(\nabla\phi(x))$ are positive. Since $J_S(\nabla\phi(x))$ is symmetric, we immediately get that $J_S(\nabla\phi(x)) = \nabla^2\varphi(\nabla\phi(x))$ is a symmetric positive definite matrix, which means that φ must have been convex. This implies that $S = \nabla\varphi$ is a transport map. □

Lemma 2.31. *Let an optimal transport map be given by* $\nabla\phi(x)$ *for some convex function* ϕ*. If* *the Hessian $\nabla^2\phi(x)$ has a spectral decomposition that does not depend on x (i.e. $P^\top D(x)P$ for a positive diagonal matrix $D(x)$), then the map $P\nabla\phi(P^\top x)$ has a diagonal Jacobian and each component of $P\nabla\phi(P^\top x)$ is a function of only a single variable.*

Proof of Lemma 2.31. If we compute the Jacobian of $P\nabla\phi(P^\top x)$ by using the chain rule twice, we get that the Jacobian of $P\nabla\phi(P^\top x)$ is given by

$$\begin{aligned} J_{P\nabla\phi(P^\top x)}(x) &= PJ_{\nabla\phi(P^\top x)}(x) = P\nabla^2\phi(P^\top x)P^\top \\ &= PP^\top D(P^\top x)PP^\top = D(P^\top x). \end{aligned}$$

This means that if we write the transport map $\nabla\phi$ in the basis given by the columns of $P^\top$ and the output is written in terms of the basis given by the columns of P, our transport map $\nabla\phi$ can be written as n single variable functions. To see this, notice that we can write the jth coordinate output of $P\nabla\phi(P^\top x)$ as some function f_j to give us

$$P\nabla\phi(P^\top x) = \begin{bmatrix} f_1(x_1,\dots,x_n) \\ f_2(x_1,\dots,x_n) \\ \vdots \\ f_n(x_1,\dots,x_n) \end{bmatrix}.$$

Recall that the (j,k)th entry of the Jacobian $J_{P\nabla\phi(P^\top x)}(x)$ is $\frac{\partial f_j}{\partial x_k}$. Because the Jacobian is diagonal, we see that $\frac{\partial f_j}{\partial x_k} = 0$ for $j \neq k$. This implies that we can actually write

$$P\nabla\phi(P^\top x) = \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_n(x_n) \end{bmatrix}.$$

So we're done. □

Now we can prove the main LOT isometry theorems for shears.

Proof of Theorem 2.6. Assume that the Jacobian of T_σ^μ has constant orthonormal basis given by an orthogonal matrix P, then Theorem 2.30 tells us that a compatible transformation S must have positive symmetric definite Jacobian J_S and has the same eigenspaces as $J_{T_\sigma^\mu}$. First, note that the corollaries of Theorem 2.30 implies that S is an optimal transport map. Second, note that since J_S commutes with $J_{T_\sigma^\mu}$, we know that $J_S = \tilde{P}^\top D(x)\tilde{P}$, where $\tilde{P}$ is a row-permutation of P from Lemma 2.29. Because S satisfies the assumptions of Lemma 2.31, we get that

$$\begin{aligned}\tilde{P}S(\tilde{P}^\top x) &= \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_n(x_n) \end{bmatrix} \\ \implies S(x) &= \tilde{P}^\top \begin{bmatrix} f_1((\tilde{P}x)_1) \\ f_2((\tilde{P}x)_2) \\ \vdots \\ f_n((\tilde{P}x)_n) \end{bmatrix}\end{aligned}$$

for f_j increasing and differentiable. Note that f_j differentiable because J_S is assumed to exist, and f_j is increasing because J_S is positive definite. The form of S, however, is exactly the form of an element of $\mathcal{F}(P)$ in Definition 2.5 (the constant vector b is a constant of integration). This proves Theorem 2.6. □

Proof of Theorem 2.8. Let us assume that our elementary transformation is $S(x) = P^\top g(Px)$, then note that the Jacobian of S can be given as $J_S(x) = P^\top J_g(Px)P$, where $J_g(z) = \mathrm{diag}((g_j'(z_j))_{j=1}^n)$ (i.e. J_g is a diagonal matrix). Now given our template μ, let's assume that there exists a reference σ

such that the compatibility $S \circ T_\sigma^\mu = T_\sigma^{S_\sharp\mu}$ holds, then we will try to get some necessary conditions that σ must satisfy. In particular, from Theorem 2.1 we can write $T_\sigma^\mu = \nabla\phi$ for some convex ϕ; moreover, we know that the Hessian can be written as $\nabla^2\phi(x) = Q^\top(x)D(x)Q(x)$ for some orthogonal matrix-valued function $Q(x)$ and diagonal matrix-valued function $D(x)$. Now, using Theorem 2.30, we know that if $S \circ T_\sigma^\mu = T_\sigma^{S_\sharp\mu}$, then

$$J_S(\nabla\phi(x))\nabla^2\phi(x) = \nabla^2\phi(x)J_S(\nabla\phi(x))$$
$$P^\top J_g(Px)PQ(x)^\top D(x)Q(x) = Q(x)^\top D(x)Q(x)P^\top J_g(Px)P.$$

Since $J_S(\nabla\phi(x))$ and $\nabla^2\phi(x)$ are two symmetric matrices that commute, we can assume without loss of generality that $Q(x)$ is a row-permutation of P for all x by invoking Lemma 2.29. We can call this matrix $\tilde{P}$. In particular, we can write $\nabla^2\phi(x) = \tilde{P}^\top D(x)\tilde{P}$, where $D(x) = \operatorname{diag}(d(x))$ for a vector-valued function $d(x)$ with $d_i(x) > 0$ (the positivity comes from the fact that the Hessian must have positive eigenvalues).

We see that since $\nabla^2\phi(x)$ has a constant eigendecomposition, we know from Lemma 2.31 that

$$\tilde{P}\nabla\phi(\tilde{P}^\top x) = \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_n(x_n) \end{bmatrix}$$
$$\implies \nabla\phi(\tilde{P}^\top x) = \tilde{P}^\top \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_n(x_n) \end{bmatrix}.$$

From Lemma 2.31, we also note that a choice of the diagonals $d_j(x_j) > 0$ gives a unique (up to a constant) anti-derivative $f_j = \int d_j(x_j)dx_j$. Thus, without loss of generality, we can consider f_j's to be completely determined by the d_j's.

If we assumed that our inputs x are actually written in the basis given by $\tilde{P}^\top$ and the outputs are written in basis given by $\tilde{P}$, then our map transport map decomposes into n single-variable functions as shown above. Moreover, note that $f_j(x_j)$ must be an increasing function since $\frac{\partial f_j}{\partial x_j} > 0$ everywhere. Thus, in principle, this map must be invertible, and we can actually compute the inverse of this map by computing

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \nabla\phi(x) = \nabla\phi(\tilde{P}^\top \tilde{P}x) = \tilde{P}^\top \begin{bmatrix} f_1((\tilde{P}x)_1) \\ f_2((\tilde{P}x)_2) \\ \vdots \\ f_n((\tilde{P}x)_n) \end{bmatrix}$$

$$\tilde{P}y = \begin{bmatrix} f_1((\tilde{P}x)_1) \\ f_2((\tilde{P}x)_2) \\ \vdots \\ f_n((\tilde{P}x)_n) \end{bmatrix}$$

$$\begin{bmatrix} f_1^{-1}((\tilde{P}y)_1) \\ f_2^{-1}((\tilde{P}y)_2) \\ \vdots \\ f_n^{-1}((\tilde{P}y)_n) \end{bmatrix} = \begin{bmatrix} (\tilde{P}x)_1 \\ (\tilde{P}x)_2 \\ \vdots \\ (\tilde{P}x)_n \end{bmatrix} = \tilde{P}x$$

$$\implies \nabla\phi^{-1}(y) = \tilde{P}^\top \begin{bmatrix} f_1^{-1}((\tilde{P}y)_1) \\ f_2^{-1}((\tilde{P}y)_2) \\ \vdots \\ f_n^{-1}((\tilde{P}y)_n) \end{bmatrix}.$$

Note that because the inverse of an increasing function is also increase, we have that $\nabla\phi^{-1} \in \mathcal{F}(P)$.

In practice, we will be given S and μ; thus, we would want to find σ such that T_σ^μ is compatible with S. Note that this will be exactly given by the map $\nabla\phi^{-1}(y)$ because $\sigma = \nabla\phi_\sharp^{-1}\mu$. This proves Theorem 2.8. □

Proof of Theorem 2.13. Given our elementary transformation $S(x) = Ax + b$, we have that $J_S = A$. Theorem 2.30, however, shows us that A must be positive symmetric definite. We will show that the only matrix A that is both positive symmetric definite and orthogonal is the identity. To see this note that since A is symmetric, we know that $A^\top = A$. Since A is assumed to be orthogonal, we know that $A^\top A = A^2 = I$. Let v be an eigenvector of A with eigenvalue λ, then $v = A^2 v = \lambda^2 v$. This means that $\lambda^2 = 1$. Since A is symmetric, we know that all the eigenvalues must be real; thus, $\lambda = \pm 1$. Moreover, because A is positive symmetric definite, the only eigenvalue it could be are $+1$. This implies that A is the identity. In particular, this means that constant rotations are not valid elementary transformations for which the compatibility condition holds. □

2.8 Proofs of Separability Results

For a set of measures μ_1 and μ_2 and a set of elementary transformations $\mathcal{H}$, the general method of showing that $F_\sigma(\mathcal{H}\star\mu_1)$ and $F_\sigma(\mathcal{H}\star\mu_2)$ are linearly separable is to

1. Show that $\mathcal{H}$ is convex,

2. Show that $\mathcal{H}\star\mu_1$ and $\mathcal{H}\star\mu_2$ are compact (or at least have their closures as being compact),

3. Show that $W_2(\mathcal{H}\star\mu_1, \mathcal{H}\star\mu_2) > \delta$ for some $\delta > 0$.

We show this now for shears, but for another class of elementary transformations, we must show that $\mathcal{H}$ is convex.

Lemma 2.32. *The set of shears $\mathcal{H}_{\gamma,M,M_b}$ described in Equation* (2.8) *is convex.*

Proof: Let $h, h' \in \mathcal{H}_{\gamma,M,M_b}$ and $s \in [0,1]$, then we want to show that $sh + (1-s)h' \in \mathcal{H}_{\gamma,M,M_b}$. We find that

$$\begin{aligned} sh(x) + (1-s)h'(x) &= s(Ax+b) + (1-s)(A'x+b') \\ &= (sA + (1-s)A')x + (sb + (1-s)b'). \end{aligned}$$

Notice first that $sA + (1-s)A'$ is symmetric. Moreover, note that

$$\begin{aligned} \lambda_{\min}(sA + (1-s)A') &= \min_{\|x\|_2=1} \langle x, (sA + (1-s)A')x \rangle \\ &= \min_{\|x\|_2=1} s\langle x, Ax \rangle + (1-s)\langle x, A'x \rangle \\ &\geq s \underbrace{\min_{\|x\|_2=1} \langle x, Ax \rangle}_{\geq \lambda_{\min}(A)} + (1-s) \underbrace{\min_{\|\tilde{x}\|_2=1} \langle \tilde{x}, A'\tilde{x} \rangle}_{\geq \lambda_{\min}(A')} \\ &\geq s\lambda_{\min}(A) + (1-s)\lambda_{\min}(A') > s\gamma + (1-s)\gamma = \gamma; \end{aligned}$$

and similarly,

$$\begin{aligned} \lambda_{\max}(sA + (1-s)A') &= \max_{\|x\|_2=1} \langle x, (sA + (1-s)A')x \rangle \\ &= \max_{\|x\|_2=1} s\langle x, Ax \rangle + (1-s)\langle x, A'x \rangle \\ &\leq s \underbrace{\max_{\|x\|_2=1} \langle x, Ax \rangle}_{\leq \lambda_{\max}(A)} + (1-s) \underbrace{\max_{\|\tilde{x}\|_2=1} \langle \tilde{x}, A'\tilde{x} \rangle}_{\leq \lambda_{\max}(A')} \\ &\leq s\lambda_{\max}(A) + (1-s)\lambda_{\max}(A') < sM + (1-s)M = M. \end{aligned}$$

This means that $sA + (1-s)A'$ is symmetric positive definite and actually has the correct bounds on its eigenvalues. We now show that $sb+(1-s)b'$ satisfies the proper bounds too. Notice that

$$\|sb+(1-s)b'\|_2 \leq s\|b\|_2+(1-s)\|b'\|_2 \leq sM_b+(1-s)M_b = M_b.$$

This implies that $sh+(1-s)h' \in \mathcal{H}$. So we're done. □

Next, given a base measure μ and set of elementary transformations $\mathcal{H}$, we ideally want to show that the set $\mathcal{H}\star\mu = \{h_\sharp\mu : h \in \mathcal{H}\}$ is compact, but the weaker condition of $\mathcal{H}\star\mu$ being precompact should be good enough for our purposes. To address compactness, we need a definition.

Definition 2.33 (Tightness)**.** *Let $(X,\mathcal{T})$ be a Hausdorff space and let $\mathcal{S}$ be a σ-algebra such that $\mathcal{T} \subseteq \mathcal{S}$. Let M be a collection of probability measures defined on $\mathcal{S}$. The collection M is called **tight** if, for any $\varepsilon > 0$, there exists a compact subset $K_\varepsilon \subset X$ such that for all measures $\mu \in M$, we have $\mu(K_\varepsilon) > 1-\varepsilon$.*

A natural theorem that relates tightness of measures to compactness is Prokhorov's theorem.

Theorem 2.34 (Prokhorov)**.** *Let (X,d) be a a separable metric space. Let $\mathcal{P}(X)$ be the collection of all probability measures defined on X with respect to the Borel σ-algebra. Then a collection $\mathcal{K} \subset \mathcal{P}(X)$ of probability measures is tight if and only if the closure of $\mathcal{K}$ is sequentially compact in $\mathcal{P}_2(X)$ equipped with the topology of weak convergence.*

According to [77, pp. 37–42], we can upgrade Prokhorov's theorem to be sequentially compact with the Wasserstein 2-metric if

$$\sup_{\mu\in\mathcal{K}} \int_{x:\|x\|_2>R} \|x\|_2^2 d\mu(x) \xrightarrow{R\to\infty} 0.$$

This is easily true if $\sup_{\mu\in\mathcal{K}}\{\|x\|_2 : x \in \mathrm{supp}(\mu)\} \leq R < \infty$.

Corollary 2.35. *Let $\mathcal{H}$ be a set of transformations such that for every $R > 0$, there exists $\tilde{R}$ such* *that* $\sup_{h\in\mathcal{H}}\{\|h(x)\|_2 : \|x\|_2 < R\} < \tilde{R}$*. Also assume that* $\mu \in \mathcal{P}_2(\mathbb{R}^n)$ *has bounded support* R_μ*, then* $\mathcal{H}\star\mu$ *is a precompact set of measures.*

Proof. For us, if μ has bounded support with bound R_μ, we should have that all measures belonging to $\mathcal{H}\star\mu$ must also have support bounded for some $\tilde{R} > 0$. To see this, note that for $\tilde{\mu} \in \mathcal{H}\star\mu$, we have $\text{supp}(\tilde{\mu})$ is bounded by $\tilde{R}$ for some $\tilde{R} > 0$. So we're done. □

For shears, we can see that every measure from $\mathcal{H}_{\gamma,M,M_b}\star\mu_1$ and $\mathcal{H}_{\gamma,M,M_b}\star\mu_2$ has bounded support since $\sup_{h\in\mathcal{H}_{\gamma,M,M_b}}\{\|h(x)\| : x\in\text{supp}(\mu), \mu\in\mathcal{K}\} \leq MR + M_b$. It's easy to see that $\mathcal{H}_{\gamma,M,M_b}\star\mu$ is tight for a big enough ball $B_R(0) = \{x : \|x\|_2 \leq R\}$ if σ has bounded support. This means that $\mathcal{H}_{\gamma,M,M_b}\star\mu$ is precompact with the Wasserstein 2-metric for any μ with bounded support.

By [91] Corollary 5.23, the stability of optimal transport maps implies that F_σ is continuous; thus, we find that $F_\sigma(\mathcal{H}\star\mu)$ is precompact if $\mathcal{H}\star\mu$ is precompact. Note also that Theorem 2.1 above gives us a corollary.

Corollary 2.36. *Let* $h : \mathbb{R}^n \to \mathbb{R}^n$ *be a transformations that can be represented as the gradient of a convex function, then for* σ*, an absolutely continuous measure with respect to the Lebesgue measure, we get that* $h = T_\sigma^{h_\sharp\sigma}$*.*

Now we must show that $F_\sigma(\mathcal{H}\star\mu)$ is convex, which will ensure that our LOT embedding is convex and precompact.

Lemma 2.37. *Let* σ *and* μ *be absolutely continuous (with respect to the Lebesgue measure) probability measures and let* $\mathcal{H} \subseteq \{h : \mathbb{R}^n \to \mathbb{R}^n | h = \nabla\phi, \phi \text{ is convex}\}$ *be a convex set of transformations that is compatible with* σ *and* μ*, then* $F_\sigma(\mathcal{H}\star\mu)$ *is convex.*

Proof. Let $h, \hat{h} \in \mathcal{H}$ and $s \in [0,1]$ so that $T_\sigma^{h_\sharp\mu}, T_\sigma^{\hat{h}_\sharp\mu} \in F_\sigma(\mathcal{H}\star\mu)$. Then we want to show that $sT_\sigma^{h_\sharp\mu} + (1-s)T_\sigma^{\hat{h}_\sharp\mu} \in F_\sigma(\mathcal{H}\star\mu)$. First notice that by Brenier's theorem, there exists convex functions ϕ and $\hat{\phi}$ such that $\nabla\phi = T_{\sigma}^{h_\sharp\mu}$ and $\nabla\hat{\phi} = T_{\sigma}^{\hat{h}_\sharp\mu}$. Note now that

$$s\nabla\phi + (1-s)\nabla\hat{\phi} = \nabla(s\phi + (1-s)\hat{\phi})$$

so that $sT_{\sigma}^{h_\sharp\mu} + (1-s)T_{\sigma}^{\hat{h}_\sharp\mu}$ is actually the gradient of a convex function. Moreover, by the uniqueness of optimal transport maps as gradients of convex functions, we know that $sT_{\sigma}^{h_\sharp\mu} + (1-s)T_{\sigma}^{\hat{h}_\sharp\mu}$ is the unique optimal transport map that transports σ to its target distribution. If this target distribution is of the form $\tilde{h}_\sharp\mu$ for some $\tilde{h} \in \mathcal{H}$, then our proof is done. Indeed, using the compatibility of of h and $\hat{h}$:

$$\begin{aligned}\left(sT_{\sigma}^{h_\sharp\mu} + (1-s)T_{\sigma}^{\hat{h}_\sharp\mu}\right)_\sharp \sigma &= \left(\left(sh + (1-s)\hat{h}\right) \circ T_{\sigma}^{\mu}\right)_\sharp \sigma \\ &= \left(sh + (1-s)\hat{h}\right)_\sharp \mu.\end{aligned}$$

Since $sh + (1-s)\hat{h} \in \mathcal{H}$, we know that $sT_{\sigma}^{h_\sharp\mu} + (1-s)T_{\sigma}^{\hat{h}_\sharp\mu}$ is the unique optimal transport map that transports σ to $(sh + (1-s)\hat{h})_\sharp\mu$. This means that

$$sT_{\sigma}^{h_\sharp\mu} + (1-s)T_{\sigma}^{\hat{h}_\sharp\mu} \in F_{\sigma}(\mathcal{H} \star \mu).$$

Thus, $F_{\sigma}(\mathcal{H} \star \mu)$ is convex. □

Using the lemma above, we get that $F_{\sigma}(\mathcal{H} \star \mu_1)$ and $F_{\sigma}(\mathcal{H} \star \mu_2)$ are both convex and have compact closures. For our linear separability result, we now only need to make sure that $\inf_{h,h' \in \mathcal{H}} \|T_{\sigma}^{h_\sharp\mu_1} - T_{\sigma}^{h'_\sharp\mu_2}\|_{\sigma} \geq \delta$ for some $\delta > 0$. Ideally, given $W_2(\mu_1, \mu_2)$ and the level of separation $\delta > 0$ we want, we should be able to find bounds on the function class $\mathcal{H}$ that we are considering. This leads us to Theorem 2.15:

Proof of Theorem 2.15. Assume that we have $\tilde{h}, \tilde{h}^{\star} \in \mathcal{H}_{\varepsilon}$, then using the triangle inequality, we have

$$\begin{aligned} W_2(\tilde{h}_\sharp\mu_1, \tilde{h}^\star_\sharp\mu_2) &\geq |W_2(\mu_1, \tilde{h}^\star_\sharp\mu_2) - W_2(\tilde{h}_\sharp\mu_1, \mu_1)| \\ &\geq \left| \left| W_2(\mu_1,\mu_2) - W_2(\mu_2, \tilde{h}^\star_\sharp\mu_2) \right| - W_2(\tilde{h}_\sharp\mu_1, \mu_1) \right|, \end{aligned}$$

provided that the quantity in the left-hand side is positive. Now, we know from [71] that $W_2(\mu,\nu) \leq \|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma$; thus, we have that

$$\left| \left| W_2(\mu_1,\mu_2) - W_2(\mu_2, \tilde{h}^\star_\sharp\mu_2) \right| - W_2(\tilde{h}_\sharp\mu_1, \mu_1) \right| \leq \|F_\sigma(\tilde{h}_\sharp\mu_1) - F_\sigma(\tilde{h}^\star_\sharp\mu_2)\|_\sigma.$$

So if we lower bound the left-hand side by $\delta > 0$, then $\|F_\sigma(\tilde{h}_\sharp\mu_1) - F_\sigma(\tilde{h}^\star_\sharp\mu_2)\|_\sigma \geq \delta > 0$. This would imply that $F_\sigma(\mathcal{H}_\varepsilon \star \mu_1)$ and $F_\sigma(\mathcal{H}_\varepsilon \star \mu_2)$ is linearly separable by the Hahn-Banach theorem.

To get this bound, let us find a generic bound for $W_2(\tilde{h}_\sharp\mu, \mu)$ when $\tilde{h} \in \mathcal{H}_\varepsilon$. In particular, there exists $h \in \mathcal{H}$ such that $\|h - \tilde{h}\|_\mu$; thus, we get

$$W_2(\tilde{h}_\sharp\mu, \mu) \leq W_2(\tilde{h}_\sharp\mu, h_\sharp\mu) + W_2(h_\sharp\mu, \mu).$$

First, since h is the gradient of convex function and Corollary 2.36, we know that $T_\mu^{h_\sharp\mu} = h$. This means that the compatibility condition holds, which further implies that

$$W_2(\tilde{h}_\sharp\mu, \mu) = \|\tilde{h} - I\|_\mu \leq L.$$

Moreover, equation 2.1 of [6] says that

$$W_2(\tilde{h}_\sharp\mu, h_\sharp\mu) \leq \|h - \tilde{h}\|_\mu < \varepsilon.$$

Because of our bounds, our results implies that

$$\begin{aligned} & L \leq \frac{W_2(\mu_1,\mu_2)-\delta}{2} - \varepsilon \leq W_2(\mu_1,\mu_2) - \delta - \varepsilon \\ \implies & W_2(\mu_1,\mu_2) - W_2(\mu_2,\tilde{h}^{\star}_{\sharp}\mu_2) \geq W_2(\mu_1,\mu_2) - W_2(\tilde{h}^{\star}_{\sharp}\mu_2, h^{\star}_{\sharp}\mu_2) - W_2(h^{\star}_{\sharp}\mu_2,\mu_2) \\ & \geq W_2(\mu_1,\mu_2) - L - \varepsilon > \delta > 0. \end{aligned}$$

Essentially, we were able to remove the absolute values because the quantity in the absolute value was positive. This positivity of the absolute value implies that we can replace

$$\left| \left| W_2(\mu_1,\mu_2) - W_2(\mu_2,\tilde{h}^{\star}_{\sharp}\mu_2) \right| - W_2(\tilde{h}_{\sharp}\mu_1,\mu_1) \right|$$

with

$$\left| W_2(\mu_1,\mu_2) - W_2(\mu_2,\tilde{h}^{\star}_{\sharp}\mu_2) - W_2(\tilde{h}_{\sharp}\mu_1,\mu_1) \right|.$$

But note that

$$\begin{aligned} W_2(\mu_1,\mu_2) - W_2(\mu_2,\tilde{h}^{\star}_{\sharp}\mu_2) - W_2(\tilde{h}_{\sharp}\mu_1,\mu_1) & \geq W_2(\mu_1,\mu_2) - 2L - 2\varepsilon \geq \delta \\ \iff L & \leq \frac{W_2(\mu_1,\mu_2)-\delta}{2} - \varepsilon. \end{aligned}$$

This implies that

$$\left| W_2(\mu_1,\mu_2) - W_2(\mu_2,\tilde{h}^{\star}_{\sharp}\mu_2) - W_2(\tilde{h}_{\sharp}\mu_1,\mu_1) \right| \geq \delta > 0.$$

So we see that if $L \leq \frac{W_2(\mu_1,\mu_2)-\delta}{2} - \varepsilon$, then we must have that $\|F_\sigma(h_\sharp\mu_1) - F_\sigma(h'_\sharp\mu_2)\|_\sigma \geq \delta$. □

Proof of Theorem 2.19. For the first statement, the linear separability result is immediate because the compatibility criteria ensures that the LOT distance and Wasserstein-2 distance are the same.

To see this, we note that $h \circ T_{\mu_1}^{\mu_2}$ is compatible with respect to the optimal transport between μ_1 and $(h \circ T_{\mu_1}^{\mu_2})_\sharp \mu_1$ because

$$T_{\mu_1}^{(h \circ T_{\mu_1}^{\mu_2})_\sharp \mu_1} = T_{\mu_1}^{h_\sharp \mu_2} = h \circ T_{\mu_1}^{\mu_2} = h \circ T_{\mu_1}^{\mu_2} \circ T_{\mu_1}^{\mu_1}.$$

This means that from [71], for $h, \tilde{h} \in \mathcal{H}$ we have that

$$\begin{aligned} \|T_{\mu_1}^{\tilde{h}_\sharp \mu_1} - T_{\mu_1}^{h_\sharp \mu_2}\|_{\mu_1} &= \|T_{\mu_1}^{\tilde{h}_\sharp \mu_1} - T_{\mu_1}^{(h \circ T_{\mu_1}^{\mu_2})_\sharp \mu_1}\|_{\mu_1} \\ &= W_2(\tilde{h}_\sharp \mu_1, (h \circ T_{\mu_1}^{\mu_2})_\sharp \mu_1) \\ &= W_2(\tilde{h}_\sharp \mu_1, h_\sharp \mu_2). \end{aligned}$$

This proves the first statement.

For the second statement, let $h_\varepsilon, \tilde{h}_\varepsilon \in \mathcal{H}_\varepsilon$ such that $\|h - h_\varepsilon\|_{\mu_1} < \varepsilon$ and $\|\tilde{h} - \tilde{h}_\varepsilon\|_{\mu_1}$ for $h, \tilde{h} \in \mathcal{H}$. We know that $\|F_{\mu_1}((\tilde{h}_\varepsilon)_\sharp \mu_1) - F_{\mu_2}((h_\varepsilon)_\sharp \mu_2)\|_{\mu_1} \geq W_2((\tilde{h}_\varepsilon)_\sharp \mu_1, (h_\varepsilon)_\sharp \mu_2)$. Now we know that

$$W_2((\tilde{h}_\varepsilon)_\sharp \mu_1, (h_\varepsilon)_\sharp \mu_2) \geq \left| \left| W_2(\tilde{h}_\sharp \mu_1, h_\sharp \mu_2) - W_2(\tilde{h}_\sharp \mu_1, (\tilde{h}_\varepsilon)_\sharp \mu_1) \right| - W_2(h_\sharp \mu_2, (h_\varepsilon)_\sharp \mu_2) \right|.$$

From equation 2.1 of [5], we have that

$$\begin{aligned} W_2(\tilde{h}_\sharp \mu_1, (\tilde{h}_\varepsilon)_\sharp \mu_1) &\leq \|\tilde{h} - \tilde{h}_\varepsilon\|_{\mu_1} < \varepsilon \\ W_2(h_\sharp \mu_2, (h_\varepsilon)_\sharp \mu_2) &\leq \|h - h_\varepsilon\|_{\mu_2} < \varepsilon. \end{aligned}$$

Note that $W_2(\tilde{h}_\sharp\mu_1, h_\sharp\mu_2) \geq \inf_{h,\tilde{h}\in\mathcal{H}} W_2(\tilde{h}_\sharp\mu_1, h_\sharp\mu_2) > 2\varepsilon$. This means that

$$W_2((\tilde{h}_\varepsilon)_\sharp\mu_1, (h_\varepsilon)_\sharp\mu_2) \geq \underbrace{\left| \underbrace{\left| \inf_{h,\tilde{h}\in\mathcal{H}} W_2(\tilde{h}_\sharp\mu_1, h_\sharp\mu_2) - \varepsilon \right|}_{>0} - \varepsilon \right|}_{>0} > 0.$$

So we have that $\|F_{\mu_1}((\tilde{h}_\varepsilon)_\sharp\mu_1) - F_{\mu_2}((h_\varepsilon)_\sharp\mu_2)\|_{\mu_1} > 0$.

For the third statement, we extend the lower bounds from above. Because $h, \tilde{h} \in \mathcal{H}$ are compatible, we have that $W_2(\tilde{h}_\sharp\mu_1, h_\sharp\mu_2) = \|T_{\mu_1}^{\tilde{h}_\sharp\mu_1} - T_{\mu_1}^{h_\sharp\mu_2}\|_{\mu_1}$. Using the triangle inequality, we get

$$\begin{aligned}
\|T_{\mu_1}^{\tilde{h}_\sharp\mu_1} - T_{\mu_1}^{h_\sharp\mu_2}\|_{\mu_1} &= \|\tilde{h} - T_{\mu_1}^{h_\sharp\mu_2}\|_{\mu_1} \\
&= \|\tilde{h} - h - (T_{\mu_1}^{h_\sharp\mu_2} - h)\|_{\mu_1} \\
&\geq \left| \|T_{\mu_1}^{h_\sharp\mu_2} - T_{\mu_1}^{h_\sharp\mu_1}\|_{\mu_1} - \|\tilde{h} - h\|_{\mu_1} \right|.
\end{aligned}$$

Because $h \in \mathcal{H}$ is chosen to be compatible with respect to μ_1 and μ_2, note that

$$\begin{aligned}
W_2(h_\sharp\mu_1, h_\sharp\mu_2) &= \|T_{\mu_1}^{h_\sharp\mu_1} - T_{\mu_1}^{h_\sharp\mu_2}\|_{\mu_1} = \|h - h \circ T_{\mu_1}^{\mu_2}\|_{\mu_1} \\
&= \|h \circ (I - T_{\mu_1}^{\mu_2})\|_{\mu_1} = \|h\|_{|\mu_1 - \mu_2|} \\
&= \left(\int \|h(x)\|_2^2 d|\mu_1 - \mu_2|(x) \right)^{1/2} \\
&\geq \sqrt{2} \left(\int \|x_0 - x\|_2^2 d|\mu_1 - \mu_2|(x) \right)^{1/2}
\end{aligned}$$

where the last equality came from a change of variable and the inequality comes from our assumption that $\|h(x)\|_2 \geq \sqrt{2}\|x - x_0\|_2$. Now we refer to Theorem 6.15 of [91], which says that for any $x_0 \in \mathbb{R}^n$, we have

$$W_2(\mu_1,\mu_2) \leq \sqrt{2}\bigg(\int \|x_0 - x\|_2^2 d|\mu_1 - \mu_2|(x)\bigg)^{1/2} = f(x_0).$$

We want to minimize the right hand side; thus, taking the derivative $\frac{d}{dx_0} f(x_0) = 0$, this reduces to

$$\begin{aligned} 0 &= 2\int (x_0 - x) d|\mu_1 - \mu_2|(x) \\ \implies x_0 &= \frac{1}{|\mu_1 - \mu_2|(\mathbb{R}^n)} \int x d|\mu_1 - \mu_2|(x). \end{aligned}$$

Essentially x_0 is the mean of the measure $|\mu_1 - \mu_2|$ after normalization. So we have that $W_2(\mu_1,\mu_2) \leq W_2(h_\sharp \mu_1, h_\sharp \mu_2)$. Since $W_2(\mu_1,\mu_2) - \sup_{h,\tilde{h}\in\mathcal{H}} \|\tilde{h} - h\|_{\mu_1} \geq \delta + 2\varepsilon > \delta + \varepsilon$, these computations imply that

$$\Big|W_2(\tilde{h}_\sharp \mu_1, h_\sharp \mu_2) - W_2(\tilde{h}_\sharp \mu_1, (\tilde{h}_\varepsilon)_\sharp \mu_1)\Big| - W_2(h_\sharp \mu_2, (h_\varepsilon)_\sharp \mu_2) \geq W_2(\tilde{h}_\sharp \mu_1, h_\sharp \mu_2) - 2\varepsilon.$$

is greater than

$$W_2(\mu_1,\mu_2) - \sup_{h,\tilde{h}\in\mathcal{H}} \|\tilde{h} - h\|_{\mu_1} - 2\varepsilon > 0.$$

This implies that

$$W_2((\tilde{h}_\varepsilon)_\sharp \mu_1, (h_\varepsilon)_\sharp \mu_2) \geq \Big|W_2(\mu_1,\mu_2) - \sup_{h,\tilde{h}\in\mathcal{H}} \|\tilde{h} - h\|_{\mu_1} - 2\varepsilon\Big| \geq \delta$$

This implies that $\|F_{\mu_1}((\tilde{h}_\varepsilon)_\sharp \mu_1) - F_{\mu_2}((h_\varepsilon)_\sharp \mu_2)\|_{\mu_1} \geq W_2((\tilde{h}_\varepsilon)_\sharp \mu_1, (h_\varepsilon)_\sharp \mu_2) \geq \delta$. So we are done.

□

Notice that Theorem 2.15 above acts as a blueprint to controlling the degree of separation in the LOT embedding via the bounds on the function class $\mathcal{H}$. For the specific setting of the set of shears above, given a desired degree of separation $0 < \delta < W_2(\mu_1, \mu_2)$, we can choose M, M_b, and γ in the definition of $\mathcal{H}_{\gamma,M}$ that guarantees that $F_\sigma(\mathcal{H}_{\gamma,M,M_b} \star \mu_1)$ and $F_\sigma(\mathcal{H}_{\gamma,M,M_b} \star \mu_2)$ are δ-separated. This leads us to Corollary 2.38:

Corollary 2.38. *Consider probability distributions μ_1 and μ_2 with Wasserstein distance $W_2(\mu_1, \mu_2)$, and let $\delta > 0$. Let us denote $R_1 = \max_{x \in supp(\mu_1)} \|x\|_2$ and $R_2 = \max_{x \in supp(\mu_2)} \|x\|_2$. Moreover, for $\varepsilon > 0$, define*

$$\mathcal{H}_{\gamma,M,M_b,\varepsilon} = \{\tilde{h} : \|h - \tilde{h}\|_{\mu_i} < \varepsilon, i \in \{1,2\}, h \in \mathcal{H}_{\gamma,M,M_b}\}$$

as the ε-tube around $\mathcal{H}_{\gamma,M,M_b}$. Assume $\sigma \ll \lambda \in \mathcal{P}_2(\mathbb{R}^n)$ is chosen such that $\mathcal{H}$ is compatible with σ and μ_1 as well as σ and μ_2. We consider the following 2 cases:

Case 1: *Assume that $W_2(\mu_1, \mu_2) > (R_1 + R_2) + \delta + 2\varepsilon$. If M_b is chosen such that*

$$0 \leq M_b < \frac{W_2(\mu_1, \mu_2) - \delta - 2\varepsilon - (R_1 + R_2)}{2},$$

then choosing M such that

$$2 < M \leq \frac{W_2(\mu_1, \mu_2) - \delta - 2\varepsilon - 2M_b + (R_1 + R_2)}{R_1 + R_2}$$

ensures that $F_\sigma(\mathcal{H}_{\gamma,M,M_b,\varepsilon} \star \mu_1)$ and $F_\sigma(\mathcal{H}_{\gamma,M,M_b,\varepsilon} \star \mu_2)$ are δ-separated.

Case 2: *Assume that $\delta + 2\varepsilon < W_2(\mu_1, \mu_2) < (R_1 + R_2) + \delta + 2\varepsilon$. If M_b is chosen such that*

$$\max\left\{0, \frac{W_2(\mu_1, \mu_2) - 2\varepsilon - \delta - (R_1 + R_2)}{2}\right\} \leq M_b < \frac{W_2(\mu_1, \mu_2) - 2\varepsilon - \delta}{2},$$

then either choosing M such that

$$1 < M \leq \frac{W_2(\mu_1, \mu_2) - \delta - 2\varepsilon - 2M_b + (R_1 + R_2)}{R_1 + R_2}$$

or choosing γ *such that*

$$\gamma \geq \frac{2M_b + 2\varepsilon + \delta - W_2(\mu_1,\mu_2) + R_1 + R_2}{R_1 + R_2}$$

ensures that $F_\sigma(\mathcal{H}_{\gamma,M,M_b,\varepsilon} \star \mu_1)$ *and* $F_\sigma(\mathcal{H}_{\gamma,M,M_b,\varepsilon} \star \mu_2)$ *are* δ*-separated.*

Proof of Corollary 2.38. From the lemma above, we need to only bound $\|h - I\|_\sigma$ appropriately and invert the bounds. First, note that because $h \in \mathcal{H}_{\gamma,M,M_b}$ can be written as the gradient of a convex function and $\mathcal{H}_{\gamma,M,M_b}$ is a convex set, we do satisfy the setting of the lemma. Moreover, we know that the compatibility condition holds, which implies that

$$\begin{aligned} W_2(h_\sharp\mu,\mu) = \|h - I\|_\mu &= \left(\int \|(A - I)x + b\|_2^2 d\mu(x) \right)^{\frac{1}{2}} \\ &\leq \underbrace{\|(A - I)x\|_\mu}_{I_1} + \underbrace{\|b\|_\mu}_{I_2}. \end{aligned}$$

Let us bound I_1 and I_2 separately. For the bound of I_1, we have

$$\begin{aligned} I_1 &= \left(\int \|(A - I)x\|_2^2 d\mu(x) \right)^{\frac{1}{2}} \\ &\leq \left(\int \lambda_{\max}(A - I)^2 \|x\|_2^2 d\mu(x) \right)^{\frac{1}{2}} \\ &\leq \left(\int \lambda_{\max}(A - I)^2 \max_{x \in \operatorname{supp}(\mu)} \|x\|_2^2 d\mu(x) \right)^{\frac{1}{2}} \\ &= \lambda_{\max}(A - I) \max_{x \in \operatorname{supp}(\mu)} \|x\|_2 \underbrace{\left(\int d\mu(x) \right)^{\frac{1}{2}}}_{1} \\ &\leq \max\{|M - 1|, |1 - \gamma|\} \max_{x \in \operatorname{supp}(\mu)} \|x\|_2. \end{aligned}$$

For the bound of I_2, we have

$$I_2 = \|b\|_\mu = \left(\int \|b\|_2^2 d\mu(x) \right)^{\frac{1}{2}} \leq \left(\int M_b^2 d\mu(x) \right)^{\frac{1}{2}} = M_b.$$

Thus, if $h \in \mathcal{H}_{\gamma,M,M_b}$, we have

$$W_2(h_\sharp \mu, \mu) \leq \max\{|M-1|, |1-\gamma|\} \max_{x \in \operatorname{supp}(\mu)} \|x\|_2 + M_b.$$

Using this for our specific choice of μ_1 and μ_2, we find that for $\tilde{h}, \tilde{h}^\star \in \mathcal{H}_{\gamma,M,M_b,\varepsilon}$, we have

$$W_2(\mu_1,\mu_2) - W_2(\tilde{h}_\sharp \mu_1, \mu_1) - W_2(\tilde{h}^\star_\sharp \mu_2, \mu_2)$$

is lower bounded (via equation 2.1 of [6]) by

$$\begin{aligned} W_2(\mu_1,\mu_2) - W_2(h_\sharp \mu_1, \mu_1) - \underbrace{W_2(h_\sharp \mu_1, \tilde{h}_\sharp \mu_1)}_{\leq \|h - \tilde{h}\|_{\mu_1} < \varepsilon} - W_2(h^\star_\sharp \mu_2, \mu_2) - \underbrace{W_2(h^\star_\sharp \mu_2, \tilde{h}^\star_\sharp \mu_2)}_{\leq \|h^\star \tilde{h}^\star\|_{\mu_2} < \varepsilon} \\ \geq W_2(\mu_1,\mu_2) - W_2(h_\sharp \mu_1, \mu_1) - W_2(h^\star_\sharp \mu_2, \mu_2) - 2\varepsilon, \end{aligned}$$

which in turn is lower bounded by

$$W_2(\mu_1,\mu_2) - 2M_b - \max\{|M-1|, |1-\gamma|\} \underbrace{\left(\max_{x \in \operatorname{supp}(\mu_1)} \|x\|_2 + \max_{x \in \operatorname{supp}(\mu_2)} \|x\|_2 \right)}_{R_1 + R_2} - 2\varepsilon.$$

Now we just need to find sufficient conditions M, M_b, and γ such that

$$W_2(\mu_1,\mu_2) - 2M_b - \max\{|M-1|, |1-\gamma|\}(R_1 + R_2) - 2\varepsilon \geq \delta > 0. \qquad (\star)$$

Notice that when $|M-1| > |1-\gamma|$, then $M > 1$ since M is the bound on the largest eigenvalue.

Moreover, note that we cannot have $\gamma - 1 > M - 1$ since $\gamma < M$; thus, the only cases we need to consider are when $M - 1 > 1 - \gamma$ and $M - 1 < 1 - \gamma$. We handle these cases separately.

Case 1 ($M - 1 > 1 - \gamma$): Note that in this case we can rewrite $(\star)$ as

$$\begin{aligned}
M(R_1 + R_2) &\leq W_2(\mu_1, \mu_2) - 2M_b - 2\varepsilon - \delta + (R_1 + R_2) \\
M &\leq \frac{W_2(\mu_1, \mu_2) - 2M_b - 2\varepsilon - \delta + (R_1 + R_2)}{R_1 + R_2}.
\end{aligned}$$

Case 2 ($1 - \gamma > M - 1$): In this case, we can rewrite $(\star)$ as

$$\begin{aligned}
\gamma(R_1 + R_2) &\geq \delta + 2M_b + 2\varepsilon + (R_1 + R_2) - W_2(\mu_1, \mu_2) \\
\gamma &\geq \frac{\delta + 2M_b + 2\varepsilon + (R_1 + R_2) - W_2(\mu_1, \mu_2)}{R_1 + R_2}.
\end{aligned}$$

Now we will investigate conditions in which case 1 and case 2 are active.

First note that if

$$\begin{aligned}
&\frac{W_2(\mu_1, \mu_2) - 2M_b - 2\varepsilon - \delta + (R_1 + R_2)}{R_1 + R_2} > 2 \\
\iff &\frac{W_2(\mu_1, \mu_2) - 2\varepsilon - \delta - (R_1 + R_2)}{2} > M_b > 0 \\
\iff &W_2(\mu_1, \mu_2) > \delta + 2\varepsilon + (R_1 + R_2),
\end{aligned}$$

we know that the first case is ensured since we can pick $M > 2$. In this case, $M - 1 > |1 - \gamma|$. To see this, we see that if $0 < \gamma < 1$, then $M - 1 > 2 - 1 = 1 > 1 - \gamma > 0$. If $1 < \gamma < 2$, we again have that $M > \gamma$ implies that $M - 1 > \gamma - 1$. Thus, in this regime, the choice of M dominates.

Now if we want $1 < M < 2$, we find that

$$\begin{aligned}
M \leq &\frac{W_2(\mu_1, \mu_2) - 2M_b - 2\varepsilon - \delta + (R_1 + R_2)}{R_1 + R_2} < 2 \\
\iff &\frac{W_2(\mu_1, \mu_2) - 2\varepsilon - \delta - (R_1 + R_2)}{2} < M_b.
\end{aligned}$$

Notice that since $M_b \geq 0$, if $W_2(\mu_1,\mu_2) < 2\varepsilon + \delta + (R_1 + R_2)$, then we definitely have the above inequality. On the other hand,

$$\begin{aligned}
\frac{W_2(\mu_1,\mu_2) - 2M_b - 2\varepsilon - \delta + (R_1 + R_2)}{R_1 + R_2} > 1 \\
\iff \frac{W_2(\mu_1,\mu_2) - 2\varepsilon - \delta}{2} > M_b > 0 \\
\iff W_2(\mu_1,\mu_2) > 2\varepsilon + \delta.
\end{aligned}$$

So we can pick appropriate M_b such that

$$\max\left\{\frac{W_2(\mu_1,\mu_2) - 2\varepsilon - \delta - (R_1 + R_2)}{2}, 0\right\} < M_b < \frac{W_2(\mu_1,\mu_2) - 2\varepsilon - \delta}{2},$$

and in this case, we pick an appropriate M such that

$$1 < \frac{W_2(\mu_1,\mu_2) - 2\varepsilon - 2M_b - \delta + (R_1 + R_2)}{R_1 + R_2} \leq M < 2.$$

In the case when $|1 - \gamma| > |M - 1|$ case, notice that

$$\frac{\delta + 2\varepsilon + 2M_b + (R_1 + R_2) - W_2(\mu_1,\mu_2)}{R_1 + R_2} < 1 \iff \exists M_b < \frac{W_2(\mu_1,\mu_2) - \delta - 2\varepsilon}{2};$$

thus, we can pick γ such that

$$1 > \gamma \geq \max\left\{\frac{\delta + 2M_b + 2\varepsilon + (R_1 + R_2) - W_2(\mu_1,\mu_2)}{R_1 + R_2}, 0\right\}.$$

So we still can satisfy the conditions for linear separability in these cases. □

2.9 Multiple References Example

Example 2.39. *Recall the setup of Example 2.26, where we have two template distributions $\mu_1 = \mathcal{N}(0, \Sigma_1)$ and $\mu_2 = \mathcal{N}(0, \Sigma_2)$, a set of shears*

$$\mathcal{H} = \{Ax : A = A^\top \in \mathbb{R}^{n \times n}, MI_n \succeq A \succeq mI_n \succ 0\}$$

as our set of transformations, and reference distributions defined to be of the form $\sigma_1 = (h_1)_\sharp \mu_1$ and $\sigma_2 = (h_2)_\sharp \mu_2$ for $h_1(x) = A_1 x$ and $h_2(x) = A_2 x$ for $h_1, h_2 \in \mathcal{H}$ so that

$$\sigma_1 = (h_1)_\sharp \mu_1 = \mathcal{N}(0, A_1 \Sigma_1 A_1^\top), \quad \sigma_2 = (h_2)_\sharp \mu_2 = \mathcal{N}(0, A_2 \Sigma_2 A_2^\top).$$

Using exercise 6.3.1 of [89], the bounds on our function class $\mathcal{H}$ is given by

$$\begin{aligned}\sup_{A \in \mathcal{H}} \|(A - I)\|_{\mu_j} &= \sup_{A \in \mathcal{H}} \left(\mathbb{E}_{\mu_j} \left[\|(A - I)x\|_2^2 \right] \right)^{1/2} = \sup_{A \in \mathcal{H}} \|(A - I)\Sigma_j^{1/2}\|_F \\ &\leq \sup_{A \in \mathcal{H}} \|A - I\|_2 \|\Sigma_j^{1/2}\|_F \leq \max\left(|M - 1|, |1 - m|\right) \max_j \|\Sigma_j^{1/2}\|_F \\ &= L.\end{aligned}$$

To ensure separation, we use $L \leq \frac{W_2(\mu_1, \mu_2) - \delta}{2}$, which implies that

$$\max\left(|M - 1|, |1 - m|\right) \leq \frac{\mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}})^{1/2})^{1/2} - \delta}{2 \max_{j=1,2} \|\Sigma_j^{1/2}\|_F}.$$

It is easy to see that $\frac{W_2(\mu_1, \mu_2)}{2 \max_{j=1,2} \|\Sigma_j^{1/2}\|_F} < 1$. This shows the bounds on M and m of Example 2.26. Now notice that

$$T_{\sigma_j}^{h_\sharp \mu_i} = (A_j \Sigma_j A_j)^{-1/2} ((A_j \Sigma_j A_j)^{1/2} (A \Sigma_i A^\top)(A_j \Sigma_j A_j)^{1/2})^{1/2} (A_j \Sigma_j A_j)^{-1/2} x;$$

thus, for $h,\tilde{h} \in \mathcal{H}$ where $h(x) = Ax$ and $\tilde{h}(x) = \tilde{A}x$, we get $\|T_{\sigma_j}^{h_\sharp\mu_1} - T_{\sigma_j}^{\tilde{h}_\sharp\mu_2}\|_{\sigma_j}^2$ is equal to

$$\mathbb{E}\Big[\|S_j^{-1/2}((S_j^{1/2}(A_1\Sigma_1A_1^\top)S_j^{1/2})^{1/2} - (S_j^{1/2}(A_2\Sigma_2A_2^\top)S_j^{1/2})^{1/2})S_j^{-1/2}x\|_2^2\Big],$$

where $S_j = A_j\Sigma_jA_j$ and the expectation is with respect to σ_j. Because $S_j^{-1/2}x = (A_j\Sigma_jA_j)^{-1/2}x \sim \mathcal{N}(0,I)$ and exercise 6.3.1 of [89], we find that the expectation above is equal to

$$\|S_j^{-1/2}((S_j^{1/2}(A_1\Sigma_1A_1^\top)S_j^{1/2})^{1/2} - (S_j^{1/2}(A_2\Sigma_2A_2^\top)S_j^{1/2})^{1/2})\|_F^2.$$

Using the Courant-Fischer min-max theorem as explained in [53] and our bounds on the eigenvalues of $h \in \mathcal{H}$, we can see that

$$\frac{1}{m}\Sigma_j^{-1/2} \succeq (A_j\Sigma_jA_j)^{-1/2} \succeq \frac{1}{M}\Sigma_j^{-1/2}.$$

Since $M^2\Sigma_i \succeq (A_i\Sigma_iA_i) \succeq \varepsilon^2\Sigma_i$, we have

$$\begin{aligned}
&((A_j\Sigma_jA_j)^{1/2}\underbrace{(A_1\Sigma_1A_1^\top)}_{\succeq m^2\Sigma_1}(A_j\Sigma_jA_j)^{1/2})^{1/2} - ((A_j\Sigma_jA_j)^{1/2}\underbrace{(A_2\Sigma_2A_2^\top)}_{\preceq m^2\Sigma_2}(A_j\Sigma_jA_j)^{1/2})^{1/2} \\
&\succeq (m^4\Sigma_j^{1/2}\Sigma_1\Sigma_j^{1/2})^{1/2} - (m^4\Sigma_j^{1/2}\Sigma_2\Sigma_j^{1/2})^{1/2} = m^2\Big[(\Sigma_j^{1/2}\Sigma_1\Sigma_j^{1/2})^{1/2} - (\Sigma_j^{1/2}\Sigma_2\Sigma_j^{1/2})^{1/2}\Big],
\end{aligned}$$

and similarly,

$$\begin{aligned}
&((A_j\Sigma_jA_j)^{1/2}\underbrace{(A_1\Sigma_1A_1^\top)}_{\preceq M^2\Sigma_1}(A_j\Sigma_jA_j)^{1/2})^{1/2} - ((A_j\Sigma_jA_j)^{1/2}\underbrace{(A_2\Sigma_2A_2^\top)}_{\preceq M^2\Sigma_2}(A_j\Sigma_jA_j)^{1/2})^{1/2} \\
&\succeq (M^4\Sigma_j^{1/2}\Sigma_1\Sigma_j^{1/2})^{1/2} - (M^4\Sigma_j^{1/2}\Sigma_2\Sigma_j^{1/2})^{1/2} = M^2\Big[(\Sigma_j^{1/2}\Sigma_1\Sigma_j^{1/2})^{1/2} - (\Sigma_j^{1/2}\Sigma_2\Sigma_j^{1/2})^{1/2}\Big].
\end{aligned}$$

This means that $\|T_{\sigma_j}^{h_\sharp\mu_1} - T_{\sigma_j}^{\tilde{h}_\sharp\mu_2}\|_{\sigma_j}^2$ *has the following bounds*

$$\left\|\frac{M^2}{m}\Sigma_j^{-1/2}\left[(\Sigma_j^{1/2}\Sigma_1\Sigma_j^{1/2})^{1/2} - (\Sigma_j^{1/2}\Sigma_2\Sigma_j^{1/2})^{1/2}\right]\right\|_F^2 \geq \|T_{\sigma_j}^{h_\sharp\mu_1} - T_{\sigma_j}^{\tilde{h}_\sharp\mu_2}\|_{\sigma_j}^2$$
$$\|T_{\sigma_j}^{h_\sharp\mu_1} - T_{\sigma_j}^{\tilde{h}_\sharp\mu_2}\|_{\sigma_j}^2 \geq \left\|\frac{m^2}{M}\Sigma_j^{-1/2}\left[(\Sigma_j^{1/2}\Sigma_1\Sigma_j^{1/2})^{1/2} - (\Sigma_j^{1/2}\Sigma_2\Sigma_j^{1/2})^{1/2}\right]\right\|_F^2.$$

Moreover, notice that since $\Sigma_j = \Sigma_1$ *or* $\Sigma_j = \Sigma_2$*, we can assume without loss of generality that* $\Sigma_j = \Sigma_1$ *so that*

$$\Sigma_j^{-1/2}\left[(\Sigma_j^{1/2}\Sigma_1\Sigma_j^{1/2})^{1/2} - (\Sigma_j^{1/2}\Sigma_2\Sigma_j^{1/2})^{1/2}\right] = \Sigma_1^{1/2} - \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}.$$

We can show, however, that the Frobenius norm of the right-hand side is actually $W_2(\mu_1,\mu_2)$*. To see this, first notice that because* $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$ *is symmetric, using the cyclic property of traces, we have*

$$\begin{aligned}\|\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\|_F^2 &= \mathrm{Tr}((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})\\ &= \mathrm{Tr}(\Sigma_1^{-1}\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}) = \mathrm{Tr}(\Sigma_2).\end{aligned}$$

Applying this result, we have that $\|\Sigma_1^{1/2} - \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\|_F^2$ *is equal to*

$$\begin{aligned}&\|\Sigma_1^{1/2}\|_F^2 + \underbrace{\|\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\|_F^2}_{\mathrm{Tr}(\Sigma_2)} - 2\,\mathrm{Tr}(\Sigma_1^{1/2}\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})\\ &\qquad = \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})) = W_2(\mu_1,\mu_2)^2.\end{aligned}$$

So we get that

$$\frac{M^2}{m}W_2(\mu_1,\mu_2) \geq \|T_{\sigma_j}^{h_\sharp\mu_1} - T_{\sigma_j}^{\tilde{h}_\sharp\mu_2}\|_{\sigma_j} \geq \frac{m^2}{M}W_2(\mu_1,\mu_2)$$

for our choices of reference distributions, of which there are infinite choices because our choices of ε *and M are constrained by*

$$1 - \frac{W_2(\mu_1,\mu_2)}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F} \leq m \leq M \leq 1 + \frac{W_2(\mu_1,\mu_2)}{2\max_{j=1,2}\|\Sigma_j^{1/2}\|_F}.$$

2.10 The shearing transformation

Algorithm 1 Procedure to produce shears of an image

```
Inputs : $28 \times 28$ matrix of pixel values corresponding to the image, matrix $A$ and shift $b$.
Output : A $28 \times 28$ matrix of pixel values corresponding to the transformation $A(x - center) + b$. (Center of the image is assumed to be $(14,14)$). Here $x = (i,j)$ where $i, j \in \{1, 2, \cdots 28\}$.
ShearedImage $\leftarrow$ An empty $28 \times 28$ array
for $i = 1, \cdots 28$ do
    for $j = 1, \cdots 28$ do
        $y \leftarrow (i,j) - center$
        $x \leftarrow A^{-1}(y - b) + center$
        if $x_1 > 28$ or $x_1 <= 0$ or $x_2 > 28$ or $x_2 <= 0$ then
            ShearedImage(i,j) $\leftarrow 0$
        else
            ShearedImage(i,j) $\leftarrow$ Interpolation of the pixel values (of the original image) of the four grid points corresponding to the grid box which $x$ belongs to.
        end if
    end for
end for
return ShearedImage
```

Following notation introduced in Section 4 of the main text, the function class $\mathcal{H}$ with respect to which we perform numerical experiments on MNIST images to study linear separability is,

$$\mathcal{H} = \left\{Ax + b : A \text{ is symmetric positive definite, } b \in \mathbb{R}^2\right\}, \tag{2.9}$$

Specifically we choose A to be,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}^T \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{2.10}$$

where, $\lambda_1, \lambda_2 > 0$ so that A is positive definite. In the subsequent sections, we present the classification results for two different choices for the range of parameter $(\lambda_1, \lambda_2, \theta, b)$ values, one representing a mild shearing of the images and the other representing a severe shearing of the images.

2.11 Standard deviation in test error of MNIST classification experiments

2.12 Numerical validation of example 2.12

To illustrate Theorem 2.8, we had provided a simple example with Gaussians (see example 2.12 of main text). Let $\mu = \mathcal{N}(m_1, I_n)$. Consider a symmetric positive definite matrix A with spectral decomposition $A = P^\top \Lambda P$ and a corresponding fixed shear $S(x) = Ax + b$ for some $b \in \mathbb{R}^n$, which yields the pushforward $S_\sharp\mu = \mathcal{N}(Am_1 + b, AA^\top)$. For simplicity, we will check that the subset of compatible affine transformations

$$\begin{aligned} \mathcal{F}_{\text{affine}}(P) &= \{f(x) = Cx + d : f \in \mathcal{F}(P)\} \\ &= \{P^\top DPx + d : D_{ij} = 0 \,\forall\, i \neq j, D_{ii} > 0, d \in \mathbb{R}^n\} \end{aligned} \tag{2.11}$$

yields reference distributions $\sigma \in \{f_\sharp\mu : f \in \mathcal{F}_{\text{affine}}(P)\}$ so that the compatibility condition hold. In particular note that for $f(x) = Cx + d = P^\top DPx + d$, our reference distributions have the

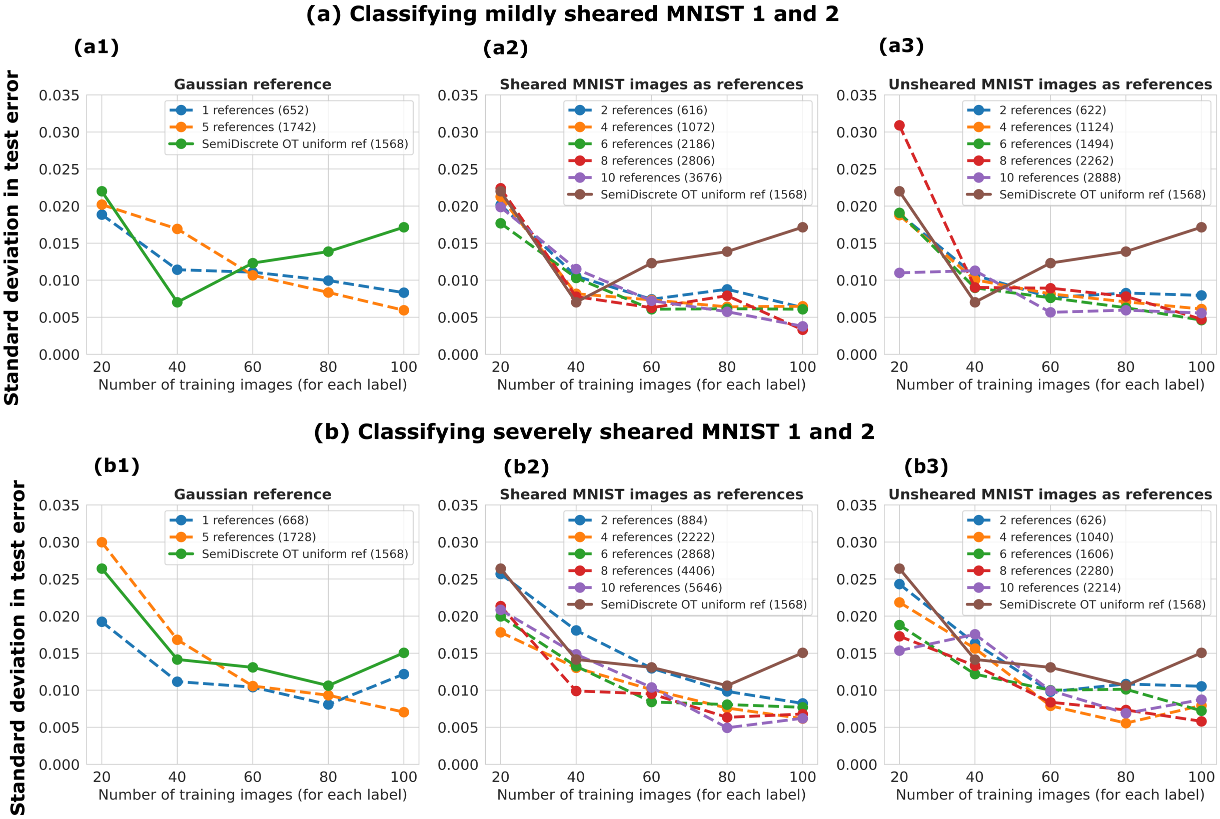


Figure 2.8: (a) Standard deviation in test errors for binary classification of mildly sheared MNIST 1s and 2s using (a1) Gaussian references (a2) sheared MNIST 1s and 2s as references (a3) unsheared MNIST 1s and 2s as references. (b) Standard deviation in test errors for binary classification of severely sheared MNIST 1s and 2s using (b1) Gaussian references (b2) sheared MNIST 1s and 2s as references (b3) unsheared MNIST 1s and 2s as references. In the cases where MNIST images are used as references, the results are reported for the cases where the number of references used is $2i$ for $i = 1, \cdots 5$ wherein i images from each class are randomly drawn to be used as references from a pool of images that do not correspond to any of the training and testing images. For each fixed number of training images per class, N_{train}, the mean test classification error averaged across 20 random choices of N_{train} training images (per class) and 1000 test images (per class) is reported. The number inside the parenthesis in the legends of the images denote the length of the LOT feature vector corresponding to the particular choice of references. In all figures, for comparison, the results for classification using the semi discrete linear optimal transport framework [68] which uses the uniform measure as the reference is also reported.

form

$$\sigma = \mathcal{N}(Cm_1 + d, CC^\top) = \mathcal{N}(Cm_1 + d, P^\top D^2 P).$$

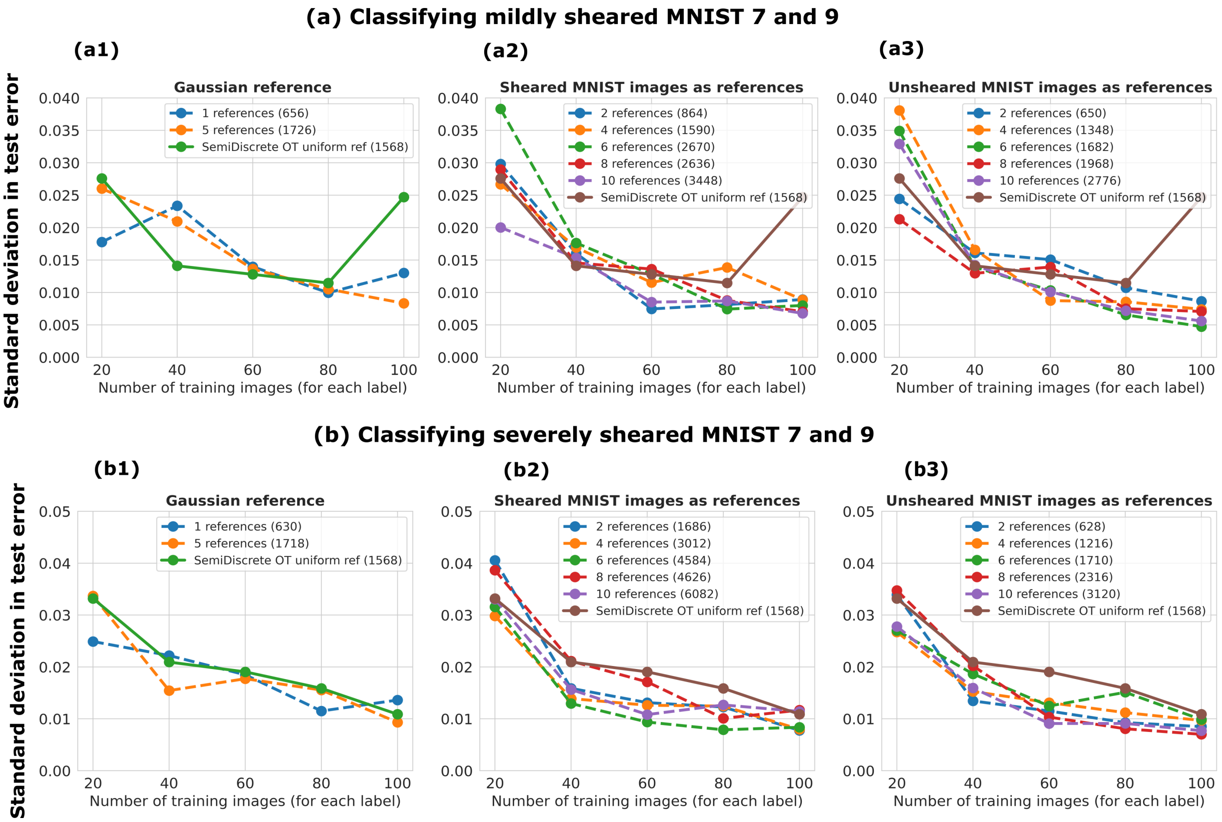


Figure 2.9: (a) Standard deviation in test errors for binary classification of mildly sheared MNIST 7s and 9s using (a1) Gaussian references (a2) sheared MNIST 7s and 9s as references (a3) unsheared MNIST 7s and 9s as references. (b) Standard deviation in test errors for binary classification of severely sheared MNIST 7s and 9s using (b1) Gaussian references (b2) sheared MNIST 7s and 9s as references (b3) unsheared MNIST 7s and 9s as references. In the cases where MNIST images are used as references, the results are reported for the cases where the number of references used is $2i$ for $i = 1, \cdots 5$ wherein i images from each class are randomly drawn to be used as references from a pool of images that do not correspond to any of the training and testing images. For each fixed number of training images per class, N_{train}, the mean test classification error averaged across 20 random choices of N_{train} training images (per class) and 1000 test images (per class) is reported. The number inside the parenthesis in the legends of the images denote the length of the LOT feature vector corresponding to the particular choice of references. In all figures, for comparison, the results for classification using the semi discrete linear optimal transport framework [68] which uses the uniform measure as the reference is also reported.

2.13 Comparison with Convolutional Neural Networks (CNNs)

2.14 Experimental Comparison of the Wasserstein distance and the LOT embedding distance

For the sheared MNIST images whose separability was studied in Section 6 of the main text, Figure 2.18 shows the comparison between the Wasserstein distance, $W_2(\mu_1, \mu_2)$, and the LOT embedding distance $W_{2,\sigma}^{\mathrm{LOT}}(\mu_1, \mu_2)$. The Wasserstein distance is approximated using the Python Optimal Transport (POT) package [43] [5] and LOT embedding distance is approximated as the l^2 distance of the difference between the *LOT feature* vectors normalized by the grid size.

2.15 Acknowledgements

This chapter, in full, is a reprint of *Supervised learning of sheared distributions using linearized optimal transport* as it appears in Sampl. Theory Signal Process. Data Anal. 21, 1 (2023). https://doi.org/10.1007/s43670-022-00038-2 by Khurana, V., Kannan, H., Cloninger, A. et al. The dissertation author was the primary investigator and author of this paper.

[5] https://pythonot.github.io/

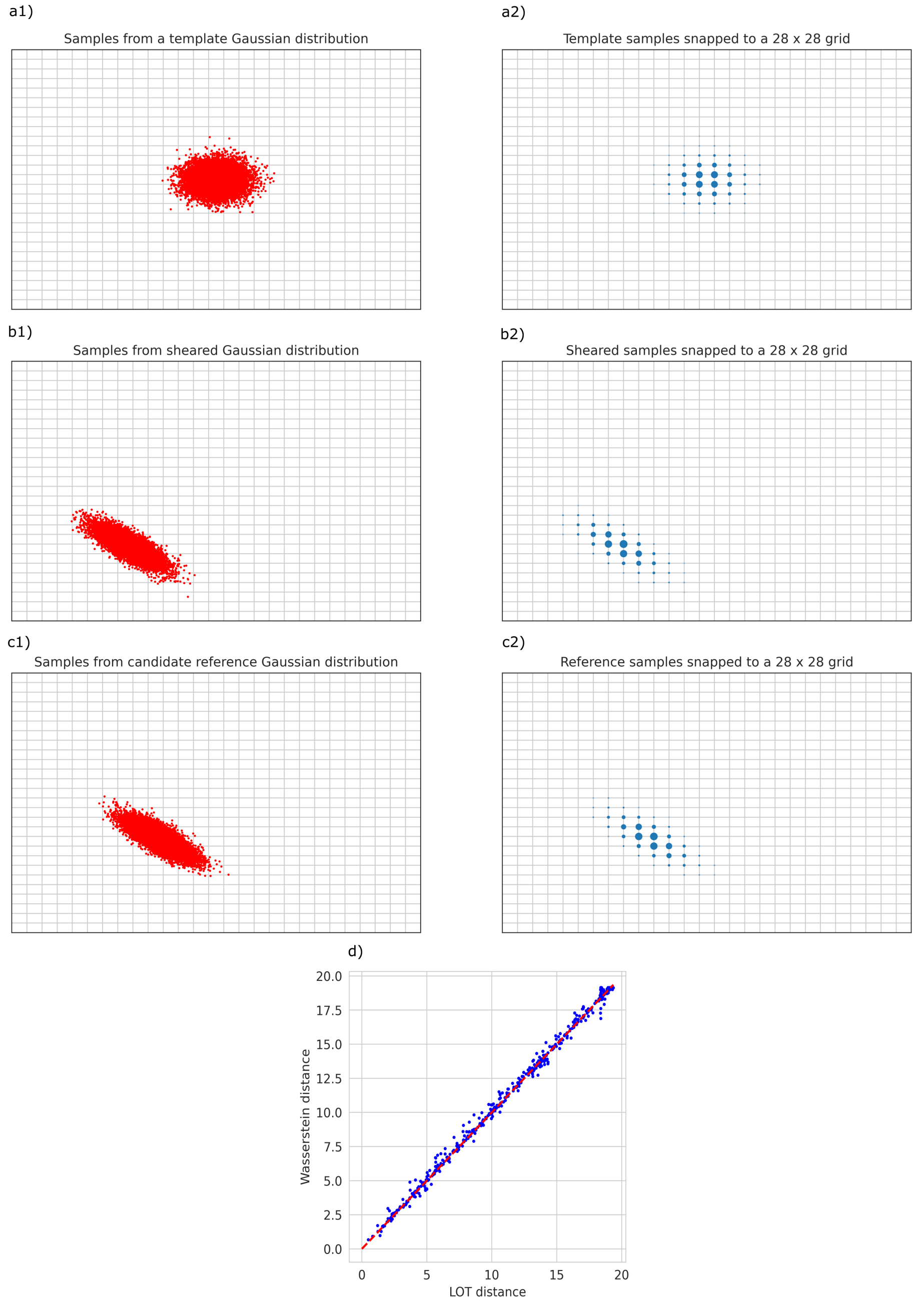


Figure 2.10: a1) Samples from a Gaussian distribution that serves as the template μ. a2) Approximation of the template distribution as a discrete distribution on a grid. b1) Samples from sheared distribution $S_\sharp\mu$. b2) Approximation of the sheared distribution as a discrete distribution on a grid. c1) Samples from a candidate referencec distribution $f_\sharp\mu \in \mathcal{F}_{\text{affine}}(P)$ (equation 2.11). c2) Approximation of the reference distribution as a discrete distribution on a grid. d) Numerical validation of the equivalence of LOT distance $W^{LOT}_{2,f_\sharp\mu}(\mu, S_\sharp\mu)$ and the Wasserstein distance $W_2(\mu, S_\sharp\mu)$ under compatibility as in example 2.12 using the LOT framework for different choices of shear S.

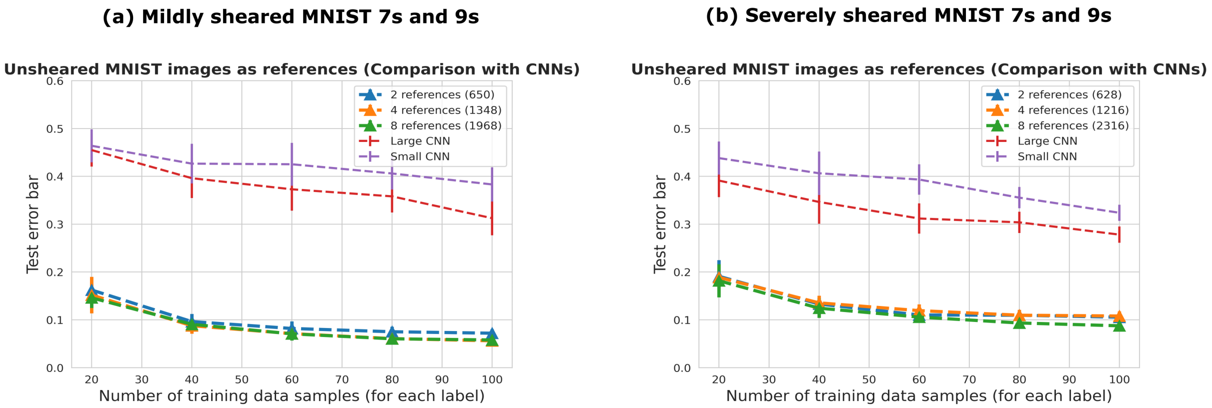


Figure 2.11: Comparison of discrete LOT classification of (a) mildly sheared MNIST 7s and 9s (b) severely sheared MNIST 7s and 9s with convolutional neural network with 1586 training parameters (labelled small CNN) and 3650 training parameters (labelled large CNN) under identical training and testing conditions.

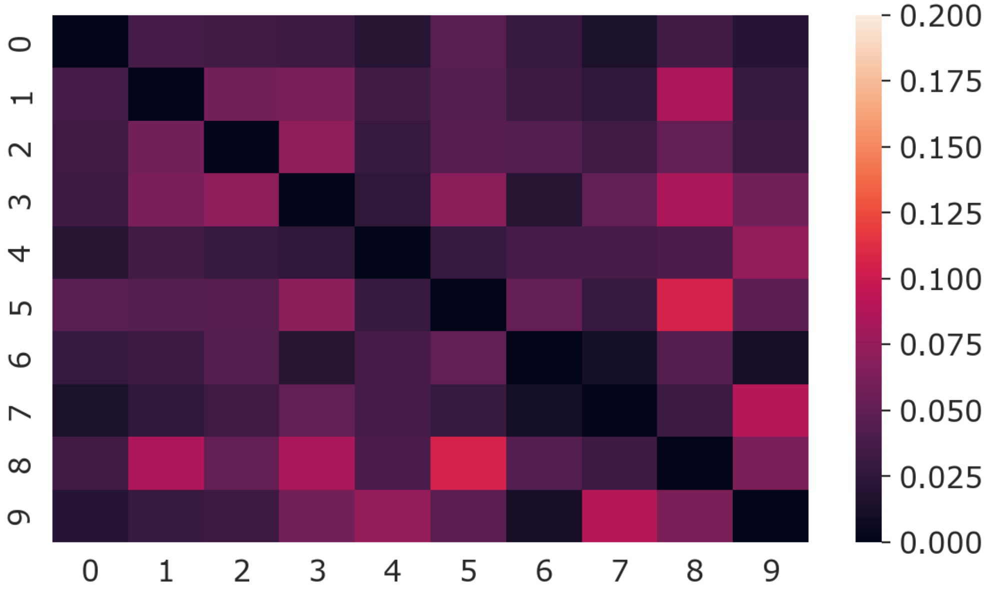


Figure 2.12: Mean test errors for pairwise binary classification of MNIST digits $0-9$ using semi-discrete LOT classifier under mild shearing conditions. For each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 1.5]$, θ is randomly chosen in the interval $[0, 360]$ and the shifts b are randomly chosen in the interval $[-5, 5]$. The number of training data samples used per digit class is 40 and the mean value of the test error is reported based on 20 sample experiments.. The standard deviation of the test errors was < 0.055.

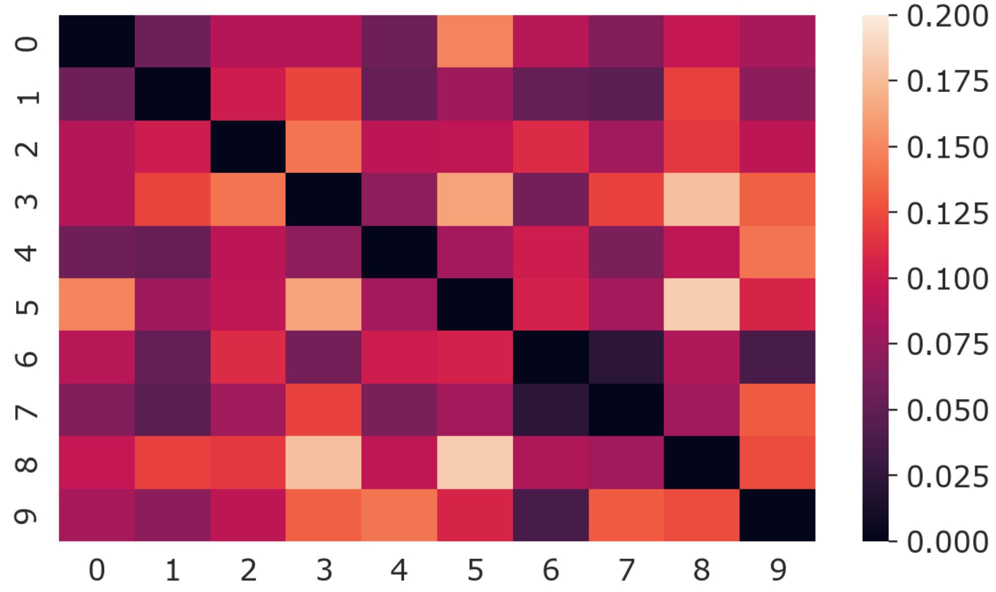


Figure 2.13: Mean test errors for pairwise binary classification of MNIST digits $0-9$ using semi-discrete LOT classifier under severe shearing conditions. For each image λ_1, λ_2 are randomly chosen in the interval $[0.5, 2.5]$, θ is randomly chosen in the interval $[0, 360]$ and the shifts b are randomly chosen in the interval $[-5, 5]$. The number of training data samples used per digit class is 40 and the mean value of the test error is reported based on 20 sample experiments.. The standard deviation of the test errors was < 0.055.

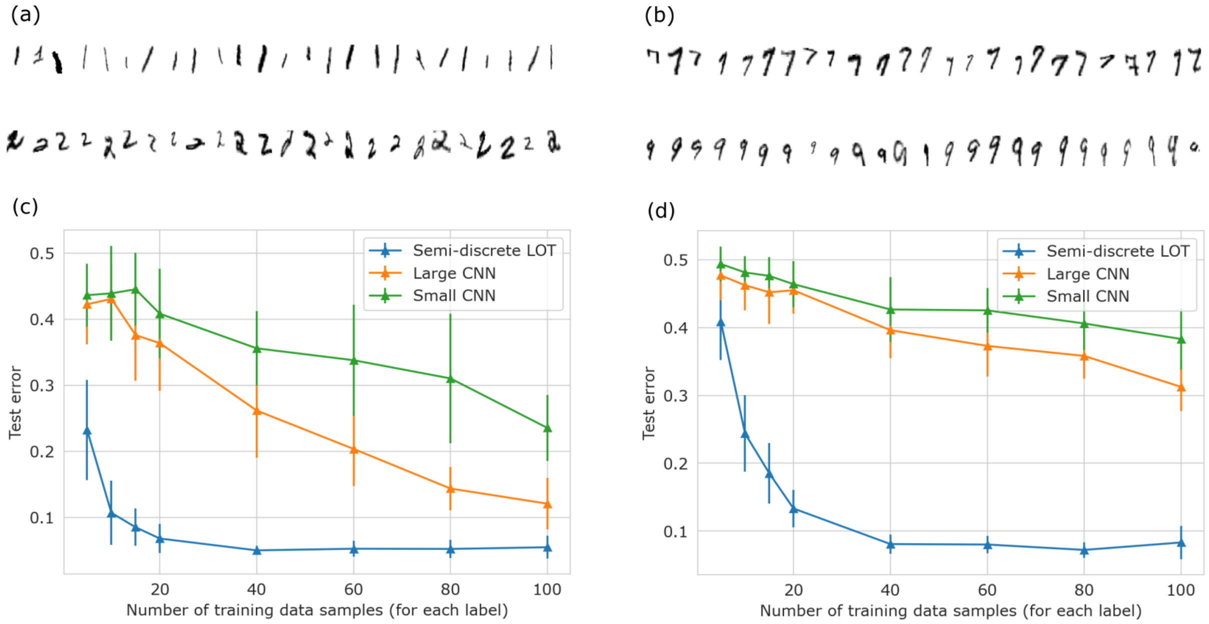


Figure 2.14: a) A randomly selected subset of mildly sheared MNIST ones and twos. b) A randomly selected subset of mildly sheared MNIST sevens and nines. c) Classifcation of mildly MNIST sheared ones and twos. For each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 1.5]$, θ is randomly chosen in the interval $[0, 360]$ and the shifts b are randomly chosen in the interval $[-5, 5]$. d) Classifcation of mildly sheared MNIST sevens and nines. For each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 1.5]$, θ is randomly chosen in the interval $[0, 360]$ and the shifts b are randomly chosen in the interval $[-5, 5]$.

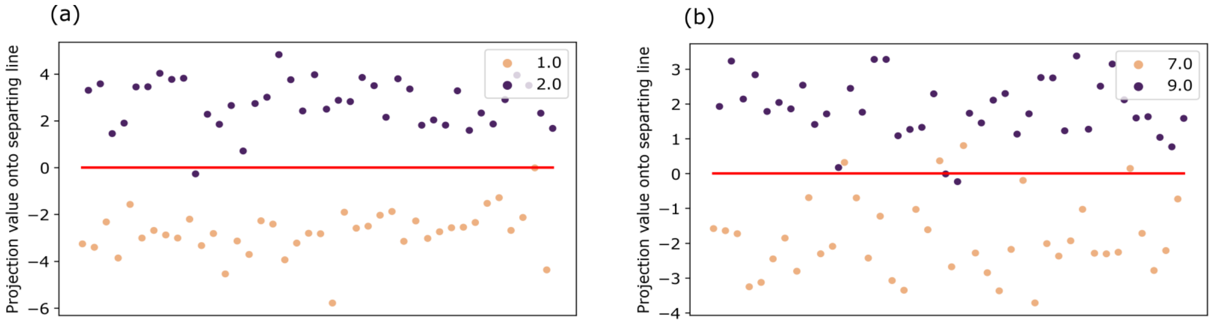


Figure 2.15: Visualization of separation in the LDA projection of 40 training per class corresponding to a) MNIST ones and twos under mild shearing b) MNIST sevens and nines under mild shearing.

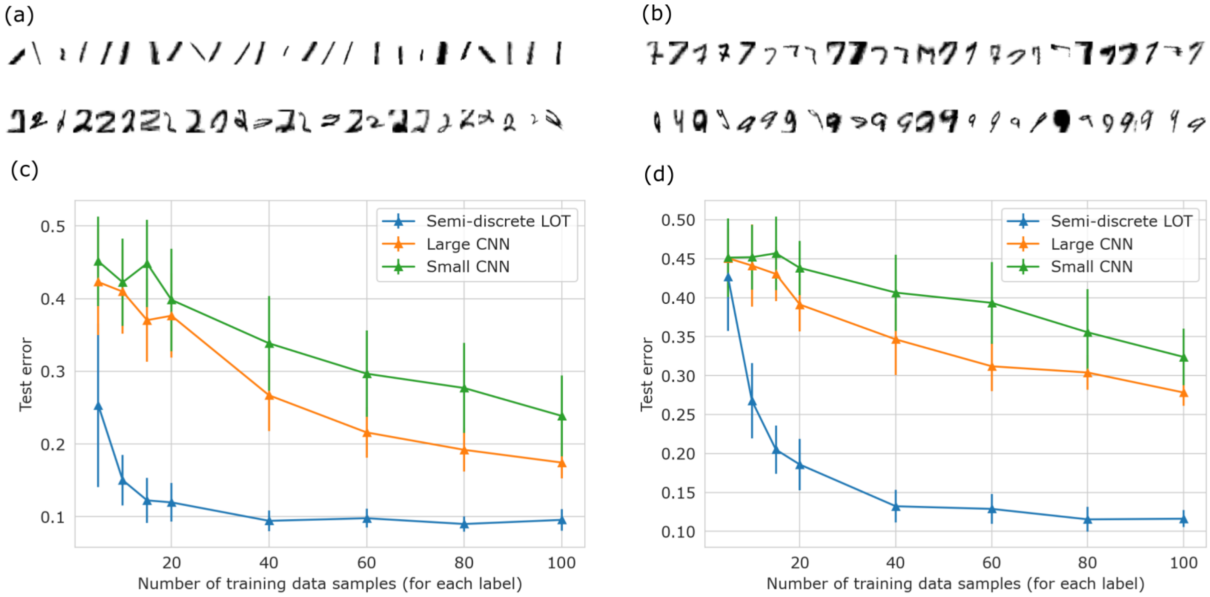


Figure 2.16: a) A randomly selected subset of severely sheared MNIST ones and twos. b) A randomly selected subset of severely sheared MNIST sevens and nines. c) Classifcation of severely sheared MNIST ones and twos. For each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 2.5]$, θ is randomly chosen in the interval $[0, 360]$ and the shifts b are randomly chosen in the interval $[-5, 5]$. d) Classifcation of severely sheared MNIST sevens and nines. For each image, λ_1, λ_2 are randomly chosen in the interval $[0.5, 1.5]$, θ is randomly chosen in the interval $[0, 360]$ and the shifts b are randomly chosen in the interval $[-5, 5]$.

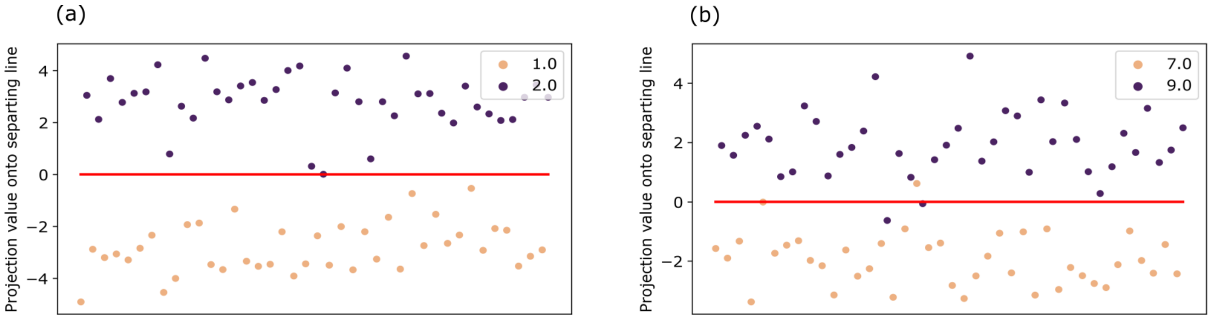


Figure 2.17: Visualization of separation in the LDA projection of 40 training samples per class corresponding to a) MNIST ones and twos under severe shearing b) MNIST sevens and nines under severe shearing.

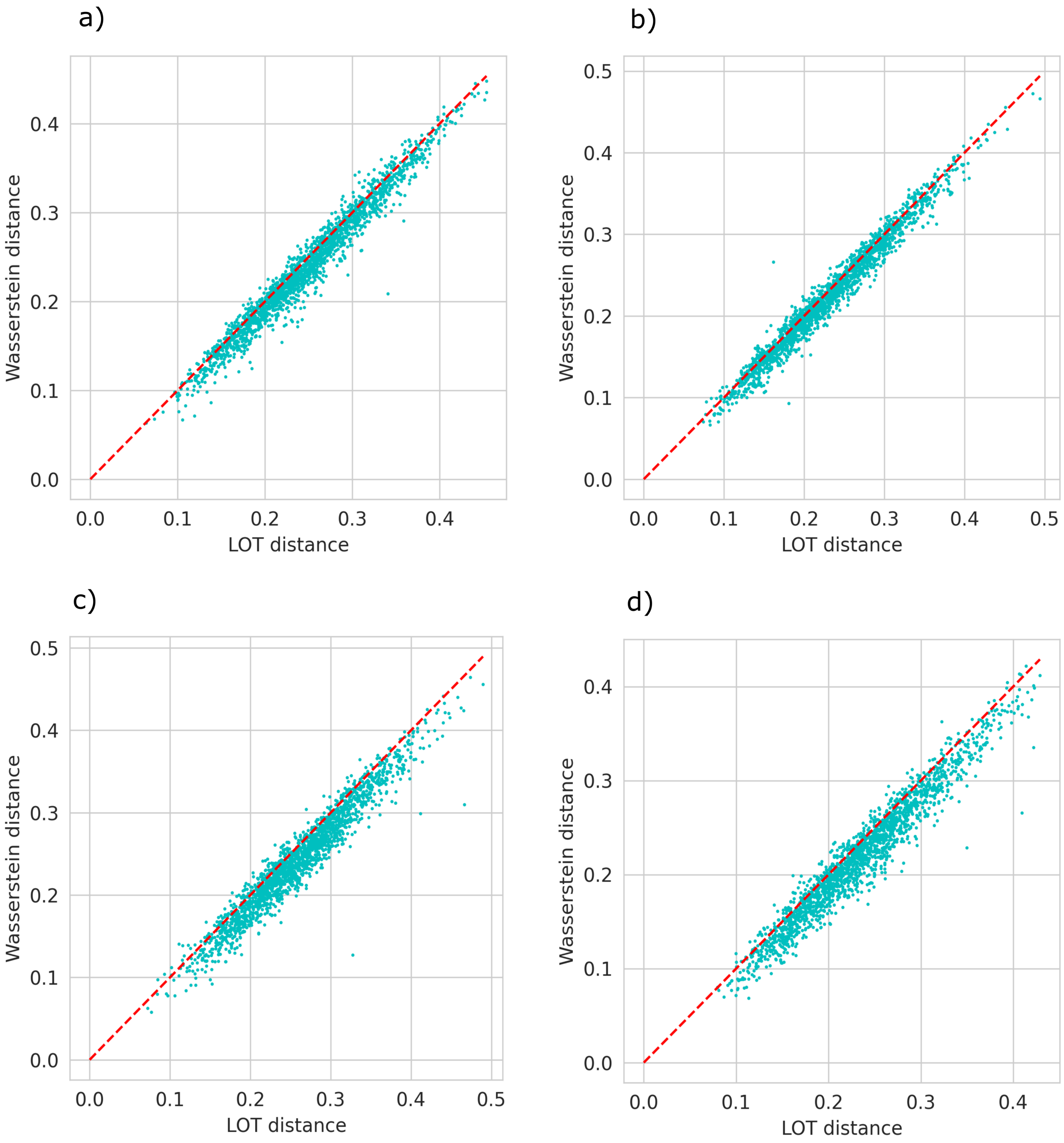


Figure 2.18: a) 2000 pairwise distances between mildly sheared MNIST ones and twos. b) 2000 pairwise distances between mildly sheared sevens and nines. c) 2000 pairwise distances between severely sheared MNIST ones and twos. d) 2000 pairwise distances between severely sheared sevens and nines.

Chapter 3

Linearized Wasserstein Embeddings

JOINT WORK WITH KEATON HAMM, CAROLINE MOOSMÜLLER, AND ALEX CLONINGER

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.

3.1 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 [13, 34, 63, 88]. 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 [82], text documents (the bag-of-word model uses word count within a text as features, creating a histogram for each document [100]), and gene expression data, which can be interpreted as a distribution over a gene network [27, 65]. 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* [90].

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

$$W_2(\mu,\nu) := \inf_{\pi\in\Gamma(\mu,\nu)} \left(\int_{\mathbb{R}^{2n}} \|x-y\|^2 d\pi(x,y) \right)^{\frac{1}{2}}, \tag{3.1}$$

where $\mathcal{P}(\mathbb{R}^{2n})$ is the set of all probability measures over $\mathbb{R}^{2n}$ and $\Gamma(\mu,\nu) := \{\gamma \in \mathcal{P}(\mathbb{R}^{2n}) : \gamma(A\times\mathbb{R}^n) = \mu(A),\ \gamma(\mathbb{R}^n \times A) = \nu(A) \text{ for all } A \subset \mathbb{R}^n\}$ is the set of all joint probability measures with marginals μ and ν. Under regularity assumptions on μ, the optimal coupling π has the form

$\pi = (\mathrm{id}, T)_\sharp \mu$, where $T \in \mathrm{L}^2(\mathbb{R}^n, \mu)$ is the "optimal transport map" [18, 90].

The Wasserstein space and optimal transport have gained popularity in the machine learning community, as they are based on a solid theoretical foundation [90] (for example, (3.1) is a metric), while providing a versatile framework for applications (for example, as a cost function for generative models [10], in semi-supervised learning [85], and in pattern detection for neuronal data [70]).

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 [48, 92], which introduces the *Wassmap* algorithm (see Section 3.2.6 for more details), a version of the Multidimensional Scaling algorithm (MDS) [64] (see Algorithm 2), or more generally, the Isomap algorithm [88].

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 (3.1). 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) [3, 36], which deal with the computational issue involving n, and in *linearized optimal transport* (LOT) [45, 92], 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) [64] (see Algorithm 2) applied to a dataset in the Wasserstein space, where the pairwise Wasserstein distances are only available up to an error τ.

- We study the size of τ 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 [48]. 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.

3.1.1 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 [99] to study shape spaces of proteins, in [65, 27] to analyze gene expression data, and in [92] 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 [48]. 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 [74, 96]. The authors propose to represent complex data in the Wasserstein space as Wasserstein barycenters of a dictionary.

3.1.2 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 3.1 (Informal version of Theorem 3.9)**.** *Assume that data points $\{\mu_i\}_{i=1}^N$ are τ_1-close to a d-dimensional submanifold $\mathcal{W}$ in the Wasserstein space, which is isometric to a subset Ω of Euclidean space $\mathbb{R}^d$. Furthermore assume that we only have access to approximations λ_{ij} of the pairwise distances $W_2(\mu_i,\mu_j)$, and that the approximation error is τ_2.*

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

Some remarks on this result:

- The first source of error, τ_1, depends on how close the data points are to the submanifold $\mathcal{W}$ isometric to a subspace of $\mathbb{R}^d$, which is completely determined by the dataset.

- The second source of error, τ_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 τ_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 (3.1), which depends on a parameter β, and leads to *Sinkhorn distances* [36]:

$$\min_{\pi\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^{2n}} \frac{1}{2}\|x-y\|^2 d\pi(x,y) + \beta D_{\mathrm{KL}}(\pi\|\mu\otimes\nu), \tag{3.2}$$

where D_{KL} is the Kullback–Leibler divergence of measures [54]. This formulation leads to a unique solution (in contrast to (3.1)), and to a significant computational speed-up in n, achieving $O(n^2 \log(n))$ through matrix scaling algorithms (Sinkhorn's algorithm) [3, 36].

(b) *Linearized Wasserstein distances*: Linearized optimal transport (LOT) [45, 92] approximates Wasserstein distances by linear L^2-distances in the tangent space at a chosen reference measure σ:

$$W_{2,\sigma}^{\mathrm{LOT}}(\mu,\nu) := \left(\int_{\mathbb{R}^n} \|T_\sigma^\mu(x) - T_\sigma^\nu(x)\|^2 d\sigma(x)\right)^{1/2}, \tag{3.3}$$

where T_σ^μ denotes the optimal transport map from σ to μ (either computed through (3.1) or (3.2), 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 σ to μ_i, and approximates pairwise maps between μ_i and μ_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 [78, 94], such as visualizing phenotypic differences between types of cells [12]. There furthermore exist error bounds for $W_{2,\sigma}^{\mathrm{LOT}}$ [15, 39, 45, 57, 68, 72].

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

$$\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu}) := \left(\frac{1}{m}\sum_{j=1}^{m} \|T_\sigma^{\widehat{\mu}}(X_j) - T_\sigma^{\widehat{\nu}}(X_j)\|^2\right)^{1/2}, \tag{3.4}$$

where $X_j \sim \sigma$ i.i.d. and the transport maps are either computed by (3.1) or (3.2) (and with barycentric projections, if necessary).

We provide values for τ_2 as in Theorem 3.1, by bounding $|W_2(\mu,\nu)^2 - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu})^2|$, 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 $\|T_\mu^{\widehat{\nu}} - T_\mu^\nu\|_\mu$, which are provided by [38] for the linear program case, and by [80] in the regularized case. Both [38] and [80] assume compactly supported μ and ν, 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 $W_{2,\sigma}^{\text{LOT}}$, which are provided in [57, 72], and are based on the idea that μ_i are generated by almost compatible functions $\mathcal{H}$ applied to a fixed generator μ. We also strengthen some of the approximation results in [57, 72].

3.1.3 Efficient algorithm: LOT Wassmap

The Wassmap algorithm of [48] 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 3), which uses LOT distances (3.3) 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 (3.4)). 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 3.1:

Corollary 3.2 (Informal version of Corollary 3.10)**.** *Assume that data points $\{\mu_i\}_{i=1}^N$ are τ_1-close to a d-dimensional submanifold $\mathcal{W}$ in the Wasserstein space, which is isometric to a subset Ω of Euclidean space $\mathbb{R}^d$. Choose a reference measure σ and compute all transport maps $T_\sigma^{\mu_i}$ (either with a linear program* (3.1) *or with Sinkhorn approximations* (3.2)*, and with barycentric projections, if necessary). Let τ_2 be the error between the empirical linearized Wasserstein-2 distance $\widehat{W}_{2,\sigma}^{\text{LOT}}(\widehat{\mu}_i,\widehat{\mu}_j)$ of* (3.4) *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 $T_\sigma^{\mu_i}$ (column-stacked) produces data points $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$, which are $C_{N,\mathcal{W}}(\tau_1+\tau_2)$-close to Ω up to rigid transformations.

We note that Corollary 3.2 is a corollary of Theorem 3.1 and that the technical assumptions and constants are the same in both results.

In Section 5.8, 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.

3.1.4 Organization of the paper

This paper is organized as follows: We start by introducing important notation and background in Section 3.2. This includes discussion of the MDS and Wassmap algorithms, (linearized) optimal transport, and plug-in estimators. Section 3.3 introduces the LOT Wassmap algorithm and provides the main results. Sections 3.4 and 3.5 provide approximation guarantees for $\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu})$ for compactly and non-compactly supported target measures, respectively. The approximation guarantees come with many technical assumptions, and Sections 3.6 and 3.7 are dedicated to discussing settings in which these assumptions hold. The paper concludes with experiments in Section 5.8, which show the effectiveness of LOT Wassmap. Proofs are provided in Sections 3.9 to 3.12.

3.2 Notation and Background

This paper has a significant amount of background and notation which is summarized categorically here. See Table 3.1 for an overview of notation used in the paper.

3.2.1 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

Table 3.1: Overview of notation used in the paper.

Notation	Definition	Reference
Δ	Square Euclidean distance matrix	Algorithm 2
Λ	Perturbed distance matrix	Corollary 3.8
$X^\dagger$	Moore–Penrose pseudoinverse of matrix X	Section 3.2.1
μ	Template measure	Section 3.2.4
$\widehat{\mu}$	Empirical measure approximating μ	(3.7)
σ	Reference measure for LOT	Section 3.2.4
$\\|\cdot\\|_{\mathrm{S}_p}$	Schatten p-norm	Section 3.2.1
$\\|\cdot\\|$	Spectral norm of a matrix or Euclidean norm of a vector	Section 3.2.1
$\\|\cdot\\|_{\mathrm{F}}$	Frobenius norm of a matrix	Section 3.2.1
$\\|\cdot\\|_{\max}$	(Entrywise) maximum norm of a matrix	Section 3.2.1
$\\|\cdot\\|_{\mu}$	Norm on $\mathrm{L}^2(\mathbb{R}^n,\mu)$	Section 3.2.3
n	Dimension of Euclidean space that probability measures are defined on	Section 3.2.3
$\mathcal{P}(\mathbb{R}^n)$	Probability measures on $\mathbb{R}^n$	Section 3.2.3
$\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^n)$	Absolutely continuous probability measures on $\mathbb{R}^n$	Section 3.2.3
$W_2(\mathbb{R}^n)$	Wasserstein-2 space over $\mathbb{R}^n$	Section 3.2.3
$W_2(\mu,\nu)$	Wasserstein-2 distance between μ and ν	(3.5)
$W_{2,\sigma}^{\mathrm{LOT}}(\mu,\nu)$	Linearized Wasserstein-2 distance between μ and ν, with σ as reference	(3.6)
$\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\mu,\nu)$	Empirical linearized Wasserstein-2 distance	(3.12)
T_σ^μ	Optimal transport (Monge) map from σ to μ	Section 3.2.3
$T_\sharp\mu$	Pushforward of μ with respect to T	Section 3.2.3
$T_\sigma^{\widehat{\mu}}$	Barycentric projection of an optimal transport plan (Kantorovich potential)	(3.10)
d	Embedding dimension of MDS	Section 3.2.2
k	Sample size that generates $\widehat{\mu}$	(3.7)
m	Sample size that generates $\widehat{\sigma}$	Algorithm 3
N	Number of data points	Algorithm 3
ε	Distance from compatibility	Definition 3.4
β	Regularizer for Sinkhorn OT	Section 3.4.2

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 $\Sigma_d = \mathrm{diag}(\sigma_1,\ldots,\sigma_d) \in \mathbb{R}^{d\times d}$. 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 $\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_{\min\{m,n\}}}$ 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 ℓ^p norms of the vector of singular values: $\|A\|_{S_p} := \|(\sigma_1, \ldots, \sigma_{\min\{m,n\}})\|_{\ell_p}$. The Frobenius norm, which is the Schatten 2-norm is denoted by $\|\cdot\|_{\mathrm{F}}$, and the spectral norm, which is the Schatten ∞-norm is denoted simply by $\|\cdot\|$. We also use $\|\cdot\|$ to denote the Euclidean norm of a vector.

3.2.2 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 2. For more details see [64].

Algorithm 2 Multidimensional Scaling (MDS) [64]

```
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 [98]. 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.

3.2.3 Optimal Transport Preliminaries

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

$$W_2(\mu,\nu) := \inf_{\pi \in \Gamma(\mu,\nu)} \left(\int_{\mathbb{R}^{2n}} \|x-y\|^2 d\pi(x,y) \right)^{\frac{1}{2}}, \tag{3.5}$$

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

In [18], Brenier showed that if μ is absolutely continuous with respect to the Lebesgue measure, the optimal coupling of (3.5) 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_μ^ν exists, then

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

Furthermore, [18] shows that when μ is absolutely continuous with respect to the Lebesgue measure, the map T_μ^ν is uniquely defined as the gradient of a convex function ϕ, i.e. $T_\mu^\nu = \nabla\phi$ (up to an additive constant).

3.2.4 Linearized optimal transport

Linearized optimal transport (LOT) [45, 68, 78, 94] defines an embedding of $\mathcal{P}(\mathbb{R}^n)$ into the linear space $\mathrm{L}^2(\mathbb{R}^n, \sigma)$, with σ 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 $\mathcal{P}(\mathbb{R}^n)$, to linearly approximate the Wasserstein distance, or for fast Wasserstein barycenter computations [2, 57, 68, 72, 78].

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

$$W_{2,\sigma}^{\mathrm{LOT}}(\mu, \nu) := \|T_\sigma^\mu - T_\sigma^\nu\|_\sigma. \tag{3.6}$$

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 3.3 (Compatibility condition [2, 72, 78])**.** *Let* $\sigma, \mu \in W_2(\mathbb{R}^n) \cap \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^n)$. *We say that the LOT embedding is compatible with the μ-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 [45].

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 3.29 and [72, 78]. In particular, this is the case when g is either a shift or scaling, or a certain type of shearing [57, 72, 78].

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

Definition 3.4 (ε-compatibility)**.** *Let* $\sigma, \mu \in W_2(\mathbb{R}^n) \cap \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^n)$. *We say that $\mathcal{H}$ is ε- compatible with respect to σ and μ, if for every $h \in \mathcal{H}$, 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 ε-compatible transformations imply “almost”-isometry between $W_{2,\sigma}^{\mathrm{LOT}}$ and W_2. Some of these results are accounted for in [72, Proposition 4.1]; however, we also extend these almost-compatibility results in Theorem 3.30. These results make use of the Hölder regularity bounds for $W_{2,\sigma}^{\mathrm{LOT}}$ of [45, 68]. 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 μ is a fixed *template measure*, and $h \in \mathcal{H}$ with $\mathcal{H}$ a space of functions in $\mathrm{L}^2(\mathbb{R}^n, \mu)$. This is similar to assumptions in [2, 57, 72, 78], where $\mathcal{H}$ consists of shifts and scalings, compatible maps, or has other properties, such as convexity and compactness. We will write $\mu_i \sim \mathcal{H}_\sharp \mu$ to indicated that μ_i is of such a form for all $i = 1, \ldots, N$, and $\mathcal{H}$ will be specified in the respective context. Note that [2] calls this data generation process an “algebraic generative model”.

3.2.5 Optimal transport with plug-in estimators

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

$$\widehat{\mu} = \frac{1}{k} \sum_{i=1}^{k} \delta_{Y_i} \tag{3.7}$$

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 (3.5) has a (possibly non-unique) solution for transporting an absolutely continuous measure σ to an empirical measure of the form (3.7). Following [38], we define the set of Kantorovich plans

$$\Gamma_{\min} := \underset{\pi\in\Gamma(\sigma,\widehat{\mu})}{\operatorname{argmin}} \int_{\mathbb{R}^{2n}} \|x-y\|^2 d\pi(x,y), \tag{3.8}$$

which may contain more than one transport plan. In practice, these optimal transport plans are exactly computed via linear programming to solve (3.8). We call optimal transport plans solved with linear programming γ_{LP}. It is much faster, however, to approximate the optimal transport plan by using an entropic regularized plan [36]. In particular, we get a unique solution by solving

$$\gamma_\beta := \underset{\pi\in\Gamma(\sigma,\widehat{\mu})}{\operatorname{argmin}} \int \frac{1}{2}\|x-y\|^2 d\pi(x,y) + \beta D_{\mathrm{KL}}(\pi\|\sigma\otimes\widehat{\mu}), \tag{3.9}$$

where D_{KL} is the Kullback–Leibler divergence of measures [54], $\sigma\otimes\widehat{\mu}$ is the measure on the product space $\mathbb{R}^n\times\mathbb{R}^n$ whose marginals are σ and $\widehat{\mu}$, and β denotes the regularizer. We solve (3.9) with Sinkhorn's algorithm, which yields entropic potentials f_β and g_β corresponding to σ and $\widehat{\mu}$, respectively.

Regardless of whether we solve the optimal transport plan using (3.8) or (3.9), we can make a transport plan $\gamma\in\Gamma$ into a map by defining the barycentric projection

$$T_\sigma^{\widehat{\mu}}(x;\gamma) := \frac{\int_y y\,d\gamma(x,y)}{\int_y d\gamma(x,y)}, \quad \text{for } x\in\operatorname{supp}(\sigma). \tag{3.10}$$

This leads to a natural way to consider linearized Wasserstein-2 distances of the form (3.6) with absolutely continuous reference σ, and for empirical distributions:

$$W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu};\gamma) := \|T_\sigma^{\widehat{\mu}}(\cdot;\gamma_{\widehat{\mu}}) - T_\sigma^{\widehat{\nu}}(\cdot;\gamma_{\widehat{\nu}})\|_\sigma, \tag{3.11}$$

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 σ 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 γ_{LP} and γ_β are understood to be in $\Gamma(\sigma,\widehat{\mu})$.

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

$$\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu};\gamma) := \left(\frac{1}{m}\sum_{j=1}^{m}\|T_\sigma^{\widehat{\mu}}(X_j;\gamma_{\widehat{\mu}}) - T_\sigma^{\widehat{\nu}}(X_j;\gamma_{\widehat{\nu}})\|^2\right)^{1/2}, \tag{3.12}$$

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

Remark 3.5. *When we use γ_β for a transport plan between $\widehat{\sigma}$ and $\widehat{\mu}$, note that our barycentric projection map is given by*

$$T_{\widehat{\sigma}}^{\widehat{\mu}}(x;\gamma_\beta) := \frac{\frac{1}{k}\sum_{i=1}^{k} y_i \exp\left(\left(g_{\beta,k}(y_i) - \frac{1}{2}\|x-y_i\|^2\right)/\beta\right)}{\frac{1}{k}\sum_{i=1}^{k} \exp\left(\left(g_{\beta,k}(y_i) - \frac{1}{2}\|x-y_i\|^2\right)/\beta\right)}, \tag{3.13}$$

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

Remark 3.6. *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 $x \in \mathrm{supp}(\widehat{\sigma})$; however, for general computation, we can just interpolate to calculate $T_\sigma^{\widehat{\mu}}(x)$ for $x \in \mathrm{supp}(\sigma)\setminus\mathrm{supp}(\widehat{\sigma})$.*

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 3.30 (Wasserstein-2 compared to LOT) and results in [38] (Wasserstein-2 compared to Wasserstein-2 on empirical distributions). This requires comparisons between all of $W_2(\mu,\nu)$, $W_{2,\sigma}^{\mathrm{LOT}}(\mu,\nu)$, $W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu};\gamma)$, and $\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu},\widehat{\nu};\gamma)$, which are discussed in Section 3.4 and Section 3.5.

3.2.6 Wassmap

Various generalizations of MDS have been explored [35] including stress minimization, which is useful in graph drawing [56, 69], Isomap [88] 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. [92] 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. [48] proved that several types of submanifolds of W_2 can be isometrically embedded via MDS with Wasserstein distances (as in [92]) and empirically studied Wassmap: a variant of Isomap that approximates nonlinear submanifolds of W_2. In particular, [48] shows that for some submanifolds of $W_2(\mathbb{R}^m)$ of the form $\mathcal{H}_\sharp \mu$ where $\mathcal{H} = \{h_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ 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).

3.2.7 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 $ca \leq b \leq Ca$ for some absolute constance $0 < c, C < \infty$.

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

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

We denote by $O(d)$ 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 $\min\limits_{Q\in O(d)} \|X - QY\|_{\mathrm{F}}$.

3.3 LOT Wassmap algorithm and Main Theorem

Here we present our main algorithm which is an LOT approximation to the Wassmap embedding of [48], and our main theorem which describes the quality of the embedding using some existing perturbation bounds for MDS.

3.3.1 The LOT Wassmap Embedding Algorithm

The algorithm presented here (Algorithm 3) 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 [48] which requires $O(N^2)$ Wasserstein distance computations, Algorithm 3 requires computation of only $O(N)$ optimal tranport maps. Given the high cost of computing a single optimal transport map for densely sampled measures, this represents a significant savings.

Note that the factor of $\frac{1}{\sqrt{m}}$ appearing in the computation of the final embedding is due to (3.12) where the $\frac{1}{m}$ appears in the definition of the empirical LOT distance. Lemma Lemma 3.27 shows that $T^\top T$ where T is as in Algorithm 3 is actually the MDS matrix $-\frac{1}{2}J\Lambda J$ where Λ

consists of the empirical LOT distances between the data, hence we absorb the $\frac{1}{m}$ into the norm in (3.12) to get the matrix T in Algorithm 3.

Algorithm 3 LOT WassMap Embedding

```
Input  : Reference point cloud $\{w_i\}_{i=1}^m \sim \sigma \in \mathcal{P}_2(\mathbb{R}^n)$
         Sample point clouds $\{x_j^k\}_{j=1}^{n_k} \sim \mu_k \in \mathcal{P}_2(\mathbb{R}^n)$ $(k = 1, \dots, N)$
         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 $\widetilde{T}_k(w_i) = \left(\sum_{j=1}^{n_k} x_j^k (\gamma_k)_{ij}\right) / \left(\sum_{j=1}^{n_k} (\gamma_k)_{ij}\right)$
$\widehat{T} = \left[\widetilde{T}_j(w_i)\right]_{i=1,j=1}^{m,n}$
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 $z_i = V_d \Sigma_d(i, :)$
```

3.3.2 MDS Perturbation Bounds

As stated above, the output of Algorithm 3 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 [9].

Theorem 3.7 ([9, Theorem 1]). *Let $Y, Z \in \mathbb{R}^{d \times N}$ with $d < N$ such that* $\operatorname{rank}(Y) = d$*, and let* $\varepsilon^2 := \|Z^\top Z - Y^\top Y\|_{S_p}$ *for some* $p \in [1, \infty]$. *Then,*

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

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

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

Corollary 3.8. *Let* $y_1,\ldots,y_N \in \mathbb{R}^d$ *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* $\|Y^\dagger\|\|\Lambda-\Delta\|_{\mathrm{S}_p}^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}}$*, then MDS (Algorithm 2) with input dissimilarities* $\{\Lambda_{ij}\}_{i,j=1}^N$ *and embedding dimension* d *returns a point set* $z_1,\ldots,z_N \in \mathbb{R}^d$ *satisfying*

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

Proof of Corollary 3.8. The proof follows along similar lines to that of [9, 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 $\|AB\|_{\mathrm{S}_p} \leq \|A\|\|B\|_{\mathrm{S}_p}$, we can estimate

$$\frac{1}{2}\|J(\Lambda-\Delta)J\|_{\mathrm{S}_p} \leq \frac{1}{2}\|J\|^2\|\Lambda-\Delta\|_{\mathrm{S}_p} \leq \frac{1}{2}\|\Lambda-\Delta\|_{\mathrm{S}_p} < \sigma_d^2(Y), \tag{3.14}$$

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 3.27). Thus by Weyl's inequality, the fact that $\|\cdot\| \leq \|\cdot\|_{\mathrm{S}_p}$ for all p, and (3.14),

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

Consequently, $-\frac{1}{2}J\Lambda J$ has rank d, so if Z contains the columns of the MDS embedding corresponding to Λ, 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 that

$$\left\|Z^\top Z + \frac{1}{2}J\Lambda J\right\|_{\mathrm{S}_p} \leq \left\|\frac{1}{2}J(\Lambda-\Delta)J\right\|_{\mathrm{S}_p}. \tag{3.15}$$

Combining (3.14) and (3.15), we have

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

Thus, $\|Y^\dagger\|\varepsilon \leq \|Y^\dagger\|\|\Lambda-\Delta\|_{\mathrm{S}_p}^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}}$, so we may apply the final bound of Theorem 3.7 to yield the conclusion. □

3.3.3 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 3.8 outright would require computing a proxy distance matrix and applying MDS; however, to make Algorithm 3 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 3, which does not require any distance matrix computation.

Theorem 3.9. *Let $\{\mu_i\}_{i=1}^N \subset W_2(\mathbb{R}^n)$. Suppose $\mathcal{W} \subset W_2(\mathbb{R}^n)$ is a subset of Wasserstein space that is isometric to a subset of Euclidean space $\Omega \subset \mathbb{R}^d$, and $\{\nu_i\}_{i=1}^N \subset \mathcal{W}$ and $\{y_i\} \subset \Omega$ 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 2) with input Λ.*

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 τ_1 and τ_2, and

$$\|Y^\dagger\| \sqrt{N} (\tau_1 + \tau_2)^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}}, \tag{3.16}$$

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

$$\min_{Q \in O(d)} \|Z - QY\|_{\mathrm{F}} \leq (1 + \sqrt{2}) \|Y^\dagger\| N (\tau_1 + \tau_2).$$

Proof. Note that

$$\|\Lambda - \Delta\|_{\mathrm{F}} \leq \|\Gamma - \Delta\|_{\mathrm{F}} + \|\Lambda - \Gamma\|_{\mathrm{F}} \leq N(\tau_1 + \tau_2).$$

Consequently, (3.16) allows us to apply Corollary 3.8 to yield the conclusion. □

Specializing Theorem 3.9 to the case of Algorithm 3 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 3.9.

Corollary 3.10. *Invoke the notations and assumptions of Theorem 3.9. 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 3. Let $U_d\Sigma_d V_d^\top$ be the truncated SVD of T, and let $z_i = V_d\Sigma_d(i,:)$ for $1 \le i \le N$ (i.e., z_i is the output of Algorithm 3). If* (3.16) *holds, then*

$$\min_{Q\in O(d)} \|Z - QY\|_{\mathrm{F}} \le (1+\sqrt{2})\|Y^\dagger\| N\,(\tau_1+\tau_2)\,.$$

Proof. Since T is centered, Lemma 3.27 implies that $T^\top T = JT^\top TJ = -\frac{1}{2}J\Lambda J$. Consequently, if $-\frac{1}{2}J\Lambda J = V\Sigma^2V^\top = T^\top T$, 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 Λ. The conclusion follows by direct application of Theorem 3.9.

□

In the rest of the paper, we will discuss how various LOT approximations to Wasserstein distances affect the value of the bound τ_2 appearing in Theorem 3.9 and Corollary 3.10. In particular, we get different values of τ_2 when we have compactly supported target measures (as in Theorem 3.12 for linear programming estimators and Theorem 3.17 for Sinkhorn estimators) and non-compactly supported target measures (as in Theorem 3.21 for linear programming estimators and Theorem 3.22 for Sinkhorn estimators).

3.4 Bounds for compactly supported target measures

To capture the bound τ_2 of Theorem 3.9, 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 μ_i, and a discretization of the reference distribution

σ. In particular, the main approximation argument consists of the following triangle inequality:

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

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

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

The error from (1) and (3) are handled in Section 3.10 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 3.12 and Theorem 3.17. 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 σ, μ, and $\mathcal{H}$:

Assumption 3.11. *Consider the following conditions on σ, μ, and $\mathcal{H}$*

i $\sigma \in \mathcal{P}_{ac}(\Omega)$ *for a compact convex set* $\Omega \subseteq B(0,R) \subset \mathbb{R}^n$ *with probability density* f_σ *bounded above and below by positive constants.*

ii *μ has finite p-th moment with bound M_p with $p > n$ and $p \geq 4$.*

iii *There exist $a, A > 0$ such that every $h \in \mathcal{H}$ satisfies $a\|x\| \leq \|h(x)\| \leq A\|x\|$.*

iv *$\mathcal{H}$ is compact and ε-compatible with respect to $\sigma, \mu \in W_2(\mathbb{R}^n)$. Moreover, $\sup_{h,h' \in \mathcal{H}} \|h - h'\|_\mu \leq M$.*

v *$\mu_i \sim \mathcal{H}_\sharp \mu$ i.i.d.*

These assumptions ensure that ε-compatible transformations are also "ε-isometric" as shown in Theorem 3.30.

3.4.1 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 3.12. *Let $\delta > 0$. Along with Assumption 3.11 and that $\mu \in \mathcal{P}_{ac}(\Omega)$ for the Ω in Assumption 3.11, assume that*

(i) $T_\sigma^{\mu_i}$ is L-Lipschitz, which may occur if T_σ^{μ} is L-Lipschitz. Note that if σ and μ are both compactly supported, then T_σ^{μ} itself is L-Lipschitz.

(ii) We estimate μ_i with an empirical measure $\widehat{\mu_i}$ using k samples and discretize σ with m samples. Let our estimator be given by (3.10) *with γ solved using linear programming.*

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

$$\left| W_2(\mu_1, \mu_2)^2 - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1}, \widehat{\mu_2}; \gamma_{LP})^2 \right| \leq (M + 2R) \left(C\varepsilon^{\frac{p}{6p+16n}} + 2O_p(r_n^{(k)} \log(1+k)^{t_{n,\alpha}}) \right.$$
$$\left. + R\sqrt{\frac{2\log(2/\delta)}{m}} \right). \quad (3.17)$$

where C is the constant from Theorem 3.30 depending on n, p, Ω, M_p, the constants a and A come from Assumption 3.11 (iv), and

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

$$t_{n,\alpha} = \begin{cases} (4\alpha)^{-1}(4 + ((2\alpha + 2n\alpha - n) \vee 0)) & n < 4 \\ (\alpha^{-1} \vee 7/2) - 1 & n = 4 \\ 2(1+n^{-1}) & n > 4 \end{cases},$$

so that $r_n^{(k)}$ and $t_{n,\alpha}$ are on the order of $k^{-1/n}$ and $2(1+n^{-1})$, respectively. In this case, τ_2 of Corollary 3.10 is bounded above by the right-hand side of (3.17).

Proof. Note that the transport plan that we are using for the following proof is γ_{LP}. Henceforth, we will suppress γ_{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 μ_1 and μ_2 are pushforwards of a fixed template distribution μ, we know that $\mu_i = h_{i\sharp}\mu$, where by [6, 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 $\mathcal{H}$ is compact, μ is compactly supported, and $\mu_i \sim \mathcal{H}_\sharp\mu$, we know that μ_i is compactly supported with $\mathrm{supp}(\mu_i) \subseteq B(0,R)$ for all i. This implies that

$$\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2}) = \left(\frac{1}{m}\sum_{j=1}^{m} \underbrace{|T_\sigma^{\widehat{\mu_1}}(X_j) - T_\sigma^{\widehat{\mu_2}}(X_j)|^2}_{\leq 4R^2}\right)^{1/2} \leq 2R.$$

Putting these estimates together, we have

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

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 3.11, we can use ε-compatibility of $\mathcal{H}$ in Theorem 3.30 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 3.30.

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

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

Note that Assumption 3.11(i) implies that there exists some $t > 0$ and $\alpha > 0$ such that $\mathbb{E}_\sigma[t\|x\|^\alpha] < \infty$. Together with T_σ^μ being Lipschitz, this allows us to use Theorem 3.31 to conclude that

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

c) From Theorem 3.33 we know that with probability at least $1-\delta$,

$$\left|W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu}_1,\widehat{\mu}_2) - \widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu}_1,\widehat{\mu}_2)\right| \le R\sqrt{\frac{2\log(2/\delta)}{m}}.$$

Putting these bounds together yields the result. □

3.4.2 Using entropic regularization (Sinkhorn) to compute transport maps

Although [38] 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 (3.9) and using the barycentric projection map (3.13) 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 3.13. *Since we can choose σ 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 σ than μ_i (i.e. $m > k$) will result in better approximation.*

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

Definition 3.14. *Consider the Wasserstein geodesic between $\sigma = \mu_0$ and $\mu = \mu_1$ with μ_t being the measure on the geodesic for $t \in (0,1)$. Let $f(t,x)$ be the density corresponding to μ_t. Then the* *integrated Fisher information along the Wasserstein geodesic between σ and μ 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 $\phi \in \mathbb{R}^n$ is given by

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

see, e.g., [8, p. 45]. Now by using Theorem 3 from [80], 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 3.15 ([80, Theorem 3]). *Assume that*

(A1) $\sigma, \mu_i \in \mathcal{P}_{ac}(\Omega)$ *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* ϕ^* *denotes the convex conjugate of* ϕ*.*

(A3) $T_\sigma^{\mu_i} = \nabla\phi$ *with* $mI \preceq \nabla^2\phi(x) \preceq LI$ *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*

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

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* σ *and* μ_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 3.15 we have that

$$\begin{aligned}\mathbb{E}[Z_k] \leq \mathbb{E}\big[Z_k^2\big]^{1/2} &\leq \sqrt{\big(1+I_0(\sigma,\mu_i)\big)k^{-\frac{\widetilde{\alpha}+1}{2n'+\widetilde{\alpha}+1}}\log k} \\ &= \sqrt{\log(k)(1+I_0(\sigma,\mu_i))}k^{-\frac{\widetilde{\alpha}+1}{2(2n'+\widetilde{\alpha}+1)}}.\end{aligned}$$

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

Corollary 3.16. *Assume that σ and μ_i satisfy (A1)–(A3) of Theorem 3.15 and let $\delta > 0$. Then with probability at least $1-\delta$, we have that*

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

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 3.15 and the asymptotic bound on the regularization parameter β used in the entropic regularization.

Theorem 3.17. *Let $\delta > 0$. Along with Assumption 3.11 and $\mu \in \mathcal{P}_{ac}(\Omega)$ for the Ω in Assumption 3.11, assume that*

(i) *σ and μ_i satisfy assumptions (A1)–(A3) from Theorem 3.15 for all i. Note that (A1), regularity of ϕ 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 [80]. Assume our estimator is $T_{\widehat{\sigma}}^{\widehat{\mu}_i}(\,\cdot\,;\gamma_\beta)$ given by* (3.13).

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

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

where C is from Theorem 3.30 and $I_0(\sigma,\mu_i)$ *is defined in Theorem 3.15. In this case,* τ_2 *in Corollary 3.10 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 γ_β. Henceforth, we will suppress γ_β 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 3.12, we find that

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

Similar to the proof of Theorem 3.12, we bound

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

The first and last term are bounded the same way as in the proof of Theorem 3.12 above. Since assumption (i) of Assumption 3.11, implies assumption (A1) of Theorem 3.15, we get that with probability at least $1-\delta$

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

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

Using Theorem 3.12 and Theorem 3.17, we see that as long as μ_i are ε-compatible pushforwards of μ 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.

3.5 Bounds for non-compactly supported target measures

In the last section, we saw that for compactly supported $\mu_i \sim \mathcal{H}_\sharp\mu$ (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 [48] via Algorithm 3. In this section, we show that we can adapt Theorem 3.12 and Theorem 3.17 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 [39].

Theorem 3.18 ([39])**.** *Let Ω be a compact convex set and let σ be a probability density on Ω, bounded from above and below by positive constants. Let $p > n$ and $p \geq 4$. Assume that $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^n)$ 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 μ 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 μ well (i.e., $W_1(\mu,\widetilde{\mu}) < \eta$.). We achieve this in the following lemma.

Lemma 3.19. *Fix $\eta > 0$, and let σ satisfy the assumptions of Theorem 3.18. Moreover, let*

$\mu \in \mathcal{P}_2(\mathbb{R}^n)$ with density f_μ 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

$$f_\mu(x) < \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{3C\|x\|^{n+2}},$$

where C denotes the constant from integrating over concentric n-spheres. Then there exists a compactly supported absolutely continuous measure $\widetilde{\mu}$ such that

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

The next lemma will be useful in establishing conditions on $\mathcal{H}$ and μ so that our truncated measure has a density that is bounded away from 0.

Lemma 3.20. *Let σ satisfy the assumptions of Theorem 3.18 and let $\mu \in \mathcal{P}_2(\mathbb{R}^n)$ 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*

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

where C_{n,p,Ω,M_p} comes from from Theorem 3.18, C' is a constant from integrating over concentric n-spheres as well as 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 3.19 and Lemma 3.20 are located in Section 3.11. With these two lemmas above, we obtain the following theorems. Note that Theorem 3.21 replaces the assumption that μ is compactly supported with one of polynomial (in the ambient dimension) tail decay; while the second assumption below is the same as Theorem 3.12, the final assumption differs from that of Theorem 3.12 by requiring the discretizations of σ and μ_i to have the same sample size to apply the lemmas above.

Theorem 3.21. *Let $\delta > 0$. Along with Assumption 3.11, assume that*

(i) Every μ_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 $x \notin B(0,R)$, we have

$$f_{\mu_i} < \left(\frac{\eta}{C_{n,p,\Omega,M_p}}\right)^{\frac{6p+16n}{p}} \frac{1}{3C\|x\|^{n+2}}.$$

Define $\widetilde{\mu}_i$ to be the truncated measure found in Lemma 3.19 or Lemma 3.20 such that $W_1(\mu_i,\widetilde{\mu}_i) < \varepsilon$.

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

(iii) Given empirical distributions $\widehat{\sigma}$ and $\widehat{\mu_i}$ with $\operatorname{supp}(\widehat{\mu_i}) \subseteq B(0,R)$ and sample sizes m and k, respectively, let our estimator be the barycentric estimator (3.10), with γ_{LP}.

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

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

where $r_n^{(k)}$ and $t_{n,\alpha}$ are defined in Theorem 3.12 and C is a constant coming from Theorem 3.30. In this case, τ_2 of Corollary 3.10 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 3.22 and Theorem 3.21 is the addition of (A1)–(A3) from

Theorem 3.15 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 3.17, but with $\widehat{\mu}_i$ replaced with $\widetilde{\mu}_i$ arising from Theorem 3.21, whereas the additional assumptions below are that μ_i have decaying tails as opposed to being compactly supported.

Theorem 3.22. *Let $\delta > 0$. Along with Assumption 3.11 and (i) of Theorem 3.21, assume that*

(i) σ and $\widetilde{\mu}_i$ satisfy assumptions (A1)–(A3) in 3.15 for all i, where $\widetilde{\mu}_i$ is the truncated measure from Theorem 3.21.

(ii) Given empirical distributions $\widehat{\sigma}$ and $\widehat{\mu}_i$ with $\operatorname{supp}(\widehat{\mu}_i) \subseteq B(0,R)$ 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 3.15. Moreover, assume our estimator is given by (3.13).

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

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

where $I_0(\sigma,\mu_i)$ is defined in Theorem 3.15 and C is a constant from Theorem 3.30. In this case, τ_2 of Corollary 3.10 is bounded above by the right-hand side of the inequality above.

The following is a proof for both theorems above.

Proof of Theorems 3.21 and 3.22. 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 γ_{LP} or the entropic estimator with γ_β) 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

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

where $\widetilde{\mu_i}$ is defined as in the theorem statement and $\widehat{\mu_i}$ denotes the empirical measure of μ_i. Since we assume that $\mathrm{supp}(\widehat{\mu_i}) \subseteq B(0,R)$, we know that $\widehat{\mu_i}$ can equivalently be thought of as being sampled from $\widetilde{\mu_i}$ rather than μ_i. 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 μ_i are compactly supported can be extended to non-compactly supported μ_i as long as their densities decay fast enough and the reference distribution σ has a compact and convex support.

3.6 Conditions on $\mathcal{H}$ and μ (Compact case)

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

Theorem 3.23 (Barycentric Map Case (Compact)). *Along with Assumption 3.11 (with $\varepsilon = 0$ so that every $h \in \mathcal{H}$ is exactly compatible with σ and μ), assume that $\mu_i \sim \mathcal{H}_\sharp\mu$ i.i.d. and that*

(i) μ is compactly supported,

(ii) σ is chosen such that T_σ^μ is Lipschitz,

Then μ_i satisfies the conditions of Theorem 3.12, i.e., each μ_i is compactly supported and $T_\sigma^{\mu_i}$ is Lipschitz.

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

Theorem 3.24 (Entropic Map Case (Compact)). *Under the assumptions of Theorem 3.23, as well as*

(iv) σ and μ satisfy (A1)-(A3),

μ_i satisfies the conditions of Theorem 3.17.

The proofs of both Theorems 3.23 and 3.24 are given in Section 3.12.1.

3.7 Conditions on $\mathcal{H}$ and μ (Non-compact case)

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

Theorem 3.25 (Barycentric Map Case (Non-Compact)). *Along with Assumption 3.11 (with $\varepsilon = 0$ so that every $h \in \mathcal{H}$ is exactly compatible with σ and μ), assume that $\mu_i \sim \mathcal{H}_\sharp\mu$ i.i.d. Assume further that*

(i) for every $h \in \mathcal{H}$, there exists an inverse $h^{-1} \in \mathcal{H}$.

(ii) The density of μ 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 $x \in B(0,RL)$. Moreover, f_μ has a decay rate as in Lemma 3.20 for $x \notin B(0,R)$.

Then μ_i satisfies the conditions of Theorem 3.21.

Theorem 3.26 (Entropic Map Case (Non-Compact))**.** *Assume that $\mu_i \sim \mathcal{H}_\sharp\mu$ i.i.d. and that μ, $\mathcal{H}$, and σ satisfy the conditions of Theorem 3.25. Then μ_i satisfies the conditions of Theorem 3.22.*

The proofs of both Theorems 3.25 and 3.26 are found in Section 3.12.2.

3.8 Experiments

We demonstrate that Algorithm 3 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 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, μ_i, which are Gaussians of various means and covariance, and a fixed reference measure σ 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 μ_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`.

3.8.1 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 μ_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 μ_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 σ. We randomly sample $m = 1000$ points from each data measure and the reference measure independently. Figure 3.1 shows the original sampled data and the reference measure (in blue), the true embedding points x_i, and the embeddings of Algorithm 3 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 [48] in similar experiments. Figure 3.2 shows the relative error vs. sampling size m of the measures, and one can see the good performance for modest sample sizes.

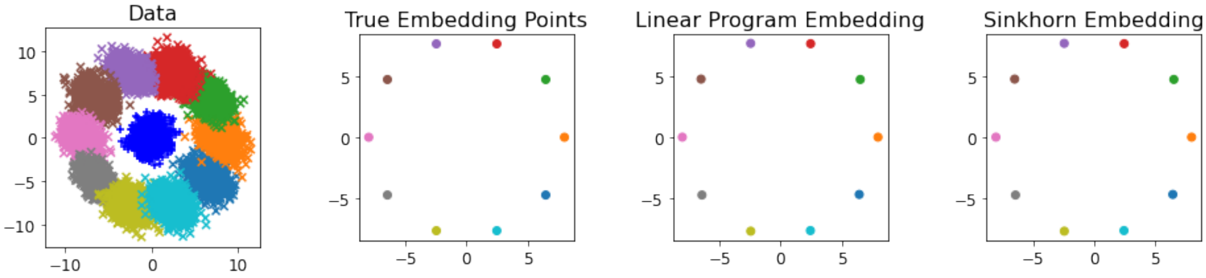


Figure 3.1: 1-D Manifold of translations: **(Left)** reference measure $\sigma \sim \mathcal{N}(0,I)$ in blue and data measures μ_i which are Gaussians with the same covariance matrix and means x_i uniformly sampled from the circle of radius 8. **(Left Middle)** Means x_i of μ_i which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 3 using the linear program. **(Right)** Embedding attained with Algorithm 3 using the Sinkhorn distance with $\lambda = 1$.

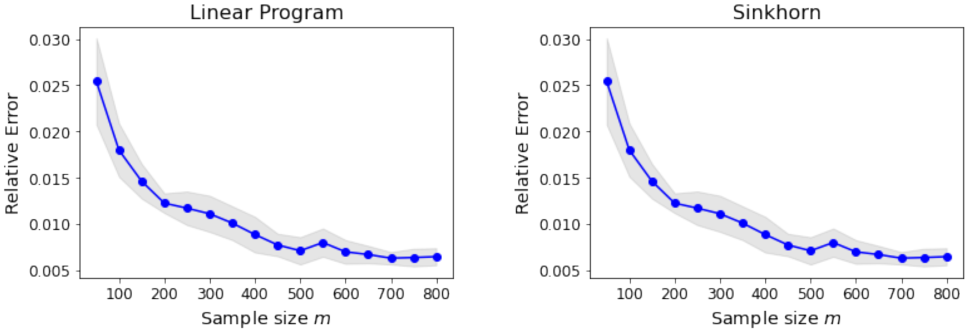


Figure 3.2: Embedding error vs. m (number of sample points from data and reference distributions for the 1-D translation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 1$ **(Right)**.

3.8.2 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 $\begin{bmatrix} 2 & 0 \\ 0 & .5 \end{bmatrix}$ by the angles θ_i. As in experiment 1, the noise level added is 0.5 and we sample $m = 1000$ points from each measure. Figure 3.3 shows the data measures, true embedding, and embeddings from Algorithm 3 using both the linear program and Sinkhorn (with $\lambda = 1$) to compute the optimal transport maps. Figure 3.4 shows the relative error vs. sample size.

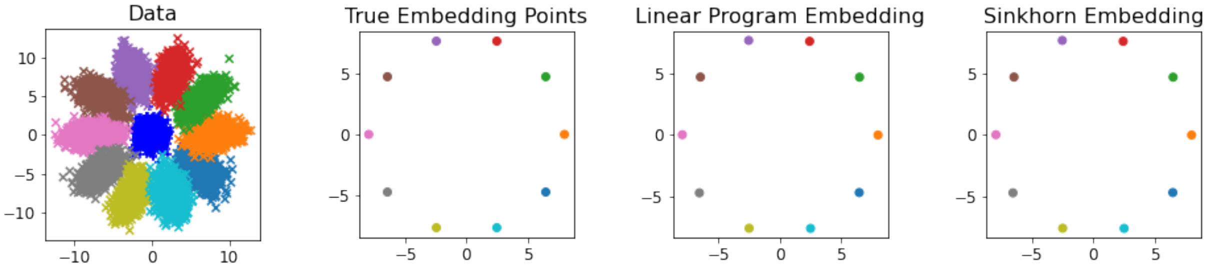


Figure 3.3: 1-D Manifold of rotations: **(Left)** reference measure $\sigma \sim \mathcal{N}(0,I)$ in blue and data measures μ_i which are Gaussians with means lying on the circle of radius 8 and covariance matrices that are rotations of each other. **(Left Middle)** Means x_i of μ_i which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 3 using the linear program. **(Right)** Embedding attained with Algorithm 3 using the Sinkhorn distance with $\lambda = 1$.

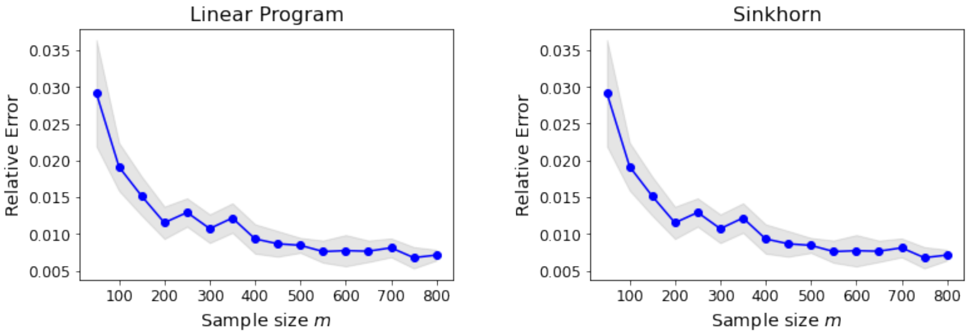


Figure 3.4: Embedding error vs. m (number of sample points from data and reference distributions for the 1-D rotation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 1$ **(Right)**.

3.8.3 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 5×5 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 $\lambda = 10$. Figures 3.5 and 3.6 show the data, embeddings, and relative error vs. sample size.

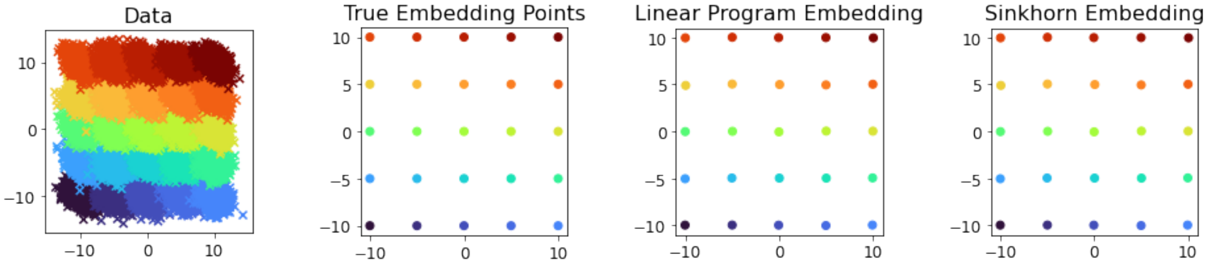


Figure 3.5: 2-D Manifold of translations: **(Left)** data measures μ_i which are Gaussians with the same covariance matrix and means x_i taken from a 5×5 uniform grid on $[-10,10]^2$. **(Left Middle)** Means x_i of μ_i which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 3 using the linear program. **(Right)** Embedding attained with Algorithm 3 using the Sinkhorn distance with $\lambda = 10$.

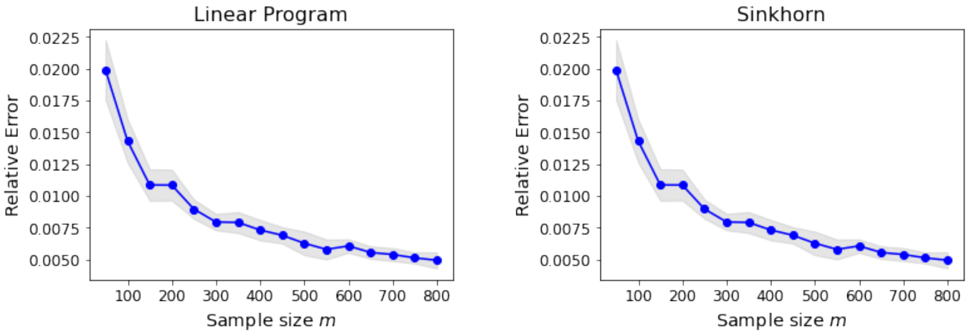


Figure 3.6: Embedding error vs. m (number of sample points from data and reference distributions for the 2-D translation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 10$ **(Right)**.

3.8.4 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 $\mathrm{diag}(\alpha_i^2, \beta_i^2)$ for (α_i, β_i) taken from a uniform 3×3 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 $\lambda = 100$. Figure 3.7 show the data measures, true embedding parameters, and embeddings from Algorithm 3. Note that the true embedding parameters are centered to allow them to be comparable to the output of Algorithm 3 which are naturally centered.

Figure 3.8 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.

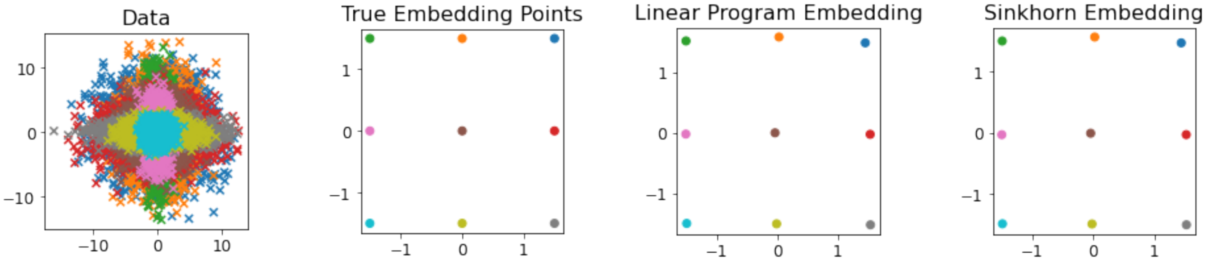


Figure 3.7: 2-D Manifold of Anisotropic Dilations: **(Left)** data measures μ_i which are Gaussians with mean 0 and anisotropically dilated covariance matrices where dilations are taken from a 3×3 uniform grid on $[1,4]^2$. **(Left Middle)** Dilation factors (x_i, y_i) of μ_i which are the true embedding points. **(Right Middle)** Embedding attained with Algorithm 3 using the linear program. **(Right)** Embedding attained with Algorithm 3 using the Sinkhorn distance with $\lambda = 100$.

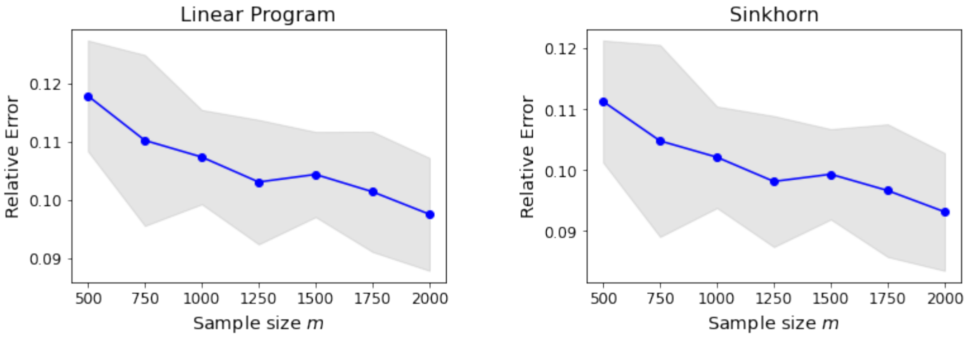


Figure 3.8: Embedding error vs. m (number of sample points from data and reference distributions for the 2-D translation manifold. Optimal transport maps are computed via the Linear Program **(Left)** and Sinkhorn with $\lambda = 10$ **(Right)**.

3.8.5 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 3 using the Linear Program or Sinkhorn with $\lambda = 1$ and the Wassmap algorithm of [48] 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 2. 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 (Figure 3.10). As expected, larger regularization parameter yields faster computation time, though the difference is relatively small even for modestly large sample size.

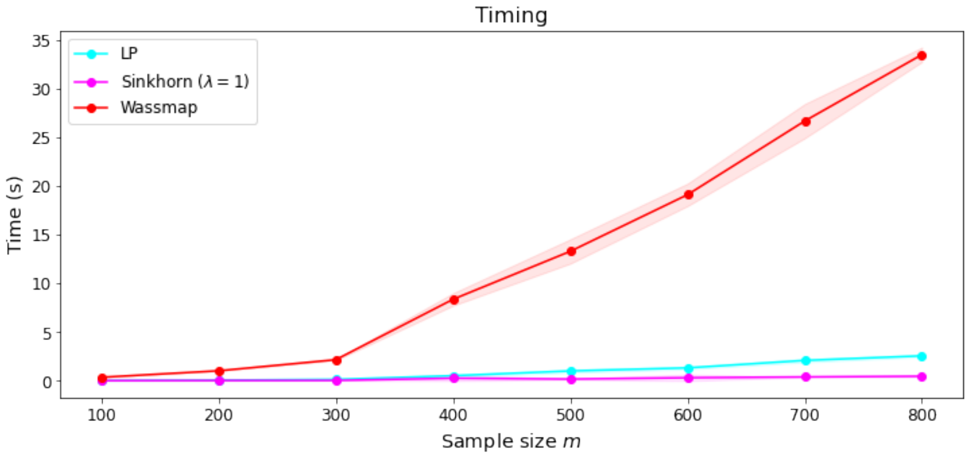


Figure 3.9: Timing vs. sample size m of the reference distribution and data measures. The data set consists of $N = 25$ measures translated on a 5×5 uniform grid on $[-10, 10]^2$ as in Experiment 3. Shown are the computation times to compute the Wassmap embedding and the embeddings of Algorithm 3 using the Linear Program (LP) and Sinkhorn with regularization parameter $\lambda = 1$.

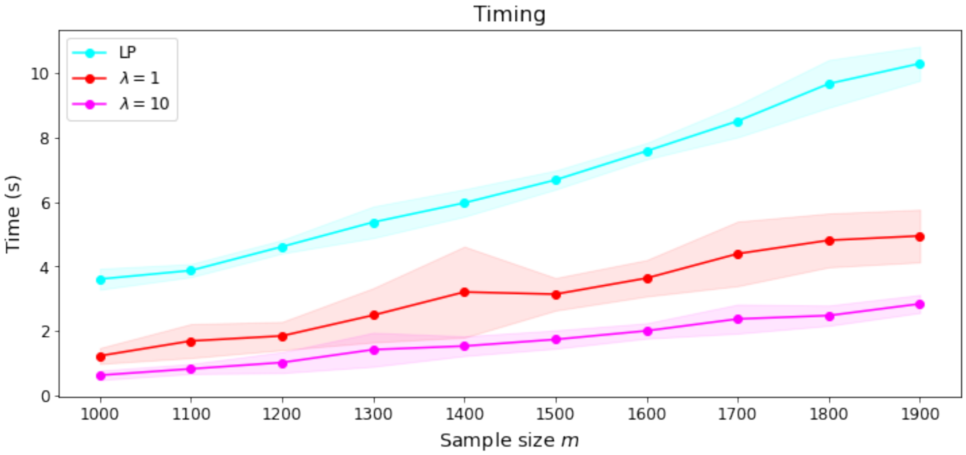


Figure 3.10: Timing vs. sample size m of the reference distribution and data measures. The data set consists of $N = 25$ measures translated on a 5×5 uniform grid on $[-10, 10]^2$ as in Experiment 3. Shown are the computation times to compute the embeddings of Algorithm 3 using the Linear Program (LP) and Sinkhorn with regularization parameters $\lambda = 1$ and $\lambda = 10$.

Acknowledgements

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 award DMS-2306064. A.C. is partially supported by NSF award DMS-2012266 and a gift from Intel research. 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.

3.9 Helper Theorems and Lemmas

We use the following lemma to extend Corollary 3.8 to get our main theorem (Theorem 3.9). The proof follows standard arguments, e.g., as in [64]; the proof is included for completeness.

Lemma 3.27 ([64, 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 2), $G = V^\top V$ be the Gram matrix for V, and D be the squared distance matrix $D_{ij} = \|v_i - v_j\|^2$. Then $G = -\frac{1}{2}JDJ$.*

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

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

Note here that $V\mathbf{1} = 0$ since $\sum_{j=1}^n v_j = 0$. After cancelling 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 ε-compatibility as well as its effects on LOT. First, we show that every ε-compatible map has a compatible map (with ε = 0) nearby whose LOT distance from the ε-compatible map is small.

Lemma 3.28. *Assume that*

(i) σ *is supported on a compact convex set* $\Omega \subset \mathbb{R}^n$ *with probability density* f_σ *bounded above and below by positive constants.*

(ii) μ *has finite p-th moment with bound* M_p *with* $p > d$ *and* $p \geq 4$.

(iii) There exist $a, A > 0$ *such that every* $h \in \mathcal{H}$ *satisfies* $a\|x\| \leq \|h(x)\| \leq A\|x\|$.

Let $\mathcal{H}$ *be* ε*-compatible with respect to* σ *and* μ*. Then for every* $h \in \mathcal{H}$ *there exists a compatible* g *such that*

$$\left\| T_\sigma^{g_\sharp\mu} - T_\sigma^{h_\sharp\mu} \right\|_\sigma \leq C_{n,p,\Omega,a^{-1}A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}$$
$$\|h \circ T_\sigma^\mu - T_\sigma^{h_\sharp\mu}\|_\sigma < \varepsilon + C_{n,p,\Omega,a^{-1}A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}.$$

Proof. Let $h \in \mathcal{H}$, 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 ε-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 $h \in \mathcal{H}$ and are Lipschitz, we know that

$$\int_\Omega \|x\|^p f_{h_\sharp\mu}(x)dx = \int_\Omega \underbrace{\|h(x)\|^p}_{\leq A^p\|x\|^p} \underbrace{|J_{h^{-1}}(x)|}_{a^{-1}} f_\mu(x)dx \leq a^{-1}A^pM_p.$$

Similarly, we have the same bound for g since $g \in \mathcal{H}$. Now using Theorem 3.18 and equation 2.1

of [6], we get that

$$\begin{aligned}\left\|T_\sigma^{g_\sharp\mu} - T_\sigma^{h_\sharp\mu}\right\|_\sigma &\leq C_{n,p,\Omega,a^{-1}A^pM_p} W_1(g_\sharp\mu, h_\sharp\mu)^{\frac{p}{6p+16n}} \\ &\leq C_{n,p,\Omega,a^{-1}A^pM_p} W_2(g_\sharp\mu, h_\sharp\mu)^{\frac{p}{6p+16n}} \\ &\leq C_{n,p,\Omega,a^{-1}A^pM_p} \|h-g\|_\mu^{\frac{p}{6p+16n}} \\ &\leq C_{n,p,\Omega,a^{-1}A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}.\end{aligned}$$

This implies that

$$\|h\circ T_\sigma^\mu - T_\sigma^{h_\sharp\mu}\|_\sigma < \varepsilon + C_{n,p,\Omega,a^{-1}A^pM_p} \cdot \varepsilon^{\frac{p}{6p+16n}}.$$

□

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

Lemma 3.29. *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

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

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

Note that Theorem 24 of [57] implies that given an exactly compatible transformation g, $J_g(T_\sigma^\mu(x))$ must share the same eigenspaces as $J_{T_\sigma^\mu}(x)$. By Corollary 4 of [57], we know that exactly compatible transformations are optimal transport maps themselves. This means that $T_\mu^{g_\sharp\mu} = g$ for exactly compatible transport maps. Moreover, for an exactly compatible $h' \in \mathcal{H}$, 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

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

In particular, we get that

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

□

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

Theorem 3.30. *Assume that*

(i) σ is supported on a compact convex set $\Omega \subset \mathbb{R}^n$ with probability density f_σ bounded above and below by positive constants.

(ii) *μ has finite p-th moment with bound M_p with $p > n$ and $p \geq 4$.*

(iii) *There exists constants $a, A > 0$ such that Every $h \in \mathcal{H}$ satisfies $a\|x\| \leq \|h(x)\| \leq A\|x\|$.*

Let $\mathcal{H}$ be ε-compatible with respect to absolutely continuous measures σ and μ and that $h_\sharp\mu$ is absolutely continuous. Then for $h_1, h_2 \in \mathcal{H}$,

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

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

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

By Lemma 3.29, we know that

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

However, by equation 2.1 of [6] 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 3.28, for $i = 1, 2$, we know that

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

This implies that

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

and the proof is complete. □

3.10 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 3.4.

3.10.1 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 [38] is used in the proofs of our main results, including Theorem 3.12.

Theorem 3.31 ([38, Theorem 2.2])**.** *Suppose that T_σ^μ is L-Lipschitz, and μ 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 μ 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

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

$$t_{n,\alpha} = \begin{cases} (4\alpha)^{-1}(4 + ((2\alpha + 2n\alpha - n) \vee 0)) & n < 4 \\ (\alpha^{-1} \vee 7/2) - 1 & n = 4 \\ 2(1+n^{-1}) & n > 4 \end{cases},$$

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

Remark 3.32. *We note that Theorem 3.31 is the "semi-discrete" version described in [38]. The paper also provides equivalent bounds in the instance that* σ *is similarly estimated. However, the bounds only guarantee that the transport maps agree when integrated against* $\widehat{\sigma}$*, whereas we need the bound for* σ *itself.*

3.10.2 Approximating with Finite Samples from the Reference Distribution

Some of the norms from Theorem 3.12 and Theorem 3.17 are assumed to be integrated against the true σ. However, we need to consider the discretized σ 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

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

where $X_j \sim \sigma$, $\gamma_{\widehat{\mu_j}}$ is a transport plan between σ 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 μ_i are supported in a ball of radius R, then McDiarmid's inequality implies

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

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

$$\mathbb{P}\left(\left|\widehat{W}_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2-W_{2,\sigma}^{\mathrm{LOT}}(\widehat{\mu_1},\widehat{\mu_2};\gamma)^2\right|>t\right)\leq 2e^{-m\frac{t^2}{32R^4}}. \tag{3.18}$$

Theorem 3.33. *Consider $\mu_i,\sigma\in W_2(\mathbb{R}^n)$ with σ absolutely continuous with respect to the Lebesgue measure. 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 σ.

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 (3.18), 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.

$\square$

3.11 Non-Compactly Supported Measures Proofs and Results

Here, we give the proofs of the lemmas preceding Theorems 3.21 and 3.22.

Proof of Lemma 3.19. We will construct the measure $\widetilde{\mu}$ by constructing a transport map that sends μ to a compactly supported absolutely continuous measure. The compact set that $\widetilde{\mu}$ will be supported on is going to be $\overline{B(0,R)}$. In particular, for some $0 < \rho \ll 1$, consider the map

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

Then let $\widetilde{\mu} = (S_{R,\rho})_\sharp \mu$, and note that

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

However, recall that $d\mu(x) = f_\mu(x)dx$; thus,

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

where C is a constant from integrating over concentric n-spheres. Invoking Theorem 3.18, this means that

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

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

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

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 λ, then

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

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 μ so that $\mu(\partial B(0,R)) \leq \lambda(\partial B(0,R)) = 0$. Moreover, note that $\mu(A \cap B(0,R)) \leq \mu(A) \leq \lambda(A) = 0$. 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 μ 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 λ, so the proof is complete. □

Proof of Lemma 3.20. Rather than constructing a transport map, we will construct a density $f_{\widetilde{\mu}}$ and will argue that the transport map from μ 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 α. 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 $0 < c \leq b \leq f_{\widetilde{\mu}} \leq B < \infty$ for some constants b and B.

Now, we would like to bound $W_1(\mu, \widetilde{\mu})$. Let us consider S such that $S_\sharp\mu = \widetilde{\mu}$ and $S(x) = x$ if $x \in B(0,R)$. Such an S exists because we can consider the pushforward that is the identity on $B(0,R)$ and pushes the rest of the mass of μ from $\mathbb{R}^d \setminus B(0,R)$ to $B(0,a) \setminus B(0,R)$. 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 $\|S(x)\| \leq \widetilde{C}\|x\|$ (if $a < 2R$, then $\widetilde{C} \leq 2$). For the following calculation, we assume that

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

where C_{sphere} denotes a constant from integrating over concentric n-spheres and C_{n,p,Ω,M_p} denotes the constant from Theorem 3.18. Now note that

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

Invoking Theorem 3.18, this means that

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

Thus, we have the desired result. □

3.12 Proofs and Results for Conditions on $\mathcal{H}$ and μ

This section provides the proofs of the results in Sections 3.6 and 3.7.

3.12.1 Compact Case Proofs and Results

Here we prove the results of Section 3.6 which provide conditions on σ, μ, and $\mathcal{H}$ which guarantee that $\mu_i \sim \mathcal{H}_\sharp \mu$ satisfy the conditions of the theorems from Section 3.4.

Proof of Theorem 3.23. For the barycentric map estimator, we need to show that the μ_i's are compactly supported within a ball of radius R and $T_\sigma^{\mu_i}$ is Lipschitz.

- **Compact Support:** To ensure that a given μ_i is compactly supported, it suffices for μ to have compact support and $\mathcal{H}$ to consist of continuous maps. Indeed, under these assumptions, μ_i is compactly supported since the image of a compact set under a continuous map is compact. Since we are considering only a finite number of measures $\{\mu_i\}_{i=1}^N$, each with compact support, there exists a sufficiently large radius R such that $\text{supp}(\mu_i) \subseteq B(0,R)$ for all i.

- **Lipschitz OT Map:** To make sure that each $T_\sigma^{\mu_i}$ is Lipschitz, we will need that h_i is Lipschitz. In particular, we note that $\mu_i = (h_i)_\sharp\mu$ for some $h_i \in \mathcal{H}$. Thus, 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_σ^{μ} is Lipschitz, then $T_\sigma^{\mu_i}$ is Lipschitz.

□

Proof of Theorem 3.24. For the entropic map estimator, the μ_i's need to again be compactly-supported, $T_\sigma^{\mu_i}$ needs to be Lipschitz, and σ and μ_i together satisfy assumptions $(A1)-(A3)$. It will turn out, that we will only need to assume that there exist constants $a, A > 0$ such that

$$aI \preceq J_h(x) \preceq AI.$$

That μ_i is compactly supported and each $T_\sigma^{\mu_i}$ are Lipschitz follow from the same analysis as in the proof of Theorem 3.23.

- **Ensuring that μ_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 [57, Corollary 4], we know that h is an optimal transport map if it is compatible. This implies that $J_h(x)$ is positive semidefinite; however, if h is positive definite and Lipschitz (i.e.

$$aI \preceq J_h(x) \preceq AI$$

for some $\widetilde{m}, M > 0$), we know that

$$A^{-1}I \preceq J_{h^{-1}}(x) \preceq a^{-1}I.$$

This implies 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

$$A^{-d} \leq |J_{h^{-1}}(x)| = \prod_{j=1}^{n} \lambda_j(J_{h^{-1}}(x)) \leq a^{-n}.$$

Finally, since μ itself adheres to (A1), this implies that

$$\frac{b}{A^n} \leq f_\mu(x)|J_{h^{-1}}(x)| \leq \frac{B}{a^n}.$$

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

$$aI \preceq J_h(x) \preceq AI.$$

- **Ensuring that μ_i satisfy** $(A2)$**:** From [52, Corollary 4.2.10], we can ensure that $(A2)$ is satisfied if $(A3)$ is satisfied, which is proved below.

- **Ensuring that μ_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 [57, Theorem 24] gives that

$$(ma)I \preceq J_{T_\sigma^{h_\sharp\mu}}(x) \preceq (AL)I$$

for all x, where m and L come from assuming σ and μ satisfy (A3) whilst a and A come from Assumption 3.11. So $(A3)$ holds for σ and $\widetilde{\mu}$.

□

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

3.12.2 Non-Compact Case Proofs and Results

Here we prove the results of Section 3.7 which provide conditions on σ, μ, and $\mathcal{H}$ which guarantee that $\mu_i \sim \mathcal{H}_\sharp\mu$ satisfy the conditions of the theorems from Section 3.5.

Proof of Theorem 3.25. Assume that $\widetilde{\mu}$ is the truncated measure approximating $h_\sharp\mu$ for $h \in \mathcal{H}$. Given the assumptions of Lemma 3.20, 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 μ 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 $h_\sharp\mu = \bar{\mu}$ also has the same decay rate up to a constant. For what follows, we must assume that $h \in \mathcal{H}$ has an inverse h^{-1}. If we assume further that $\mathcal{H}$ satisfies Assumption 3.11 (iv) (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,

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

The bi-Lipshitz assumption further implies that

$$A^{-1}I \preceq J_{h^{-1}}(x) \preceq a^{-1}I.$$

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

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

The constants a and A can be absorbed into the other decay rate constants; thus, Assumption 3.11 (iv) gives us the decay rate we want. Noting that the form of the density $f_{\overline{\mu}}$ also implies that $ca^{-n} \leq f_{\overline{\mu}}(x)$ 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 3.20.

- **Uniformly convex support:** If μ is supported on all of $\mathbb{R}^n$, we would want $h \in \mathcal{H}$ 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

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

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 3.11 (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 3.26. From the proof of Theorem 3.25 above, we easily see that if Assumption 3.11 is fulfilled and μ fulfills the conditions of Lemma 3.20 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 3.15. We get $(A1)$ for free since the density $f_{\widetilde{\mu}}$ is lower bounded from the proof of Lemma 3.20. We also get $(A2)$ since $T_\sigma^{\widetilde{\mu}}$ is differentiable from Caffarelli's regularity theorem [23, 24, 25] and if (A3) is satisfied, which comes from [52, 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 ϕ 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 $x \in \operatorname{supp}(\sigma)$, which is compact, we know that $0 < \lambda_{\min}(\sigma) = \min_{x\in\operatorname{supp}(\sigma)} \lambda_{\min}(\nabla^2\phi(x))$, which implies that $J_{T_\sigma^{\widetilde{\mu}}}(x) \succeq \lambda_{\min}(\sigma)I$. This guarantees that $(A3)$ is satisfied for σ and $\widetilde{\mu}$. □

3.13 Acknowledgements

This chapter, in full, is a reprint of *Linearized Wasserstein dimensionality reduction with approximation guarantees* by Cloninger, A., Hamm, K., Khurana, V., and Moosmüller, C. February 14, 2023. as it appears in `https://arxiv.org/abs/2302.07373`. The paper is currently under revision for the Applied Computational Harmonic Analysis (ACHA) Journal. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Lattice-Based Approximations

JOINT WORK WITH KEATON HAMM

We consider structured approximation of measures in Wasserstein space $W_p(\mathbb{R}^d)$ for $p \in [1,\infty)$ by discrete and piecewise constant measures based on a scaled Voronoi partition of $\mathbb{R}^d$. We show that if a full rank lattice Λ is scaled by a factor of $h \in (0,1]$, then approximation of a measure based on the Voronoi partition of $h\Lambda$ is $O(h)$ regardless of d or p. We then use a covering argument to show that N-term approximations of compactly supported measures is $O(N^{-\frac{1}{d}})$ which matches known rates for optimal quantizers and empirical measure approximation in most instances. Finally, we extend these results to noncompactly supported measures with sufficient decay.

4.1 Introduction

This short chapter considers N-term approximations of measures in the Wasserstein distance W_p for $p \in [1,\infty)$. We utilize structured approximations based on a Voronoi partition of $\mathbb{R}^d$ with respect to a lattice, and the approximation rates are governed by a scaling factor applied to the lattice. When translated to N-term approximations of compactly supported measures, we show that these structured approximations match the known rates of approximation for optimal quantizers and empirical measures.

Our structured approximations are motivated by orthographic projection camera models from computer vision [86] in which points are orthogonally projected onto the camera plane. An N-pixel grayscale image is typically considered as a matrix or vector, and is an array of N pixel intensity values. However, in machine learning applications, it has been observed that treating images as vectors in Euclidean space can fail to accurately reflect the structure that appears in them. Many recent works have proposed understanding images as probability measures, for instance by mapping pixel intensities to a uniform grid in $\mathbb{R}^2$ [32, 48, 57, 61, 66, 75, 92]. This viewpoint has been used for manifold learning and supervised classification in these references with success, as Wasserstein distances between images treated as measures are more meaningful than Euclidean distances.

Optimal quantization of measures and empirical measure approximation have been studied in a variety of works [21, 46, 49] and [26, 41, 44, 83, 95], respectively. It is well known that in most instances, without stricter assumptions, both problems yield N-term approximations μ_N of an absolutely continuous measure μ such that $W_p(\mu,\mu_N) = \Theta(N^{-\frac{1}{d}})$. Further assumptions on μ sometimes yields more refined estimates which we discuss in the sequel. We will show that a concrete approximation $\mu_N = \sum_{\lambda\in\Lambda_N}\alpha_\lambda\delta_\lambda$ or $\mu_N = \sum_{\lambda\in\Lambda_N}\beta_\lambda\mathbb{1}_{V_\lambda}$ will match this rate for compactly supported measures, i.e., $W_p(\mu,\mu_N) = O(N^{-\frac{1}{d}})$. Here, Λ_N will be N terms of a full rank lattice $\Lambda\subset\mathbb{R}^d$ and V_λ are the associated Voronoi cells of the lattice. These results are obtained in two stages: first, we consider an approximation on all of $\mathbb{R}^d$ using the scaled lattice $h\Lambda$ whereby we show approximation rate $O(h)$, and then we use a covering number argument to verify the N-term approximation rate for compactly supported measures. We next generalize the approximation rates to non-compactly supported measures with suitable tail decay. Additionally, we provide general rates for nonuniform approximations.

Once these rates are established, we focus on extending the compatibility condition for discrete measures similar to compatibility showcased in Chapter 2. In practice, if we are given a compact measure, the concrete approximations μ_N use of Dirac masses will ensure that our approximated measure will lie on a finite grid of the lattice specified above. So computationally, we end up considering discrete measures on a finite grid. This discrete measure compatibility does not stem from ensuring that isometry holds but rather to show that certain pushforwards on a finite grid (such as a lattice on a torus) can be absorbed into the Sinkhorn solution of the regularized optimal transport problem. The compatibility on the level of discrete measures shows properly that the reference measure used in this discretized linearized optimal transport (dLOT) as well as the pushforwards need to respect the geometry and symmetries of the underlying grid.

4.2 Background

The Wasserstein-p space, denoted $W_p(\mathbb{R}^d)$, is the set of probability measures with finite p-th moment, equipped with the Wasserstein distance

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

where $\mathcal{P}(\mathbb{R}^{2d})$ is the set of all probability measures over $\mathbb{R}^{2d}$ and $\Gamma(\mu,\nu) := \{\gamma \in \mathcal{P}(\mathbb{R}^{2d}) : \gamma(A\times\mathbb{R}^d) = \mu(A),\ \gamma(\mathbb{R}^d\times A) = \nu(A) \text{ for all } A\subset\mathbb{R}^d\}$ is the set of all joint probability measures with marginals μ and ν. We will write $M_p(\mu) = \int_{\mathbb{R}^d} |x|^p d\mu(x)$ for the p-th moment of μ. Given a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$, we denote by $T_\sharp\mu$ the pushforward measure which satisfies $T_\sharp\mu(A) = \mu(T^{-1}(A))$.

There are two main types of approximations considered in the literature for measures in $W_p(\mathbb{R}^d)$: optimal quantizers and empirical measures. The approximation rates are quite similar in both instances.

4.2.1 Optimal Quantization in W_p

Let $Q_N := \{\nu \in W_p(\mathbb{R}^d) : |\text{supp}(\nu)| \leq N\}$ be the set of all discrete measures in W_p supported on at most N points. Then the optimal quantization problem for a given measure $\mu \in W_p$ is to find a solution to

$$\mathcal{E}(\mu, Q_N)_{W_p} := \inf_{\nu \in Q_N} W_p(\mu, \nu).$$

Existence of a minimizing measure is guaranteed for compactly supported μ.

Graf and Luschgy [46, Lemma 3.1] show that

$$\mathcal{E}(\mu, Q_N)_{W_p} = \inf_{\substack{\alpha \subset \mathbb{R}^d \\ \#\alpha \leq N}} \left(\int_{\mathbb{R}^d} \min_i |x - \alpha_i|^p d\mu(x) \right)^{\frac{1}{p}}.$$

That is, finding an optimal quantizer (measure) is equivalent to the problem of approximating μ with N centers in $\mathbb{R}^d$.

Most of the estimates for optimal quantizers are asymptotic estimates. Bucklew and Wise [21] prove that if $\mu \in W_p(\mathbb{R}^d)$ has finite $p + \varepsilon$ moment for some $\varepsilon > 0$, then

$$\mathcal{E}(\mu, Q_N)_{W_p} = O(N^{-\frac{1}{d}}).$$

Interestingly, their analysis shows that the rate above only depends on the absolutely continuous part of μ. Indeed, for any singular μ, $\mathcal{E}(\mu, Q_N)_{W_p} = o(N^{-\frac{1}{d}})$, which is a stronger condition.

Hardin et al. [49] merged approaches of optimal quantization and optimizing Riesz energies, which lead to further results on asymptotics of optimal quantifiers. To the best of our knowledge, optimal quantization results in this vein are all asymptotic estimates.

4.2.2 Empirical measure approximation

Another large segment of the literature considers approximations of the form $\mu_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ where x_i are drawn i.i.d. from μ. The random measure μ_N is typically called the empirical measure of μ. Most works estimate $\mathbb{E}[W_p(\mu,\mu_N)]$, and in contrast to optimal quantization, the bounds hold for all $N \in \mathbb{N}$, but sometimes require restricted assumptions on the measure μ.

Fournier and Guillin [44] show that if $\mu \in W_p(\mathbb{R}^d)$ has finite q-th moment ($M_q(\mu) < \infty$) for some $q > p$, then for all $N \in \mathbb{N}$,

$$\mathbb{E}[W_p(\mu,\mu_N)] \leq CM_q^{\frac{p}{q}}(\mu) \begin{cases} N^{-\frac{1}{2}} + N^{-\frac{q-p}{q}} & p > \frac{d}{2},\ q \neq 2p \\ N^{-\frac{1}{2}} \log(1+N) N^{-\frac{q-p}{q}} & p = \frac{d}{2},\ q \neq 2p \\ N^{-\frac{p}{d}} + N^{-\frac{q-p}{q}} & p \in (0, \frac{d}{2}),\ q \neq \frac{d}{d-p}, \end{cases}$$

for some constant C depending only on p, q, and d. This result generalized those of Dereich et al. [41].

For measures on the d-dimensional torus, [42] showed if μ is absolutely continuous with density bounded above and below (away from ∞ and 0, respectively), then for all $N \in \mathbb{N}$,

$$\mathbb{E}[W_p(\mu,\mu_N)] \leq C \begin{cases} N^{-\frac{1}{d}} & d \geq 3 \\ N^{-\frac{1}{2}}(\log(N))^{\frac{1}{2}} & d = 2 \\ N^{-\frac{1}{2}} & d = 1. \end{cases}$$

If no bounds are assumed on the density, then $\mathbb{E}[W_p(\mu,\mu_N)] \leq C(N^{-\frac{1}{2p}} + N^{-\frac{1}{d}})$, and moreover the bound is tight [83].

Cañas and Rosasco [26] consider $\mu \in W_p(\mathcal{M})$, with $\mathcal{M}$ a compact smooth d-dimensional manifold with bounded curvature and C^1 metric and volume measure $\lambda_{\mathcal{M}}$. They show that if μ has absolutely continuous part with density $f \neq 0$, then $W_p(\mu,\mu_N) = \Omega(N^{-1/d})$ uniformly over

μ_N with constants depending only on d and f. They additionally show a probabilistic bound on rate of convergence of the empirical measure. In particular, for sufficiently large N, and any $\tau > 0$, we have

$$W_2(\mu,\mu_N) \leq C \cdot \left(\int_{\mathcal{M}} \mu_A(x)^{\frac{d}{d+2}} d\lambda_{\mathcal{M}}(x) \right) \cdot N^{-\frac{1}{2d+4}} \cdot \tau, \text{ with probability } 1 - e^{-\tau^2},$$

where C depends only on d.

Weed and Bach [95] prove both asymptotic estimates and finite-sample estimates in the following two scenarios: (m,Δ)-clusterability and approximate low-dimensional support [95, Definitions 7, 8]. A measure μ is (m,Δ)-clusterable if $\text{supp}(\mu)$ lies in the union of m balls of at most radius Δ. Moreover, μ, is approximately low-dimensional if $\text{supp}(\mu) \subseteq S_\varepsilon$ for $S_\varepsilon = \{y : \|y - S\| \leq \varepsilon\}$ where S is low-dimensional. Their main results show that

$$\mathbb{E}[W_p(\mu,\mu_N)] \leq C \begin{cases} (\frac{N}{m})^{-\frac{1}{2p}} & \mu \text{ is } (m,\Delta)\text{-clusterable} \\ N^{-\frac{1}{d}} & \mu \text{ is approximately low-dimensional} \end{cases},$$

where for the (m,Δ)-clusterable case we assume that $N \leq m(2\Delta)^{-2p}$ and for the approximately low-dimensional case we assume that $N \leq (3\varepsilon)^{-d}$.

For 1-dimensional measures, [14, 97] investigate the best uniform approximation of a measure μ by $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ where the x_i's are chosen to minimize $W_p(\mu_N,\mu)$. Similar to optimal quantizer results, they show that

$$\liminf_{N\to+\infty} N W_p(\mu_N,\mu) \geq \frac{1}{2(p+1)^{\frac{1}{p}}} \left(\int_{\mathbb{R}} \frac{\mathbb{1}_{f(x)>0}}{f^{p-1}(x)} dx \right)^{1/p},$$

where f denotes the density of the absolutely continuous part of μ with respect to the Lebesgue measure.

4.2.3 Structured approximations

We focus our attention on structured approximations of measures in Wasserstein space similar to those well known in approximation theory. We consider approximations similar to lattice quantizers in [46, Section 8.3].

A full-rank *lattice* $\Lambda \subset \mathbb{R}^d$ is a discrete subgroup of the additive group $\mathbb{R}^d$ which spans the space. A particular type of fundamental domain for a lattice is its *Voronoi cell*, which is the domain centered at the origin (which is evidently an element of the lattice) consisting of all points which are closer to the origin than any other lattice point. More formally, if Λ is a lattice, then its Voronoi cell is

$$V_0 = \{x \in \mathbb{R}^d : |x| \leq |x-\lambda|, \text{ for all } \lambda \in \Lambda\}.$$

Any element of $\mathbb{R}^d$ may be written as the sum of a point in V_0 and an element of Λ. The Voronoi cell centered at $\lambda \in \Lambda$ is defined by

$$V_\lambda = \{x \in \mathbb{R}^d : |x-\lambda| \leq |x-\lambda'|, \text{ for all } \lambda' \neq \lambda\}$$

and we have $V_\lambda = V_0 + \lambda$. The Voronoi cells tile the space by translation, i.e., $\mathbb{R}^d = \cup_{\lambda\in\Lambda} V_0 + \lambda$. Voronoi cells are convex polytopes, and defined as above, distinct Voronoi cells may intersect on their faces, but their intersection has Lebesgue measure 0. However, for our purposes, since we want to construct approximations to measures which may not be absolutely continuous, we will use the fact that one can remove faces from V_0 in such a way that it still tiles $\mathbb{R}^d$ by translation, but $V_\lambda \cap V_{\lambda'} = \emptyset$ for all $\lambda \neq \lambda'$. Therefore, we will assume that we have a Voronoi cell V_0 and $V_\lambda = V_0 + \lambda$ such that $\mathbb{R}^d = \bigsqcup_{\lambda\in\Lambda} V_\lambda$ (disjoint union).

Our requirement that the Voronoi cells be disjoint is simply avoid issues when singular portions of the measures we are approximating lie on the boundary of any particular Voronoi cell. An alternative method to remedy this problem is to simply approximate the boundary-supported singular measure with an ε-shifted measure. In particular, if μ_d denotes the singular portion of a measure, let $S(\mu_d) = \{x \in \operatorname{supp}(\mu_d) : x \in \partial V_\lambda, \lambda \in \Lambda\}$. Then for each $x \in S(\mu_d)$, define $\Lambda(x) = \{\lambda \in \Lambda : x \in \partial V_\lambda\}$. Now we simply split $\mu_d(x)$ across $\Lambda(x)$ by $\frac{\mu_d(x)}{|\Lambda(x)|}\sum_{\lambda\in\Lambda(x)}\delta_{x+\varepsilon(\lambda-x)}$ for $\varepsilon > 0$. Applying this procedure to all $x \in S(\mu_d)$ results in an updated measure $\widehat{\mu}_d$ which is close in W_p to μ_d. In practice, we would gain an extra ε error in the approximation bounds below by employing this method, but as ε can be taken as small as needed, the characteristic of the bounds remains the same.

We will consider two types of approximations akin to piecewise constant approximation of functions. In particular, we consider approximations of the forms

$$\sum_{\lambda\in\Lambda} \alpha_\lambda \delta_\lambda,$$

$$\sum_{\lambda\in\Lambda} \beta_\lambda \mathbb{1}_{V_\lambda}.$$

These are particular cases of a more general approximation method relying on a given class of functions $\mathcal{F} = \{f_\lambda : \lambda \in \Lambda\}$ giving rise to approximating measures of the form

$$\widetilde{\mu}(A) = \sum_{\lambda\in\Lambda} \int_{A\cap V_\lambda} f_\lambda(x)dx.$$

Common approximation schemes take $\mathcal{F}$ to be piecewise polynomials of a certain degree or shifted radial basis functions, for example.

In what follows, we will consider approximations at a dilated lattice $h\Lambda$ for $h > 0$. We will let $\{V_{h\lambda} : \lambda \in \Lambda\}$ be its disjoint Voronoi cells such that $V_{h\lambda} = V_{h0} + h\lambda = h(V_0 + \lambda)$. We typically write hV_0 for V_{h0} for clarity. Scaling the lattice by $0 < h \leq 1$ corresponds to a finer-scale covering of the space, and allows for more precise approximations of a measure μ. We begin by understanding approximations of μ by measures of the form $\sum_{\lambda\in\Lambda} \alpha_{h\lambda}\delta_{h\lambda}$ on the whole lattice, and then utilize a covering number argument to give N-term approximation bounds for compactly supported measures. We then show how such bounds can be extended to non-compactly supported measures.

4.2.4 Regularized Optimal Transport

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the probability measures with bounded second moment. In particular, $\sigma \in \mathcal{P}_2(\mathbb{R}^d)$ satisfies

$$\int |x|_2^2 d\sigma(x) < \infty.$$

Given a probability measure, we can consider the space $L^2(\mathbb{R}^d, \sigma)$ with norm

$$\|f\|_\sigma^2 = \int |f(x)|_2^2 d\sigma(x).$$

Unless otherwise stated, let ζ denote the Lebesgue measure, then if $\sigma \in \mathcal{P}_2(\mathbb{R}^d)$ is absolutely continuous with respect to ζ, denoted as $\sigma \ll \zeta$, then there exists a density $f_\sigma : \mathbb{R}^d \to \mathbb{R}$ such that

$$\sigma(A) = \int_A f_\sigma(x) d\zeta(x), \quad A \subseteq \mathbb{R}^d \text{ measurable.}$$

For the most part, we will be restricting our research to the case of probability measures that are absolutely continuous with respect to λ. Given an elementary transformation $S : \mathbb{R}^d \to \mathbb{R}^d$ and a measure σ, we can define the push forward measure by

$$S_\sharp \sigma(A) = \sigma(S^{-1}(A)) \tag{4.1}$$

where $A \subset \mathbb{R}^d$. If $\sigma \ll \zeta$, then in terms of densities, the pushforward relation $\nu(A) = \sigma(S^{-1}(A))$ is given by

$$\int_{S^{-1}(A)} f_\sigma(x) d\zeta(x) = \int_A f_\nu(y) d\zeta(y), \;\; A \subseteq \mathbb{R}^d \text{ measurable.}$$

Given a source distribution μ and a target distribution ν, the optimal transport problem aims at minimizing the cost to “move” μ into ν:

$$\min_{T \in \mathcal{T}_\mu^\nu} \int_{\mathbb{R}^d} \|T(x) - x\|^2 d\mu(x), \tag{4.2}$$

where $\mathcal{T}_\mu^\nu$ is the collection of all measure preserving maps from μ to ν. The minimum of (4.2) exists and is unique subject to certain regularity assumption on μ and ν [17, 91]. In particular, the following theorem explains the regularity conditions needed on the measures.

Theorem 4.1 ((Brenier))**.** *Let $\sigma, \nu \in \mathcal{P}_2(\mathbb{R}^n)$. If $\sigma \ll \lambda$, then there exists a unique map $T_\sigma^\nu \in L^2(\mathbb{R}^n, \sigma)$ that pushes σ to ν and achieves the 2-Wasserstein distance. Furthermore, the map T_σ^ν is uniquely defined as the gradient of a convex function φ so that $T_\sigma^\nu(x) = \nabla\varphi(x)$, where φ is the unique (up to an additive constant) convex function such that $(\nabla\varphi)_\sharp \sigma = \nu$.*

Note that (4.2) is the transport map version of the W_p-distance (with $p = 2$) defined earlier and gives rise to a natural distance between distributions, the Wasserstein-2 distance $W_2(\mu, \nu)^2$. The $\arg\min$ of (4.2) is referred to as the “optimal transport map” and we denote it by T_μ^ν. The optimization problem (4.2) can be formulated for different cost functions and on geometric or manifold domains [67, 5].

For practical purposes, (4.2) needs to be formulated for discrete measures on finite domains. Denote the domain by points $x_i \in \mathbb{R}^d, i = 1, \dots, n$, then a discrete measure on this domain is a vector $a \in \mathbb{R}_+^n$ such that $a^T \mathbb{1} = 1$ ($\mathbb{1}$ denotes the vector containing only ones). We denote the set of all such discrete measures by Σ_n.

The optimal transportation problem in terms of maps T_μ^ν (4.2) has proven to be too restrictive. Often an optimal map does not exist; for example, consider a discrete problem where a Dirac measure δ_{x_i} should be transformed into $0.5\,\delta_{x_j} + 0.5\,\delta_{x_k}$. To overcome this problem, the notion of “mass splitting” has been introduced by Kantorovich [55]. This relaxation looks for optimal couplings instead of optimal maps. A discrete coupling is a matrix $P \in \mathbb{R}_+^{n \times n}$, where P_{ij} describes how much mass flows from x_i to x_j. In this set up, the optimal transport formulation now reads: For a source measure a, and target measure b, find $P \in \mathbb{R}_+^{n\times n}$ that minimizes

$$\min_{P\in\Pi_a^b} \operatorname{tr}(C^T P) \tag{4.3}$$

where $\Pi_a^b = \{P : P\mathbb{1} = a \text{ and } P^T\mathbb{1} = b\}$. Here C denotes the cost matrix; in analogy to (4.2) we use $C_{ij} = |x_i - x_j|^2$. The minimum is the Wasserstein-2 distance, again denoted by $W_2(a,b)^2$. The problem (4.3) is a linear program, and thus the minimum might not be unique.

One of the main drawbacks for application purposes is the computational cost of computing (4.3). Significant computational speed-up (from $O(n^3\log(n))$ to $O(n^2\log(n)))$ for (4.3) can be achieved by adding a constraint on the entropy of P, which adds a regularization term to (4.3) [36, 4]. The regularized version of (4.3) is

$$\min_{P\in\Pi_a^b} \operatorname{tr}(C^T P) - \beta h(P), \tag{4.4}$$

where $h(P) = -\sum_{i,j=1}^n p_{ij}(\log p_{ij})$ and $\beta > 0$ is the regularizer. The optimal coupling of (4.4), which we denote by $P_{a,\beta}^b(C)$, is unique and has the form

$$\operatorname{diag}(u)e^{-\beta C}\operatorname{diag}(v)$$

where $e^{-\beta C}$ denotes the Hadamard (entrywise) exponential and $u,v \in \mathbb{R}^d$ with $u,v > 0$. The minimum is the Sinkhorn distance, denoted by $W_{2,\beta}(a,b)$. As $\beta \to 0$, $P_{a,\beta}^b$ converges to the optimal solution of (4.3) with maximal entropy [36, 79]. Therefore, also $W_{2,\beta}(a,b) \to W_2(a,b)$ as $\beta \to 0$. The optimal coupling $P_{a,\beta}^b(C)$ of the regularized problem can be easily computed via Sinkhorn-Knopp's fixed point iteration [36, 84].

4.2.5 Linearized Optimal Transport

Although Linearized Optimal Transport (LOT) is introduced and talked about in Chapter 2, we reintroduce here as a refresher. LOT, introduced by [94, 78, 45], is a method to embed $\mathcal{P}_2(\mathbb{R}^d)$ into an L^2 space in a very natural manner. In particular, if we fix a reference measure σ, the LOT embedding, denoted F_σ, is

$$F_\sigma : \begin{cases} \mathcal{P}_2(\mathbb{R}^d) \to L^2(\mathbb{R}^d, \sigma) \\ \mu \to T_\sigma^\mu \end{cases} .$$

The so-called *compatibility condition* [71, 1] describes when LOT and the pushforward commute.

Definition 4.2. *Fix $\sigma, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\sigma \ll \zeta$. For a function $S \in L^2(\mathbb{R}^d, \mu)$, we say that (σ, μ, S) form a* compatible-triple *if*

$$F_\sigma(S_\sharp \mu) = S \circ F_\sigma(\mu).$$

Note that the compatibility condition of Definition 4.2 can also be written as

$$T_\sigma^{S_\sharp \mu} = S \circ T_\sigma^\mu.$$

From past results in Chapter 2, the LOT embedding between $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ is an isometry when $\mu_1 = (S_1)_\sharp \mu$, $\mu_2 = (S_2)_\sharp \mu$ and (σ, μ, S_j) form a compatible-triple for $j = 1, 2$.

4.2.6 Outline

We begin by considering some general lemmas that we will use throughout, then focus on Dirac train approximations for compactly supported measures in Section 4.4 and piecewise constant approximations of the second form above in Section 4.5. Section 4.6 extends these results to non-compactly supported measures. In Section 4.7, we generalize our approximation methods to nonuniform meshes by using the mesh norm and minimum separation radius as a stand-in. Section 4.8 showcases how to extend the idea of pushforwards to discrete measures as well as regularized compatibility for discrete measures.

4.3 Lemmas

A set $K \subset \mathbb{R}^d$ is a convex body if it is convex and compact, and a convex body is (centrally) symmetric if $K = -K$. The Voronoi cell V_0 of a lattice Λ is a symmetric convex body, as is the Euclidean ball of any radius. Below, denote by $B_R = B(0,R)$ the ball of radius R (in the Euclidean metric) in $\mathbb{R}^d$. Given two convex bodies K and T, let $\mathcal{N}(K,T)$ be the minimal number of translates of T it takes to cover K, i.e., $\mathcal{N}(K,T) = \min\{\#(\alpha_i) \subset \mathbb{R}^d : K \subset \bigcup_i T + \alpha_i\}$.

The diameter of a convex set $A \subset \mathbb{R}^d$ is given by $\operatorname{diam}(A) = \sup\{|x-y| : s,y \in A\}$. The radius of a centrally symmetric convex body K is $\operatorname{rad}(K) = \sup\{|x| : x \in K\}$. Our estimates below will involve both $\operatorname{diam}(V_0)$ and $\operatorname{rad}(V_0)$ for the central Voronoi cell of a lattice Λ.

Lemma 4.3. *Let V_0 be the Voronoi cell centered at 0 of a full-rank lattice $\Lambda \subset \mathbb{R}^d$. Then for $h \in (0,1]$,*

$$\mathcal{N}(B_R, hV_0) \leq 3^d h^{-d} \mathcal{N}(B_R, V_0).$$

Proof. We will use the following fact about covering numbers [11, Theorem 4.1.13 and Corollary 4.1.14]: if $T \subset K \subset \mathbb{R}^d$ are convex bodies and T is symmetric ($T = -T$), then for all $h > 0$,

$$\mathcal{N}(K, hT) \leq (1+2h^{-1})^d \mathcal{N}(K,T).$$

Consequently, for $0 < h \leq 1$,

$$\mathcal{N}(B_R, hV_0) \leq (1+2h^{-1})^d \mathcal{N}(B_R, V_0) \leq 3^d h^{-d} \mathcal{N}(B_R, V_0).$$

□

We have not optimized the estimate in Lemma 4.3 because it is not required for our subsequent analysis, so it is possible that the bound therein could be improved.

Lemma 4.4. *Let $\mu \in W_p(\mathbb{R}^d)$, $p \in [1,\infty)$. Let Λ be a full rank lattice on $\mathbb{R}^d$ with Voronoi cells $\{V_\lambda\}_{\lambda\in\Lambda}$, and let $h \in (0,1]$. Then the following hold:*

(i) $\sum\limits_{\lambda\in\Lambda} |h\lambda|^p \mu(V_{h\lambda}) \leq 2^{p-1}h^p \operatorname{rad}(V_0)^p + 2^{p-1}M_p(\mu)$,

(ii) $\sum\limits_{\lambda\in\Lambda} \|x\|^p_{L_\infty(V_{h\lambda})}\mu(V_{h\lambda}) \leq 2^{p-1}\sum\limits_{\lambda\in\Lambda} |h\lambda|^p \mu(V_{h\lambda}) + 2^{p-1}h^p \operatorname{rad}(V_0)^p$,

(iii) $\sum\limits_{\lambda\in\Lambda} \|x\|^p_{L_\infty(V_{h\lambda})}\mu(V_{h\lambda}) \leq (2^{2p-2}+2^{p-1})h^p \operatorname{rad}(V_0)^p + 2^{p-1}M_p(\mu)$.

Proof. Proof of (i): Note that for any $\lambda \in \Lambda$ and any $x \in V_h\lambda$, we have

$$|h\lambda| \leq |x| + |x-h\lambda| \leq |x| + \operatorname{rad}(V_{h\lambda}) + h\operatorname{rad}(V_0).$$

Therefore, $|h\lambda|^p \leq 2^{p-1}(|x|^p + h^p \operatorname{rad}(V_0)^p)$. Integrating this inequality over $V_{h\lambda}$ with respect to μ and summing over i gives

$$\begin{aligned}\sum_{\lambda\in\Lambda} |h\lambda|^p \mu(V_{h\lambda}) &\leq 2^{p-1}\sum_{\lambda\in\Lambda}\int_{V_{h\lambda}} |x|^p d\mu + 2^{p-1}h^p \operatorname{rad}(V_0)^p \sum_{\lambda\in\Lambda}\mu(V_{h\lambda}) \\ &= 2^{p-1}M_p(\mu) + 2^{p-1}h^p \operatorname{rad}(V_0)^p.\end{aligned}$$

Proof of (ii): We estimate

$$\begin{aligned}\sum_{\lambda\in\Lambda} \|x\|^p_{L_\infty(V_{h\lambda})}\mu(V_{h\lambda}) &\leq \sum_{\lambda\in\Lambda} ||h\lambda| + \mathrm{rad}(V_{h\lambda})|^p \mu(V_{h\lambda}) \\ &\leq 2^{p-1}\left(\sum_{\lambda\in\Lambda} |h\lambda|^p\mu(V_{h\lambda}) + \sum_{\lambda\in\Lambda} \mathrm{rad}(V_{h\lambda})^p\mu(V_{h\lambda})\right) \\ &= 2^{p-1}\left(\sum_{\lambda\in\Lambda} |h\lambda|^p\mu(V_{h\lambda}) + h^p\,\mathrm{rad}(V_0)^p \sum_{\lambda\in\Lambda} \mu(V_{h\lambda})\right),\end{aligned}$$

which yields the desired conclusion.

Proof of (iii): Combine (i) and (ii). □

4.4 Dirac train approximations

First, we consider approximating an arbitrary measure $\mu \in W_p(\mathbb{R}^d)$ by a Dirac approximation as follows. Let $h \in (0,1]$ and we utilize the scaled lattice $h\Lambda = \{h\lambda : \lambda \in \Lambda\}$. Then we approximate μ by

$$\mu_h := \sum_{\lambda\in\Lambda} \alpha_{h\lambda}\delta_{h\lambda}. \tag{4.5}$$

In other words, we utilize a discrete measure with Dirac masses at the scaled lattice $h\Lambda$. Our goal is to determine the approximation rate (in terms of h) of μ via μ_h.

The first requirement is that μ_h is a probability measure, which requires that

$$\mu_h(\mathbb{R}^d) = \sum_{\lambda\in\Lambda} \alpha_{h\lambda} = 1.$$

A natural candidate would therefore be to take $\alpha_{h\lambda}$ to be the measure on the Voronoi region $V_{h\lambda}$, i.e.,

$$\alpha_{h\lambda} = \int_{V_{h\lambda}} d\mu(x) = \mu(V_{h\lambda}).$$

Indeed, we show that this choice works and provides approximation rates that match what one expects to be optimal in terms of the lattice spacing.

Theorem 4.5. *Let $\mu \in W_p(\mathbb{R}^d)$, $p \in [1,\infty)$ be fixed but arbitrary. Let $\mu_h := \sum_{\lambda\in\Lambda} \mu(V_{h\lambda})\delta_{h\lambda}$. Then for all $h \in (0,1]$, $\mu_h \in W_p(\mathbb{R}^d)$ and*

$$W_p(\mu,\mu_h) \leq \operatorname{rad}(V_0)h.$$

Proof. First, note that μ_h is clearly a measure, and

$$\mu_h(\mathbb{R}^d) = \sum_{\lambda\in\Lambda} \mu(V_{h\lambda}) = \mu\left(\bigcup_{\lambda\in\Lambda} V_{h\lambda}\right) = \mu(\mathbb{R}^d) = 1,$$

where the second equality comes from countable additivity of μ and the third equality from the fact that $\mathbb{R}^d = \bigsqcup_{\lambda\in\Lambda} V_{h\lambda}$. Hence μ_h is a probability measure. To show that μ_h has finite p-th moment, we notice that (via Tonelli's Theorem) and Lemma 4.4(i),

$$\int_{\mathbb{R}^d} |x|^p d\mu_h = \sum_{\lambda\in\Lambda} \int_{V_{h\lambda}} |h\lambda|^p d\mu \leq 2^{p-1}h^p \operatorname{rad}(V_0)^p + 2^{p-1}M_p(\mu) < \infty.$$

Using the Kantorovich formulation of W_p, we define a (non-optimal) coupling between μ and μ_h via

$$\widetilde{\pi}(A,B) = \sum_{\lambda\in\Lambda} \mu(B\cap V_{h\lambda})\delta_{h\lambda}(A) = \int_{A\times B} \sum_{\lambda\in\Lambda} \mathbb{1}_{V_{h\lambda}}(y)\, \delta_{h\lambda}(x) dx d\mu(y).$$

It is straightforward to check that $\widetilde{\pi}$ is a measure on $\mathbb{R}^d \times \mathbb{R}^d$. Noting that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} d\widetilde{\pi}(x,y) = \sum_{\lambda\in\Lambda} \mu(V_{h\lambda}) = \mu(\mathbb{R}^d) = 1,$$

we see that $\widetilde{\pi}$ is a probability measure. Computing the marginals, we see

$$\widetilde{\pi}(\mathbb{R}^d, B) = \sum_{\lambda\in\Lambda} \mu(B\cap V_{h\lambda}) = \mu(B)$$

and

$$\widetilde{\pi}(A, \mathbb{R}^d) = \sum_{\lambda\in\Lambda} \mu(V_{h\lambda})\delta_{h\lambda}(A) = \mu_h(A),$$

for all Borel measurable sets $A, B \in \mathbb{R}^d$. Therefore, $\widetilde{\pi}$ is a coupling of μ and μ_h.

Notice that

$$\widetilde{\pi}(V_{h\lambda}, V_{h\lambda'}) = \sum_{\bar{\lambda}\in\Lambda} \mu(V_{h\lambda}\cap V_{h\bar{\lambda}})\delta_{h\bar{\lambda}}(V_{h\lambda'}) = 0,$$

so that $\widetilde{\pi}$ only evaluates mass on sets of the form $V_{h\lambda}\times V_{h\lambda}$. Thus, we have

$$\begin{aligned}
W_p(\mu,\mu_h)^p &\leq \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p d\widetilde{\pi}(x,y) \\
&= \sum_{\lambda\in\Lambda} \int_{V_{h\lambda}\times V_{h\lambda}} |x-y|^p d\tilde{\pi}(x,y) \\
&= \sum_{\lambda\in\Lambda} \int_{V_{h\lambda}} |h\lambda - y|^p d\mu(y) \\
&\leq \sum_{\lambda\in\Lambda} \operatorname{rad}(V_{h\lambda})^p \mu(V_{h\lambda}) \\
&= h^p \operatorname{rad}(V_0)^p \sum_{\lambda\in\Lambda} \mu(V_{h\lambda}) \\
&= h^p \operatorname{rad}(V_0)^p.
\end{aligned}$$

The inequality arises by noting that $|h\lambda - y| \leq \operatorname{rad}(V_{h\lambda})$ for $y\in V_{h\lambda}$, while the subsequent equality is due to the fact that $\operatorname{rad}(V_{h\lambda}) = h\operatorname{rad}(V_\lambda)$ for all λ. The conclusion follows by taking the p-th root on both sides of the above expression. □

To provide N-term approximation rates, we assume that μ is supported on a compact set which is a subset of the interior of some ball B_R. We state bounds in terms of the covering number of the support of μ with Voronoi regions.

Theorem 4.6. *Let $\mu \in W_p(\mathbb{R}^d)$, $p \in [1,\infty)$ have compact support contained in the interior of B_R. Let $N \in \mathbb{N}$ be fixed, $\mathcal{N} = \mathcal{N}(B_R, V_0)$, and set $h = 3\left(\frac{\mathcal{N}}{N}\right)^{\frac{1}{d}}$. Then $\mu_h = \sum_{\lambda\in\Lambda}\mu(V_{h\lambda})\delta_{h\lambda}$ is an at most N-term approximation of μ, which satisfies*

$$W_p(\mu,\mu_h) \leq 3\,\mathrm{rad}(V_0)\mathcal{N}^{\frac{1}{d}}N^{-\frac{1}{d}}.$$

Proof. Note that for any $h \in (0,1]$, μ_h has at most

$$\mathcal{N}(B_R, hV_0) \leq 3^d h^{-d}\mathcal{N}(B_R, V_0) = 3^d h^{-d}\mathcal{N}$$

terms (the first inequality is Lemma 4.3), whereupon setting h as in the statement of the theorem and applying Theorem 4.5 yields the conclusion. □

From Theorem 4.6, we may deduce the following result which is related to a simple camera model from computer vision. The orthographic projection camera model projects objects orthogonally onto the camera plane [86]. This model is a simplification, as it does not accurately treat perspective of the object being imaged, but it is relatively accurate for objects being imaged from a distance. If we take a scaled integer lattice in $\mathbb{R}^2$ with N elements, then the shifted Voronoi regions which are cubes of side length $N^{-\frac{1}{2}}$ and a discretization of the form (4.5) corresponds to an orthographic projection model of an N-pixel camera in which pixel intensity values are the average of the intensity over the given region of the imaging window.

Corollary 4.7. *Let $\mu \in W_p(\mathbb{R}^d)$ have support in the interior of $[-\frac{1}{2},\frac{1}{2}]^d$, and let $\Lambda_N = (N^{-\frac{1}{2}}\mathbb{Z})^d$ with Voronoi cells $(V_\lambda)_{\lambda\in\Lambda_N}$. Then $\mu_N = \sum_{\lambda\in\Lambda_N}\mu(V_\lambda)\delta_\lambda$ is an at most N-term approximation to μ which satisfies*

$$W_p(\mu,\mu_N) \leq \frac{\sqrt{d}}{2}N^{-\frac{1}{d}}.$$

Proof. Setting $h = N^{-\frac{1}{d}}$ yields N cubes of length $N^{-\frac{1}{d}}$ in the unit cube, so appealing to the proof of Theorem 4.5 yields the desired rate upon noticing that $\operatorname{rad}([-\frac{1}{2},\frac{1}{2}]^d) = \frac{\sqrt{d}}{2}$ and $\mathcal{N}([-\frac{1}{2},\frac{1}{2}]^d, [-\frac{1}{2},\frac{1}{2}]^d) = 1$. Note that in this case, the bound of Theorem 4.6 yields an overestimate as it assumes support in a unit ball instead of a cube. □

4.5 Piecewise Constant Approximation

In this section we consider approximating a measure $\nu \in W_p(\mathbb{R}^d)$ by a piecewise constant approximation of the form

$$\nu_h = \sum_{\lambda \in \Lambda} \beta_{h\lambda} \mathbb{1}_{V_{h\lambda}}. \tag{4.6}$$

For ν_h to be a probability measure, the following is required:

$$\sum_{\lambda \in \Lambda} \beta_{h\lambda} \int_{V_{h\lambda}} dx = \sum_{\lambda \in \Lambda} \beta_{h\lambda} |V_{h\lambda}| = 1,$$

hence a natural choice is

$$\beta_{h\lambda} = \frac{\nu(V_{h\lambda})}{|V_{h\lambda}|},$$

which corresponds to a piecewise constant approximation of ν where each Voronoi region is assigned the value of the ratio of the mass that ν assigns to the region to its Lebesgue measure.

Theorem 4.8. *Let $\nu \in W_p(\mathbb{R}^d)$ be fixed but arbitrary. Let $\nu_h := \sum_{\lambda \in \Lambda} \frac{\nu(V_{h\lambda})}{|V_{h\lambda}|} \mathbb{1}_{V_{h\lambda}}$. Then for all $h \in (0,1]$, $\nu_h \in W_p(\mathbb{R}^d)$ and*

$$W_p(\nu, \nu_h) \leq \operatorname{diam}(V_0) h.$$

Proof. The analysis above shows that $\nu_h \in \mathcal{P}(\mathbb{R}^d)$, so it remains to show that it has finite p-th moment, which can be seen as follows:

$$\sum_{\lambda\in\Lambda}\frac{\nu(V_{h\lambda})}{|V_{h\lambda}|}\int_{V_{h\lambda}}|x|^p dx \le \sum_{\lambda\in\Lambda}\frac{\nu(V_{h\lambda})}{|V_{h\lambda}|}\|x\|^p_{L_\infty(V_{h\lambda})}|V_{h\lambda}|$$
$$\le 2^{p-1}M_p(\mu)+(2^{2p-2}+2^{p-1})h^p\,\mathrm{rad}(V_0)^p<\infty,$$

where we have used Lemma 4.4(iii) (here and throughout this proof, Tonelli's Theorem justifies the interchange of sum and integral). Therefore $\nu_h\in W_p(\mathbb{R}^d)$.

To estimate the convergence rate, we again form a (non-optimal) coupling in the Kantorovich sense, as follows: for Borel measurable $A,B\subset\mathbb{R}^d$, define

$$\widetilde{\pi}(A,B):=\sum_{\lambda\in\Lambda}\frac{\nu(A\cap V_{h\lambda})}{|V_{h\lambda}|}|B\cap V_{h\lambda}|=\int_{A\times B}\sum_{\lambda\in\Lambda}\frac{1}{|V_{h\lambda}|}\mathbb{1}_{V_{h\lambda}}(x)\mathbb{1}_{V_{h\lambda}}(y)d\nu(x)dy.$$

It is straightforward to check that $\widetilde{\pi}$ is a measure on $\mathbb{R}^d\times\mathbb{R}^d$. To see it is a probability measure, note that

$$\widetilde{\pi}(\mathbb{R}^d,\mathbb{R}^d)=\sum_{\lambda\in\Lambda}\frac{\nu(V_{h\lambda})}{|V_{h\lambda}|}|V_{h\lambda}|=\sum_{\lambda\in\Lambda}\nu(V_{h\lambda})=\nu(\mathbb{R}^d)=1.$$

Additionally, the marginals can be computed as follows:

$$\widetilde{\pi}(A,\mathbb{R}^d)=\sum_{\lambda\in\Lambda}\frac{\nu(A\cap V_{h\lambda})}{|V_{h\lambda}|}|V_{h\lambda}|=\nu(A)$$

as before, and

$$\widetilde{\pi}(\mathbb{R}^d,B)=\sum_{\lambda\in\Lambda}\frac{\nu(V_{h\lambda})}{|V_{h\lambda}|}|B\cap V_{h\lambda}|=\sum_{\lambda\in\Lambda}\frac{\nu(V_{h\lambda})}{|V_{h\lambda}|}\int_B\mathbb{1}_{V_{h\lambda}}(x)dx=\nu_h(B).$$

Therefore, $\widetilde{\pi}$ is a coupling of ν and ν_h.

Thus,

$$\begin{aligned}
W_p(\nu,\nu_h)^p &\leq \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p d\widetilde{\pi}(x,y) \\
&= \sum_{\lambda\in\Lambda} \frac{1}{|V_{h\lambda}|}\int_{V_{h\lambda}\times V_{h\lambda}} |x-y|^p d\nu(x)dy \\
&\leq \sum_{\lambda\in\Lambda} \frac{\operatorname{diam}(V_{h\lambda})^p}{|V_{h\lambda}|}\nu(V_{h\lambda})|V_{h\lambda}| \\
&= h^p \operatorname{diam}(V_0)^p \sum_{\lambda\in\Lambda} \nu(V_{h\lambda}) \\
&= h^p \operatorname{diam}(V_0)^p,
\end{aligned}$$

and the conclusion follows. □

Corollary 4.9. *Let* $\nu \in W_p(\mathbb{R}^d)$ *have compact support contained in the interior of* B_R. *Let* $N \in \mathbb{N}$ *be fixed,* $\mathcal{N} = \mathcal{N}(V_0, B_R)$, *and set* $h = 3\left(\frac{\mathcal{N}}{N}\right)^{\frac{1}{d}}$. *Then* $\nu_h = \sum_{\lambda\in\Lambda} \frac{\nu(V_{h\lambda})}{|V_{h\lambda}|}\mathbb{1}_{h\lambda}$ *is an N-term approximation of* ν, *which satisfies*

$$W_p(\nu,\nu_h) \leq 3\operatorname{diam}(V_0)\mathcal{N}^{\frac{1}{d}}N^{-\frac{1}{d}}.$$

Proof. Mimic the proof of Theorem 4.6 *mutatis mudandis* applying Theorem 4.8. □

Corollary 4.10. *Let* $\nu \in W_p(\mathbb{R}^d)$ *have compact support contained in the interior of* $[-\frac{1}{2},\frac{1}{2}]^d$, *and let* $\Lambda_N = (N^{-\frac{1}{d}}\mathbb{Z})^d$ *with Voronoi cells* $(V_\lambda)_{\lambda\in\Lambda_N}$. *Then* $\nu_N = \sum_{\lambda\in\Lambda_N} \frac{\nu(V_\lambda)}{|V_{h\lambda}|}\mathbb{1}_\lambda$ *is an N-term approximation to* ν *which satisfies*

$$W_p(\nu,\nu_N) \leq \sqrt{d}N^{-\frac{1}{d}}.$$

Proof. Set $h = N^{-\frac{1}{d}}$ while noting that $\mathcal{N}([-\frac{1}{2},\frac{1}{2}]^d,[-\frac{1}{2},\frac{1}{2}]^d) = 1$ and $\operatorname{diam}([-\frac{1}{2},\frac{1}{2}]^d) = \sqrt{d}$. □

While Corollary 4.7 corresponds to mapping pixel intensity values from an orthographic camera image to a discrete grid in $\mathbb{R}^2$ (or more generally $\mathbb{R}^d$), Corollary 4.10 corresponds to a voxel representation of the image in which each pixel intensity value takes up the whole cube rather than just a single point in the center.

4.6 Non-Compactly Supported Measures

So far, we have assumed that μ is compactly supported in a ball $B_R \subset \mathbb{R}^d$, but here we extend the results above to non-compactly supported measure with suitable decay. If the measure μ decays fast enough away outside of a ball B_R, we first estimate μ with a compactly supported measure $\widehat{\mu}$, and then apply our approximation schemes above to $\widehat{\mu}$.

We want to create a non-optimal coupling that will send μ to itself when restricted to sets inside B_R but that will project the part of μ outside of B_R to the boundary of the ball. To do this, we define the projection operator

$$P_{B_R}(x) = \underset{y \in B_R}{\operatorname{argmin}} \|x - y\|.$$

In particular, given any set $B \subseteq \mathbb{R}^d$, this projection operator has a preimage $P_{B_R}^{-1}(B) = \{x \in \mathbb{R}^d : P_{B_R}(x) \in B\}$. If $B \cap B_R = \emptyset$, then, $P_{B_R}^{-1}(B) = \emptyset$. Finally, notice that

$$P_{B_R}^{-1}(B) = P_{B_R}^{-1}(B \cap B_R) \cup P_{B_R}^{-1}(B \cap \partial B_R)$$

We use this definition in our construction to define the following coupling:

$$\pi = \left(I \times P_{B_R}\right)_{\sharp}\mu = \mu \times (P_{B_R})_{\sharp}\mu =: \mu \times \widehat{\mu}.$$

Notice first that this coupling sends μ to itself when restricted to sets in B_R. Secondly, for $A \subseteq B_R^c$, it projects the measure of A to the boundary ∂B_R. In essence, this acts as approximation through a truncated measure supported on the ball B_R. In particular, the measure $\pi(\mathbb{R}^d, B) = \widehat{\mu}(B)$ is supported entirely on the ball B_R.

Recalling the Lebesgue decomposition theorem, we know that $\mu = \mu_< + \mu_\perp + \mu_d$ where $\mu_<$ is absolutely continuous with respect to the Lebesgue measure (and has a density f_μ), $\mu_\perp$ is the singular continuous measure such that $\mu_\perp\{x\} = 0$ for $x \in \mathbb{R}^d$, and μ_d purely atomic discrete measure such that $\mu_d = \sum_{i=1}^{\infty} c_i \delta_{x_i}$. We will show that $W_p(\mu, \widehat{\mu}) < \varepsilon$ if we assume some decay conditions on $\mu_<$, $\mu_\perp$, and μ_d. Apart from natural decay conditions on $\mu_<$ and μ_d, if $\mu_\perp$ decays on concentric shells $B_{(a,b]} = B_b \setminus B_a$ for $a > R$, then we get the approximation result. These ideas are laid out in the following theorem:

Theorem 4.11. *Let μ have refined Lebesgue decomposition, $\mu = \mu_< + \mu_\perp + \mu_d$, where $\mu_<$ has density f_μ, and let $\varepsilon > 0$. Assume that*

1. *$f_\mu(x) \leq \frac{\varepsilon^p}{3C|x|^{p+d+1}}$ where C is the integration constant from integrating over concentric d-spheres,*

2. *For every $j \geq \lfloor R \rfloor$, we have*

$$\mu_\perp\left(B_{\left(j,j+1\right]}\right) \leq \frac{\varepsilon^p}{3\left(j+1-R\right)^{p+2}} \frac{6}{\pi^2},$$

3. *the c_k's in $\mu_d = \sum_{k=1}^{\infty} c_k \delta_{x_k}$ decay like*

$$c_k \leq \frac{1}{(|x_k| - R)^p} \cdot \frac{1}{k^q} \cdot \frac{\varepsilon^p}{3} \cdot \frac{1}{\sum_{\ell=1}^{\infty} \ell^q}$$

for some $q > 1$ and $|x_k| > R$.

Then $W_p(\mu, \widehat{\mu}) < \varepsilon$, where $\widehat{\mu}$ is the compactly supported measure $(P_{B_R})_\sharp \mu$.

Proof. Notice that

$$\begin{aligned}
W_p(\mu,\widehat{\mu})^p &\leq \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-P_{B_R}(x)|^p d\mu(x) \\
&= \underbrace{\int_{B_R} |x-P_{B_R}(x)|^p d\mu(x)}_{0} + \int_{B_R^c} |x-P_{B_R}(x)|^p d\mu(x) \\
&= \int_{B_R^c} \left| x - R\frac{x}{|x|} \right|^p d\mu(x) = \int_{B_R^c} \left(1-\frac{R}{|x|}\right)^p |x|^p d\mu(x) \\
&= \int_{B_R^c} \big(|x|-R\big)^p d\mu(x).
\end{aligned}$$

This means that

$$\begin{aligned}
W_p(\mu,\widehat{\mu})^p \leq& \int_{B_R^c} \big(|x|-R\big)^p d(\mu_< + \mu_\perp + \mu_d)(x) \\
=& \underbrace{\int_{B_R^c} \big(|x|-R\big)^p f_\mu(x) dx}_{I_1} + \underbrace{\int_{B_R^c} \big(|x|-R\big)^p d\mu_\perp(x)}_{I_2} \\
&+ \underbrace{\int_{B_R^c} \big(|x|-R\big)^p d\mu_d(x)}_{I_3}.
\end{aligned}$$

In particular, we need that both $I_1, I_2, I_3 < \frac{\varepsilon^p}{3}$. For I_1, this is ensured if

$$f_\mu(x) \leq \frac{\varepsilon^p}{3C|x|^{p+d+1}},$$

where C is the integration constant from integrating over concentric d-spheres. To see this, notice that

$$I_1 = \int_{B_R^c} \big(|x|-R\big)^p f_\mu(x) dx \leq \frac{\varepsilon^p}{3C} \int_{B_R^c} \underbrace{\frac{\big(|x|-R\big)^p}{|x|^p}}_{<1} \frac{1}{|x|^{d+1}} dx < \frac{\varepsilon^p}{3} \underbrace{\int_{r\geq R} \frac{r^{d-1}}{r^{d+1}} dr}_{\leq 1} \leq \frac{\varepsilon^p}{3}.$$

To bound I_2, we assumed decay rates on the measure of concentric annuli emanating out from B_R. In particular, we assume that for $j \geq \lfloor R \rfloor$, we have

$$\mu_\perp\left(B_{\left(j,j+1\right]}\right) \leq \frac{\varepsilon^p}{3\left(j+1-R\right)^{p+2}}\,\frac{6}{\pi^2}.$$

This implies that

$$\begin{aligned}
\int_{B_R^c}\left(|x|-R\right)^p d\mu_\perp(x) &= \int_{B_{[R,\lceil R\rceil)}} \underbrace{(|x|-R)^p}_{\leq(\lceil R\rceil-R)^p}\mu_\perp(x) + \sum_{j>R}^{\infty}\int_{B_{[j,j+1)}} \underbrace{(|x|-R)^p}_{(j+1-R)^p}\mu_\perp(x) \\
&\leq \int_{B_{[R,\lceil R\rceil)}} (\lceil R\rceil - R)^p\mu_\perp(x) + \sum_{j>R}^{\infty}\int_{B_{[j,j+1)}} (j+1-R)^p\mu_\perp(x) \\
&\leq \sum_{j\geq\lfloor R\rfloor} (j+1-R)^p\mu_\perp\left(B_{\left(j,j+1\right]}\right) \\
&\leq \sum_{j\geq\lfloor R\rfloor} (j+1-R)^p\frac{\varepsilon^p}{3\left(j+1-R\right)^{p+2}}\,\frac{6}{\pi^2} \\
&\leq \frac{\varepsilon^p}{3}\frac{6}{\pi^2}\underbrace{\sum_{j\geq\lfloor R\rfloor}\frac{1}{(j+1-R)^2}}_{\leq\frac{\pi^2}{6}} \leq \frac{\varepsilon^p}{3}.
\end{aligned}$$

Finally, let us bound I_3. Recalling that $\mu_d = \sum_{i=1}^{\infty} c_i\delta_{x_i}$, we get

$$\begin{aligned}
\int_{B_R^c}\left(|x|-R\right)^p d\mu_d(x) &= \sum_{k=1}^{\infty}\mathbf{1}_{|x_k|\geq R}\left(|x_k|-R\right)^p\mu_d(x_k) \\
&= \sum_{k=1}^{\infty}\mathbf{1}_{|x_k|\geq R}(x_k)\cdot\left(|x_k|-R\right)^p c_k.
\end{aligned}$$

If we assume that c_k decays at the following rate

$$c_k \leq \frac{1}{(|x_k|-R)^p}\cdot\frac{1}{k^q}\cdot\frac{\varepsilon^p}{3}\cdot\frac{1}{\sum_{\ell=1}^{\infty}\ell^q}$$

for some $q > 1$ and $|x_k| > R$, then

$$\begin{aligned}\int_{B_R^c} \big(|x| - R\big)^p d\mu_d(x) &= \sum_{k=1}^{\infty} \mathbf{1}_{|x_k|\geq R}(x_k)\big(|x_k| - R\big)^p c_k \\ &\leq \frac{\varepsilon^p}{3} \frac{1}{\sum_{\ell=1}^{\infty} \ell^q} \sum_{k=1}^{\infty} \mathbf{1}_{|x_k|\geq R}(x_k) \frac{(|x| - R)^p}{(|x| - R)^p} \frac{1}{k^q} \\ &\leq \frac{\varepsilon^p}{3}.\end{aligned}$$

With these bounds for our decomposed measure μ, we get that

$$W_p(\mu, \widehat{\mu})^p \leq \frac{\varepsilon^p}{3} + \frac{\varepsilon^p}{3} + \frac{\varepsilon^p}{3} = \varepsilon^p,$$

and therefore $W_p(\mu, \widehat{\mu}) \leq \varepsilon$. □

Now, we can use the theorems above to approximate $\widehat{\mu}$ with the Voronoi cell approximations. In particular, we get the following corollary for Dirac train approximations and the piecewise constant approximations.

Corollary 4.12. *Assume that μ satisfies the assumptions of Theorem 4.11 and let $\widehat{\mu}$ be the compactly supported measure $(P_{B_R})_\sharp\mu$. Moreover, let $N \in \mathbb{N}$, $\mathcal{N} = \mathcal{N}(B_R, V_0)$, and set $h = 3\big(\frac{\mathcal{N}}{N}\big)^{\frac{1}{d}}$. Then*

$$W_p(\mu, \widehat{\mu}_h) \leq 3\,\mathrm{rad}(V_0)\mathcal{N}^{\frac{1}{d}} N^{-\frac{1}{d}} + \varepsilon,$$

$$W_p(\mu, \widehat{\mu}_h') \leq 3\,\mathrm{diam}(V_0)\mathcal{N}^{\frac{1}{d}} N^{-\frac{1}{d}} + \varepsilon,$$

where $\widehat{\mu}_h = \sum_{\lambda\in\Lambda} \widehat{\mu}(V_{h\lambda})\delta_{h\lambda}$ and $\widehat{\mu}_h' = \sum_{\lambda\in\Lambda} \frac{\widehat{\mu}(V_{h\lambda})}{|V_{h\lambda}|} \mathbb{1}_{h\lambda}$.

Proof. Letting $\widehat{\mu}$ be the measure from Theorem 4.11, we have

$$W_p(\mu, \widehat{\mu}_h) \leq W_p(\mu, \widehat{\mu}) + W_p(\widehat{\mu}, \widehat{\mu}_h).$$

By Theorem 4.11, $W_p(\mu,\widehat{\mu}) < \varepsilon$. Since $\widehat{\mu}$ is compactly supported within some ball of radius R, we have by Theorem 4.6 and Theorem 4.8 that $W_p(\widehat{\mu},\widehat{\mu}_h) \le 3\,\mathrm{rad}(V_0)\mathcal{N}^{\frac{1}{d}}N^{-\frac{1}{d}}$ and $W_p(\widehat{\mu},\widehat{\mu}_h') \le 3\,\mathrm{diam}(V_0)\mathcal{N}^{\frac{1}{d}}N^{-\frac{1}{d}}$. Putting these together yields the desired result. □

4.7 Nonuniform Approximations

Let $X := \{x_i\}_{i=1}^{\infty} \subset \mathbb{R}^d$ be a set of (finite or infinite) points that is separated ($x_i \neq x_j$, $i \neq j$). We define two quantities governing these points: the *mesh norm* given by

$$h_X := \sup_{y\in\mathbb{R}^d} \inf_{x_i\in X} |x_i - y|,$$

and the *minimum separation radius*,

$$q_X := \frac{1}{2}\inf_{i\neq j}|x_i - x_j|.$$

We denote by $V_i \subset \mathbb{R}^d$ the Voronoi region centered at x_i, and enforce $V_i \cap V_j = \emptyset$, $i \neq j$.

Lemma 4.13. *Let $\mu \in W_p(\mathbb{R}^d)$, $p \in [1,\infty)$. Let $X \subset \mathbb{R}^d$ be such that $0 < q_X \le h_X < \infty$. with Voronoi cells $\{V_i\}_{i=1}^{\infty}$. Then the following hold:*

(i) $\displaystyle\sum_{i=1}^{\infty} |x_i|^p \mu(V_i) \le 2^{p-1}h_X^p + 2^{p-1}M_p(\mu),$

(ii) $\displaystyle\sum_{i=1}^{\infty} \|x\|_{L_\infty(V_i)}^p \mu(V_i) \le 2^{p-1}\sum_{i=1}^{\infty} |x_i|^p \mu(V_i) + 2^{p-1}h_X^p,$

(iii) $\displaystyle\sum_{i=1}^{\infty} \|x\|_{L_\infty(V_i)}^p \mu(V_i) \le (2^{2p-2} + 2^{p-1})h_X^p + 2^{p-1}M_p(\mu).$

Proof. Proof of (ii): Note that

$$\begin{aligned}
\sum_{i=1}^{\infty} \|x\|_{L_\infty(V_i)}^p \mu(V_i) &\leq \sum_{i=1}^{\infty} (|x_i| + h_X)^p \mu(V_i) \\
&\leq 2^{p-1} \sum_{i=1}^{\infty} |x_i|^p \mu(V_i) + 2^{p-1} h_X^p \sum_{i=1}^{\infty} \mu(V_i) \\
&= 2^{p-1} \sum_{i=1}^{\infty} |x_i|^p \mu(V_i) + 2^{p-1} h_X^p.
\end{aligned}$$

Proof of (i): By the definition of the mesh norm and the triangle equality, the following holds for every i and every $x \in V_i$:

$$|x_i| \leq |x| + |x - x_i| \leq |x| + h_X,$$

hence $|x_i|^p \leq 2^{p-1}(|x|^p + h_X^p)$. Integrating this inequality over V_i with respect to μ and summing over i gives

$$\sum_{i=1}^{\infty} |x_i|^p \mu(V_i) \leq 2^{p-1} \sum_{i=1}^{\infty} \int_{V_i} |x|^p d\mu(x) + 2^{p-1} h_X^p \sum_{i=1}^{\infty} \mu(V_i) = 2^{p-1} M_p(\mu) + 2^{p-1} h_X^p,$$

which is the desired conclusion.

Proof of (iii): Combine (i) and (ii). □

Theorem 4.14. *Let $\mu \in W_p(\mathbb{R}^d)$, $p \in [1, \infty)$ be fixed but arbitrary, and let $X \subset \mathbb{R}^d$ be such that $0 < q_X \leq h_X < \infty$. Let $\mu_X := \sum_{i=1}^{\infty} \mu(V_i) \delta_{x_i}$. Then*

$$W_p(\mu, \mu_X) \leq h_X.$$

Proof. First, note that μ_X is clearly a measure, and we have

$$\mu_X(\mathbb{R}^d) = \sum_{i=1}^{\infty} \mu(V_i) = \mu\left(\bigcup_{i=1}^{\infty} V_i\right) = \mu(\mathbb{R}^d) = 1,$$

where the second and third equalities comes from countable additivity of μ and the fact that $\mathbb{R}^d = \sqcup_{i=1}^{\infty} V_i$. To show that μ_X has finite p-th moment, we notice that (via Tonelli's Theorem) and Lemma 4.13(i),

$$\int_{\mathbb{R}^d} |x|^p d\mu_X = \sum_{i=1}^{\infty} |x_i|^p \mu(V_i) \leq 2^{p-1} h_X^p + 2^{p-1} M_p(\mu) < \infty.$$

Using the Kantorovich formulation of W_p, we define the following is a (non-optimal) coupling of μ and μ_X:

$$\widetilde{\pi}(A,B) := \sum_{i=1}^{\infty} \mu(B \cap V_i) \delta_{x_i}(A) = \int_{A \times B} \sum_{i=1}^{\infty} \mathbb{1}_{V_i}(y) \delta_{x_i}(x) dx d\mu(y).$$

It is straightforward to check that $\widetilde{\pi}$ is a measure on $\mathbb{R}^d \times \mathbb{R}^d$. Noting that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} d\widetilde{\pi}(x,y) = \sum_{i=1}^{\infty} \mu(V_i) = \mu(\mathbb{R}^d) = 1,$$

we see that $\widetilde{\pi}$ is a probability measure, and its marginals are

$$\widetilde{\pi}(\mathbb{R}^d, B) = \sum_{i=1}^{\infty} \mu(B \cap V_i) = \mu(B),$$

and

$$\widetilde{\pi}(A, \mathbb{R}^d) = \sum_{i=1}^{\infty} \mu(V_i) \delta_{x_i}(A) = \mu_X(A),$$

for all Borel measurable sets $A, B \in \mathbb{R}^d$. Therefore, $\widetilde{\pi}$ is a coupling of μ and μ_X.

Notice that if $k \neq j$, then

$$\widetilde{\pi}(V_k, V_j) = \sum_{i=1}^{\infty} \mu(V_k \cap V_i) \delta_{x_i}(V_j) = 0,$$

that is, $\widetilde{\pi}$ only evaluates mass on sets intersecting $V_i \times V_i$. Therefore, we have

$$\begin{aligned}
W_p(\mu,\mu_X)^p &\leq \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p d\widetilde{\pi}(x,y)\\
&= \sum_{i=1}^{\infty} \int_{V_i\times V_i} |x-y|^p d\widetilde{\pi}(x,y)\\
&= \sum_{i=1}^{\infty} \int_{V_i} |x_i-y|^p d\mu(y)\\
&\leq h_X^p \sum_{i=1}^{\infty} \mu(V_i)\\
&= h_X^p.
\end{aligned}$$

□

Theorem 4.15. *Let $\nu \in W_p(\mathbb{R}^d)$, $p \in [1,\infty)$ be fixed but arbitrary, and let $X \subset \mathbb{R}^d$ be such that $0 < q_X \leq h_X < \infty$. Let $\nu_X := \sum_{i=1}^{\infty} \frac{\nu(V_i)}{|V_i|} \mathbb{1}_{V_i}$. Then $\nu_X \in W_p(\mathbb{R}^d)$ and*

$$W_p(\nu,\nu_X) \leq 2h_X.$$

Proof. First, notice that

$$\nu_X(\mathbb{R}^d) = \sum_{i=1}^{\infty} \frac{\nu(V_i)}{|V_i|} \int_{V_i} dx = \nu(\mathbb{R}^d) = 1,$$

so indeed ν_X is a probability measure. Next, we see that

$$\begin{aligned}
\int_{\mathbb{R}^d} |x|^p d\nu_X(x) &= \sum_{i=1}^{\infty} \frac{\nu_X(V_i)}{|V_i|} \int_{V_i} |x|^p dx\\
&\leq \sum_{i=1}^{\infty} \frac{\nu_X(V_i)}{|V_i|} \|X\|_{L_\infty(V_i)}^p |V_i|\\
&\leq (2^{p-2}+2^{p-1})h_X^p + 2^{p-1}M_p(\mu) < \infty,
\end{aligned}$$

whereby $\nu_X \in W_p(\mathbb{R}^d)$. Here, and throughout the proof, Tonelli's Theorem justifies the interchange of sum and integral.

To estimate the convergence rate, we again form a (non-optimal) coupling in the Kantorovich sense by

$$\widetilde{\pi}(A,B) := \sum_{i=1}^{\infty} \frac{\nu(A\cap V_i)}{|V_i|}|B\cap V_i| = \int_{A\times B} \sum_{i=1}^{\infty} \frac{1}{|V_i|} \mathbb{1}_{V_i}(x)\mathbb{1}_{V_i}(y) d\nu(x)dy.$$

It is straightforward to check that $\widetilde{\pi}$ is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$. Its marginals are

$$\widetilde{\pi}(A,\mathbb{R}^d) = \sum_{i=1}^{\infty} \frac{\nu(A\cap V_i)}{|V_i|}|V_i| = \nu(A)$$

as before, and

$$\widetilde{\pi}(\mathbb{R}^d,B) = \sum_{\lambda\in\Lambda} \frac{\nu(V_i)}{|V_i|}|B\cap V_i| = \sum_{i=1}^{\infty} \frac{\nu(V_i)}{|V_i|}\int_B \mathbb{1}_{V_i}(x)dx = \nu_X(B).$$

Therefore, $\widetilde{\pi}$ is a coupling of ν and ν_h.

Finally,

$$\begin{aligned} W_p(\nu,\nu_X)^p &\le \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p d\widetilde{\pi}(x,y) \\ &= \sum_{i=1}^{\infty} \frac{1}{|V_i|}\int_{V_i\times V_i} |x-y|^p d\nu(x)dy \\ &\le \sum_{i=1}^{\infty} \frac{\operatorname{diam}(V_i)^p}{|V_i|}\nu(V_i)|V_i| \\ &\le 2^p h_X^p \sum_{i=1}^{\infty} \nu(V_i) \\ &= 2^p h_X^p. \end{aligned}$$

□

Theorem 4.16. *Let $\mu \in W_p(\mathbb{R}^d)$, $p \in [1,\infty)$ have compact support contained in the interior of*

B_R. Let $X \subset \mathbb{R}^d$ be such that $0 < q_X \leq h_X < \infty$ and such that the cardinality of $X \cap B_R$ is N. If $RN^{-\frac{1}{d}} \leq h_X \leq CN^{-\frac{1}{d}}$, then $\mu_X = \sum_i \mu(V_i)\delta_{x_i}$ is an at most N-term approximation of μ, which satisfies

$$W_p(\mu, \mu_X) \leq CN^{-\frac{1}{d}}.$$

Proof. Note that N being the number of points in X contained in B_R implies also that N is at least the number of Voronoi regions intersecting B_R. The mesh norm h_X is the largest radius of one of the Voronoi regions, which means that N balls of radius h_X must cover B_R; that is

$$N \geq \mathcal{N}(B_R, B_{h_X}).$$

By volumetric arguments, $\mathcal{N}(B_R, B_{h_X}) \geq R^d h_X^{-d}$ [11, Theorem 4.1.13]. Rearranging yields the assumed lower bound on h_X. Next, with the upper bound on h_X, applying Theorem 4.16 implies that

$$W_p(\mu, \mu_X) \leq h_X \leq CN^{-\frac{1}{d}},$$

as required. □

Theorem 4.17. *Let $\nu \in W_p(\mathbb{R}^d)$, $p \in [1, \infty)$ have compact support contained in the interior of B_R. Let $X \subset \mathbb{R}^d$ be such that $0 < q_X \leq h_X < \infty$ and such that the cardinality of $X \cap B_R$ is N. If $RN^{-\frac{1}{d}} \leq h_X \leq CN^{-\frac{1}{d}}$, then $\nu_X = \sum_i \frac{\nu(V_i)}{|V_i|} \mathbb{1}_{V_i}$ is an at most N-term approximation of ν, which satisfies*

$$W_p(\nu, \nu_X) \leq 2CN^{-\frac{1}{d}}.$$

Proof. Mimic the proof of Theorem 4.16 *mutatis mudandis* applying Theorem 4.15. □

Corollary 4.18. *Assume that $\mu \in W_p * \mathbb{R}^d)$ satisfies the assumptions of Theorem 4.11, and let $\widehat{\mu}$ be the compactly supported measure $(P_{B_R})_\sharp \mu$. Moreover, invoke the assumptions of Theorems*

4.16 and 4.17. Then

$$W_p(\mu, \widehat{\mu}_X) \le CN^{-\frac{1}{d}} + \varepsilon,$$
$$W_p(\mu, \widehat{\mu}'_X) \le 2CN^{-\frac{1}{d}} + \varepsilon,$$

where $\widehat{\mu}_h = \sum_i \widehat{\mu}(V_i)\delta_{x_i}$ *and* $\widehat{\mu}'_X = \sum_i \frac{\widehat{\mu}(V_i)}{|V_i|}\mathbb{1}_{V_i}$.

Proof. Mimic the proof of Corollary 4.12 using the results from this section. □

4.8 Linear optimal transport in the Kantorovich setting for Dirac train approximations

In this section, we extend the ideas used in LOT to the Kantorovich formulation of optimal transport and investigate an analogue to compatibility when we work with regularized optimal transport. All the work done in the Dirac train approximations is essentially an infinite grid, but cutting off the grid at a certain portion, we get a finite grid akin to the ones discussed in

Given two discrete measure $a, r \in \Sigma_n$, where r is the reference measure, we define the regularized LOT embedding for a fixed $\beta > 0$ by

$$F_{r,C}(a) = P^b_{a,\beta}(C)^\top, \tag{4.7}$$

where $P^b_{a,\beta}(C)$ is the unique solution to the regularized problem (4.4). Note that this is a mapping from Σ_n to $M_{n\times n}(\mathbb{R})$, the space of matrices of size $(n \times n)$.

4.8.1 Pushforward of discrete measures

Given a discrete measure $a \in \Sigma_n$, we consider pushforwards of $a \in \Sigma_n$ by a transport plan $S \in \Pi_a$, where

$$\Pi_a = \{S \in \mathbb{R}^{n\times m} : S^\top \mathbb{1}_n \in \Sigma_m, S\mathbb{1}_m = a\}$$

and m is subject to change depending on which output grid we would like to push forward into. In this sense, we can specify pushing a forward by the following definition:

Definition 4.19. *Let $a \in \Sigma_n$ and let $S \in \Pi_a$. Then, the pushforward of a by S is defined by*

$$S_\sharp a := S^\top \operatorname{diag}(a^{-1})a = S^\top \mathbb{1}_n \in \Sigma_m.$$

This definition of pushforward directly extends the definition of the pushforward operator and pushforward measure in [79]. In particular, given a discrete measure $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ where $x_i \in \mathcal{X}$ and a continuous map $T : \mathcal{X} \to \mathcal{Y}$, [79] defines the pushforward of α under T as

$$T_\sharp \alpha = \sum_{i=1}^n a_i \delta_{T(x_i)}.$$

This definition is restricted to “transport maps” rather than transport plans. With Definition 4.19, we can expand the definition of a pushforward to include mass splitting.

4.8.2 Regularized compatibility for discrete measures

One of the core necessities for LOT embeddings is understanding how push-forwards of distributions create changes in the LOT embedding space. Ideally, we would like to relate a push-forward S directly to the embedding $F_{r,C}(a)$ so that $F_{r,C}(S_\sharp a)$ and $F_{r,C}(a)$ are related.

In the process of showing this relationship, we derive a few results. First, given a discrete reference measure $r \in \Sigma_n$ and a target measure $a \in \Sigma_n$ and an a-pushforward denoted by $S \in \Pi_a$, we will see how to transform Π_r^a into $\Pi_r^{S_\sharp a}$.

Lemma 4.20. *Let $r, a \in \Sigma_n$ and assume that $S \in \Pi_a \subseteq \mathbb{R}^{n\times m}$ with is chosen such that the columns of S form a spanning set for $\mathbb{R}^n$. Then the transformation*

$$\begin{cases} \mathcal{G}(S) : \Pi_r^a \to \Pi_r^{S_\sharp a} \\ P \to P \operatorname{diag}(a^{-1}) S \end{cases}$$

is a one-to-one map from Π_r^a to $\Pi_r^{S_\sharp a}$.

Proof. We need to check first that for $P \in \Pi_r^a$, we have $\mathcal{G}(S)(P) \in \Pi_r^{S_\sharp a}$. Indeed, note that

$$P \operatorname{diag}(a^{-1}) S \mathbb{1} = P \operatorname{diag}(a^{-1}) a = P \mathbb{1} = r$$
$$S^\top \operatorname{diag}(a^{-1}) P^\top \mathbb{1} = S^\top \operatorname{diag}(a^{-1}) a = S^\top \mathbb{1} = S_\sharp a.$$

Now, we want to show that S is one-to-one. To show this, let $s_1, \ldots, s_m$ denote the columns of S and consider $P, P' \in \Pi_r^a$ with $P \neq P'$. Let $p_1, \ldots, p_n$ denote the rows of P, and likewise, let $p'_1, p'_2, \ldots, p'_n$ denote the rows of P'. If we assume towards a contradiction that $\mathcal{G}(S)(P) = \mathcal{G}(S)(P')$, then

$$\begin{aligned} (P - P') \operatorname{diag}(a^{-1}) S = 0 \\ \implies s_j \in \ker\Big((P - P') \operatorname{diag}(a^{-1})\Big) \\ \implies a_\ell^{-1} \cdot \langle p_\ell - p'_\ell, s_j \rangle = 0 \ \ \forall j, \ell. \end{aligned}$$

But since $\{s_j\}$ are the columns of S and form a spanning set, we know that $\langle p_\ell - p'_\ell, s_j \rangle = 0$ for all j implies that $p_\ell = p'_\ell$, which contradicts that $P \neq P'$. This means that $\mathcal{G}(S)$ is a one-to-one map. □

A simple, yet important example of pushforwards are permutations because given a permutation Q, $\mathcal{G}(Q)$ will form a permutation. Moreover, permutations are the analogue of optimal transport maps in the context of discrete measures. Given a permutation Q, we can construct a corresponding transport plan in $S_Q(a) \in \Pi_a$ by defining

$$S_Q(a) = \operatorname{diag}(a) Q^\top.$$

Note, here that

$$\begin{aligned} S_Q(a)\mathbb{1}_n = \operatorname{diag}(a)Q^\top \mathbb{1}_n = \operatorname{diag}(a)\mathbb{1}_n = a \\ S_Q(a)^\top \mathbb{1}_n = Q\operatorname{diag}(a)\mathbb{1}_n = Qa. \end{aligned}$$

This is exactly what we expect from a permutation push-forward. In light of the previous lemma, we get the following corollary.

Corollary 4.21. *The map $\mathcal{G}(S_Q(a)) : \Pi_r^a \to \Pi_r^{S_Q(a)} = \Pi_r^{Qa}$ is a bijection.*

Proof. We just need to find an inverse map to $\mathcal{G}(S_Q(a))$. The map that will end up working is $\mathcal{G}(S_{Q^\top}(Qa)) : \Pi_r^{Qa} \to \Pi_r^a$. Indeed, note that if $P \in \Pi_r^a$, then

$$\begin{aligned} \mathcal{G}(S_{Q^\top}(Qa)) \circ \mathcal{G}(S_Q(a))(P) &= P\operatorname{diag}(a^{-1})\underbrace{\operatorname{diag}(a)Q^\top}_{S_Q(a)}\operatorname{diag}((Qa)^{-1})\underbrace{\operatorname{diag}(Qa)Q}_{S_{Q^\top}(Qa)} \\ &= P. \end{aligned}$$

So we're done. □

Permutations are special in the context of relating $F_{r,C}((S_Q(a))_\sharp a)$ and $F_{r,C}(a)$ because of the following lemma.

Lemma 4.22. *Let $r, a \in \Sigma_n$ and Q a permutation, then $P^a_{r,\beta}(C)Q^\top = P^{Qa}_{r,\beta}(CQ^\top)$, $P^{Qa}_{r,\beta}(C)Q =$*

$P^{a}_{r,\beta}(CQ)$*, and* $P^{a}_{r,\beta}(QC) = QP^{a}_{Q^{\top}r,\beta}(C)$

Proof. First, recall that $\mathcal{G}(S_Q(a))(P^{a}_{r,\beta}(C)) \in \Pi^{Qa}_{r}$ from the corollary above. We know that $P^{a}_{r,\beta}(C)$ has the form $\operatorname{diag}(u)e^{-\beta C}\operatorname{diag}(v)$; thus,

$$\begin{aligned}
\mathcal{G}(S_Q(a))(P^{a}_{r,\beta}(C)) &= \operatorname{diag}(u)e^{-\beta C}\operatorname{diag}(v)\operatorname{diag}(a^{-1})\operatorname{diag}(a)Q^{\top} \\
&\overset{(1)}{=} \operatorname{diag}(u)e^{-\beta C}Q^{\top}Q\operatorname{diag}(v)Q^{\top} \\
&\overset{(2)}{=} \operatorname{diag}(u)e^{-\beta CQ^{\top}}\operatorname{diag}(Qv),
\end{aligned}$$

where (1) comes from the fact that $Q^{\top}Q = I$ and (2) comes from that fact that permutation matrices are the normalizer of diagonal matrices (i.e. $Q\operatorname{diag}(v)Q^{\top} = \operatorname{diag}(Qv)$) and $e^{-\beta C}Q^{\top} = e^{-\beta CQ^{\top}}$ since the exponentiation entry-wise. Given the form of the last equation, we can see that

$$P^{a}_{r,\beta}(C)Q^{\top} = \operatorname{diag}(u)e^{-\beta CQ^{\top}}\operatorname{diag}(Qv) = P^{Qa}_{r,\beta}(CQ^{\top}).$$

Using the same exact reasoning, we also get that

$$P^{Qa}_{r,\beta}(C)Q = P^{a}_{r,\beta}(CQ).$$

Finally, for the last equality, we have that $P^{a}_{r,\beta}(QC)$ can be written in the form

$$\begin{aligned}
P^{a}_{r,\beta}(QC) &= \operatorname{diag}(\tilde{u})e^{-\beta QC}\operatorname{diag}(\tilde{v}) \\
&= QQ^{\top}\operatorname{diag}(\tilde{u}Qe^{-\beta C}\operatorname{diag}(\tilde{v}) \\
&= Q\underbrace{\operatorname{diag}(Q^{\top}\tilde{u})e^{-\beta C}\operatorname{diag}(\tilde{v})}_{P}.
\end{aligned}$$

We want to see which marginals P has. In particular, note that

$$\begin{aligned} P^\top \mathbb{1} &= P^\top Q^\top \mathbb{1} = P^a_{r,\beta}(QC)^\top \mathbb{1} = a \\ QP\mathbb{1} &= P^a_{r,\beta}(QC)\mathbb{1} = r \\ \implies P\mathbb{1} &= Q^\top r. \end{aligned}$$

Because of the form P, we see that $P^a_{r,\beta}(QC) = QP^a_{Q^\top r,\beta}(C)$. So we're done. □

Combining some of the equalities above, we immediately get the following corollary.

Corollary 4.23. *Assume that C is symmetric, and let Q be a permutation matrix that commutes with C, then $P^a_{r,\beta}(C) = Q^\top P^{Qa}_{Qr,\beta}(C)Q$ (or equivalently $QP^a_{r,\beta}(C)Q^\top = P^{Qa}_{Qr,\beta}(C)$).*

Proof. If $CQ = QC$, then notice that $Q^\top C = CQ^\top$. Then using the equalities from above, we have

$$P^a_{r,\beta}(C) = P^a_{r,\beta}(Q^\top QC) = P^a_{r,\beta}(Q^\top CQ) = P^{Qa}_{r,\beta}(Q^\top C) = Q^\top P^{Qa}_{Qr,\beta}(C)Q.$$

This finishes the proof. □

We finally get another corollary in the circumstance that $Qr = r$.

Corollary 4.24. *Let r and a be discrete measures. If C is a symmetric matrix and Q is a permutation matrix such that $CQ = QC$ and $Qr = r$, then $QP^a_{r,\beta}(C)Q^\top = P^{Qa}_{r,\beta}(C)$.*

4.9 Acknowledgments

The authors thank Oleksandr Vlasiuk for introducing them to optimal quantizers for measures and their relation to Wasserstein distance.

Both authors were partially supported by the National Science Foundation under Grant No. DMS-1929284 while visiting the Institute for Computational and Experimental Research in Mathematics (ICERM) for the Workshop on Optimal Transport in Data Science. The authors thank ICERM for their hospitality and the organizers for a stimulating workshop.

KH was partially supported by a Research Enhancement Program grant from the College of Science at the University of Texas at Arlington. Research was sponsored by the Army Research Office and was accomplished 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.

This chapter, in full, is a reprint of *Lattice approximations in wasserstein space* (2023) by K. Hamm and V. Khurana as it appears in `https://arxiv.org/abs/2310.09149`. The dissertation author was the primary investigator and author of this paper.

Chapter 5

A Neural Network Two-Sample Test

JOINT WORK WITH ALEX CLONINGER AND XIUYUAN CHENG

We construct and analyze a neural network two-sample test to determine whether two datasets came from the same distribution or not. We perform some time-analysis on a neural tangent kernel (NTK) two-sample test and extend the analysis to the regular neural network two-sample test by approximating the neural network dynamics with the NTK dynamics. Although the approximation relies on a small-time training regime, the complexity of the neural network in relation to the complexity of the two-sample problem considered still allows for the approximation to hold. We particularly show the theoretical minimum time needed for the neural network two-sample test to sense a difference $\varepsilon > 0$ between the datasets and the theoretical maximum time before the two-sample test senses a difference $\varepsilon > 0$. Additionally, we run some experiments showcasing a two-layer neural network two-sample test on a hard two-sample test problem. We show the statistical power of the test in relation to the time it takes to train and how complex the network is.

5.1 Introduction

The ability to compare whether two datasets $\widehat{P} \sim p$ and $\widehat{Q} \sim q$ came from the same data-generating process (i.e. checking if $p = q$ or $p \neq q$) is a problem studied for many years. Traditionally, the methods to answer this question are called two-sample tests. As a non-exhaustive list of applications, two-sample testing is widely used in testing drug efficacy [37], studying behavioral differences in psychology [16], pollution impact studied in environmental science research [20], and market research impact studies [22]. The most basic method to compare distributions is by comparing means with a t-test, proportions with a z-test, variances with Levene's test, medians with a Mann-Whitney U test, or overall distributions with a Kolmogorov-Smirnov test. The advent of complex, high-dimensional data in fields like genomics, finance, and social media analytics has exposed limitations in these traditional methods, particularly in terms of handling non-linearity, complex interactions, and the curse of dimensionality. The flexibility and scalability of neural networks make them particularly suited to tackle the challenges posed by modern datasets, suggesting their potential to revolutionize two-sample testing.

This paper is not the first to explore this idea of using neural networks or classifiers for two-sample testing. In particular, [62] shows properties and analyzes performance of the so-called Classifier Two-Sample Test (C2ST) and specifically showcasing theoretically what the statistical power of such two-sample tests. To go further in the neural network direction, [30] expanded [47]'s work and used the neural tangent kernel (NTK) for the kernel involved in a maximum mean discrepancy (MMD) problem. Yet their analysis still did not relate the NTK MMD performance to the behavior of neural network two-sample tests. Moreover, [28] introduced a neural network-based two sample test statistic using the classification logit and show theoretical guarantees for test power for sub-exponential densities problems. One may be hesitant to use a neural network for two-sample tests since with a big enough neural network and long enough training time, a neural network could find a separation for data coming from the same distribution. Our approach alleviates this hesitation since we train the neural network on a small time-scale and ensure our network is initialized to output 0 for all values. We also conduct time analysis on two levels. First, we analyze the time needed for achieving a desired level of deviation or detection in the two-sample test. Second, we provide time approximations between different training regimes, which extends the analysis of the time needed for detection to different training regimes.

5.1.1 Main Contributions

Our main contributions to the field are the following:

1. We perform some time analysis on the neural tangent kernel (NTK) derived from our neural network and show that the time it takes for the neural network two-sample test to learn does *not* depend on the entire spectrum of the NTK but rather only a subset of the spectrum on which the labels or witness function $f^* = \frac{p-q}{p+q}$ non-trivially projects onto. This behavior is a result of averaging behavior of the neural network two-sample test.
2. We approximate the population-level neural network dynamics and finite-sample neural network dynamics with the population-level NTK dynamics. This allows the time analysis performed on the NTK dynamics to transfer to the other two training regimes. Additionally, we notice here that there is a balancing act of not training the neural network too long so that the the approximations hold but long enough to detect differences in the datasets. This balancing act is further informed by the complexity of the neural network considered in relation to the difficulty of the two-sample test problem.

Our main result essentially shows that as long as p and q are "separated enough", our neural network two-sample test can detect the difference before the same detection would take place if $p = q$. In particular, we can summarize the main result of value as the following informal theorem.

Theorem 5.1 (Informal)**.** *Assume the alternative hypothesis that f^* nontrivially projects onto the first k eigenfunctions of the zero-time NTK K_0 holds true. Given a desired detection level $\varepsilon > 0$ and time separation level $C\varepsilon \geq \gamma > 0$, further assume that the projection of f^* onto the first k eigenfunctions has a "large enough norm." Then with high probability,*

$$t^+(\varepsilon) - t^-(\varepsilon) \geq \gamma > 0,$$

where $t^-(\varepsilon)$ and $t^+(\varepsilon)$ are the minimum times needed for the neural network two-sample test to detect a deviation ε under the alternative hypothesis and the null hypothesis, respectively.

In the formal version of this theorem (as shown in Corollary 5.33), the detection level $\varepsilon > 0$ possible is perturbed by a time-approximation error between the actual neural network two-sample test and the zero-time NTK two-sample test. This adds a small amount of complexity to the informal theorem above and there are lower bound conditions on f^* to ensure detectability. We also discuss (in a subsequent remark to Corollary 5.33) which detection level is the most trustworthy. A visual for this graph is given in Figure 5.1.

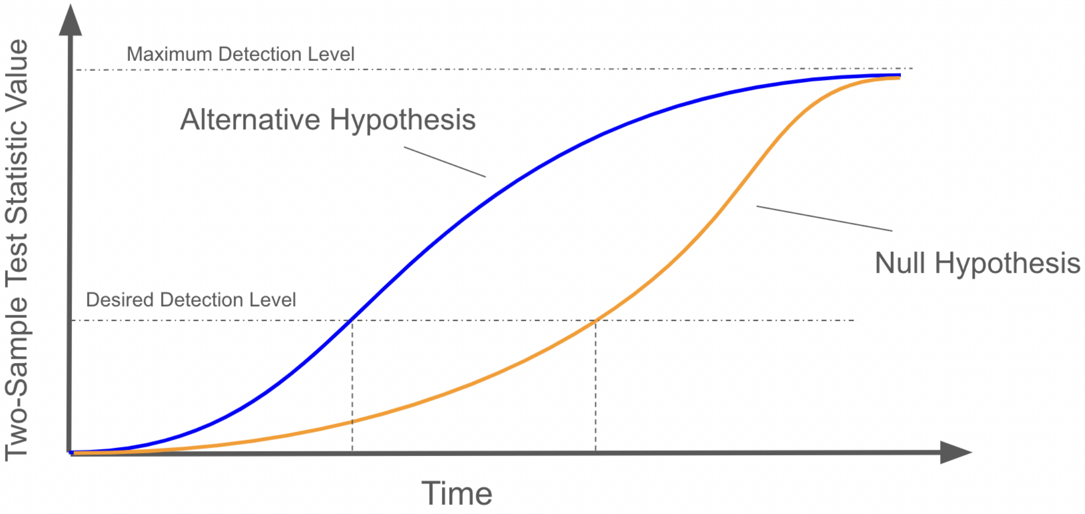


Figure 5.1: Visual for detection levels $t^+(\varepsilon)$ and $t^-(\varepsilon)$ being well-separated.

5.1.2 Structure of the paper

We review some papers in Section 5.2 that study the same topic that we explore in this paper. In Section 5.3, we introduce the main notation, a motivating example, and concepts that we will need for the rest of the paper. We discuss specifics of how the training time-scale interplays with the network complexity in Section 5.4.

In Section 5.5, we describe in detail three training regimes that we consider for the two-sample test. First, we consider the case of finite-sample time-varying dynamics; second, we consider population-level time-varying dynamics; and finally, we consider the zero-time neural tangent kernel (NTK) dynamics training regime where the analysis is easier to understand. By solving for the actual solution of the zero-time NTK dynamics, we are able to find an exact form for the two-sample test in this regime by using the spectrum of the NTK. The exact form of the two-sample test further allows us to conduct some time analysis. The time analysis is done to show guarantees for when the null hypothesis is correct or when the alternative hypothesis is correct. For the alternative hypothesis, we can estimate the minimum time needed for sensing an error level $\varepsilon > 0$. For the null hypothesis statement, we can estimate the maximum time needed before we are able to sense past an error level $\varepsilon > 0$. Using proof techniques similar to [81], these time-analysis results are adapted later to the other training regimes by using approximation and estimation between the different regimes.

In Section 5.6, we estimate the population-level time-varying dynamics with the zero-time NTK dynamics. We are essentially able to approximate the population-level time-varying dynamics with the zero-time NTK dynamics up to a factor of $t^{3/2}$ where t denotes time. This means that if the zero-time NTK dynamics two-sample test are able to detect f^* faster than the approximation guarantees in this section, then all the time analysis for the NTK dynamics also holds for the population-level time-varying dynamics two-sample test.

Section 5.7 is the main section for establishing time analysis for the finite-sample time-varying dynamics case that we usually see in practice. The approximation guarantees between the finite-sample time-varying dynamics and the zero-time NTK dynamics hold up to a factor of $t^{5/2}$ and depends on how many data points are sampled from both p and q.

The results shown in Section 5.6 and Section 5.7 showcase that our neural network two-sample test is more useful in identifying when the alternative hypothesis is correct. Moreover, we see a sort of balancing act of training on short time scales and increasing the complexity of the neural network. These specifics are discussed in more detail in Section 5.4. Finally, in Section 5.8 we show empirical evidence of the statistical power of the neural network two-sample test on a hard two-sample test problem.

5.2 Previous Works

In recent years, there has been growing interest in developing two-sample tests based on neural networks, leveraging the power of deep learning to address some of the limitations of classical two-sample tests. More traditional two-sample methods use methods such as kernel two-sample test and maximum mean discrepancy (MMD) [47]. [30] took this idea of using the MMD a bit further by changing the kernel to be the neural tangent kernel (NTK) of a neural network which resulted in an NTK MMD two-sample test. The

Variations of a neural network-based two-sample test are present in [58] and [28], and the analysis done in this paper goes further by using small time approximations between the NTK-based kernel machines and the actual neural network training dynamics. To accomplish this, we use very similar proof techniques to [81], however, rather than using a loss rescaling to get into the lazy training regime [31], we are able to use small time approximations for the NTK.

5.3 Notation and Background

We will study a neural network two-sample test, which will test whether two datasets came from the same distribution or not. In particular, assume that we are given datasets $X = \{x_i\}_{i=1}^{n_p} \subseteq \mathbb{R}^d$ and $Z = \{z_j\}_{j=1}^{n_q} \subseteq \mathbb{R}^d$. We will endow samples from X to have labels 1 whilst samples from Z will have labels -1. To give some more structure to our problem, we will moreover assume that the datasets X and Z are sampled from distributions $p(x)dx$ and $q(x)dx$, respectively, where p and q are associated density functions. From X and Z, note that we can construct finite-sample empirical measures

$$\widehat{p}(x)dx = \frac{1}{n_p}\sum_{i=1}^{n_p} \delta_{x_i}(x)dx$$
$$\widehat{q}(x)dx = \frac{1}{n_q}\sum_{j=1}^{n_q} \delta_{z_j}(x)dx$$

respectively. In the same fashion, we can assume that we are given **independent** test samples from each of p and q to generate $X_{test} = \{x_i^*\}_{i=1}^{m_p}$ and $Z_{test} = \{z_j\}_{j=1}^{m_q}$ as well as corresponding *test* empirical measures $\widehat{p}_{test}(x)dx$ and $\widehat{q}_{test}(x)dx$. These test sets will be used when considering the finite-sample two-sample test on test data. We now introduce the following notation

$$\|f\|_{L^2(p+q)} = \left(\int_{\mathbb{R}^d} |f(x)|^2 (p(x)+q(x))dx\right)^{1/2}$$
$$\|f\|_{L^2(\widehat{p}+\widehat{q})} = \left(\int_{\mathbb{R}^d} |f(x)|^2 (\widehat{p}(x)+\widehat{q}(x))dx\right)^{1/2}.$$

Assume that our neural network architecture has associated parameters space $\Theta \subseteq \mathbb{R}^{M_\Theta}$ so that our neural network is given as $f : \mathbb{R}^d \times \Theta \to \mathbb{R}$ and will be trained on an ℓ_2 loss function against the labels as shown here

$$\begin{aligned}\widehat{L}(\theta) &= \frac{1}{2}\left(\frac{1}{n_p}\sum_{i=1}^{n_p}\big(f(x_i,\theta)-1\big)^2 + \frac{1}{n_q}\sum_{j=1}^{n_q}\big(f(z_j,\theta)+1\big)^2\right) \\ &= \frac{1}{2}\left(\int_{\mathbb{R}^d}\big(f(x,\theta)-1\big)^2\widehat{p}(x)dx + \int_{\mathbb{R}^d}\big(f(x,\theta)+1\big)^2\widehat{q}(x)dx\right).\end{aligned}$$

As a precursor to the more concrete notation introduced in Section 5.5, we will use the general rule of thumb of distinguishing mathematical objects in different training regimes by:

1. Finite-sample time-varying mathematical objects are adorned with hats, such as $\widehat{u}$.

2. Population-level time-varying mathematical objects are not adorned with any specific notation, such as u.

3. Population-level zero-time NTK mathematical objects are adorned with bars, such as $\overline{u}$.

5.3.1 Motivating Example

For our motivating example two-sample test scenario, we consider when our probability distributions of interest are two multivariate normals with the same covariance matrix but different means. In particular, with a fixed covariance matrix Σ, we let $p \sim N(\mu_1, \Sigma)$ and $q \sim N(\mu_2, \Sigma)$ with labels 1 and -1 respectively. Assume that we work with a linear neural network given by

$$f(x; a, W, b) = \frac{1}{M_\Theta} a^\top \Big(Wx + b\Big)$$

where $x \in \mathbb{R}^d, a \in \mathbb{R}^{M_\Theta}$, $W \in \mathbb{R}^{M_\Theta \times d}$, and $b \in \mathbb{R}^{M_\Theta}$. For ease assume that M_Θ is even, then for initialization, let $b = 0$ and opt to make $a_i = 1$ for $i \leq M_\Theta/2$ and $a_i = -1$ otherwise. For W, we will generate a random matrix $\widetilde{W} \in \mathbb{R}^{(M_\Theta/2)\times d}$ and let $W = \begin{bmatrix} \widetilde{W} \\ -\widetilde{W} \end{bmatrix}$. These choices will ensure that $f(x) = 0$ for all x.

Recall the gradient of f with respect to its parameters is given by

$$\begin{aligned}\frac{\partial f}{\partial a}(x) &= \frac{1}{M_\Theta}(Wx+b)\\ \frac{\partial f}{\partial W}(x) &= \frac{1}{M_\Theta}ax^\top\\ \frac{\partial f}{\partial b}(x) &= \frac{1}{M_\Theta}a.\end{aligned}$$

We will show that just one population-level gradient descent step with this setup will allow the two-sample test to detect the difference in distributions with high probability. In particular, recall that with our initialization

$$\frac{\partial \widehat{L}(\theta_0)}{\partial f}(x) = \begin{cases} -1 & x \sim p \\ 1 & x \sim q \end{cases}.$$

Now with learning rate η, one gradient descent step gives us

$$\begin{aligned} a^{(1)} &= a - \eta \int \frac{\partial f}{\partial a}(x)\frac{\partial \widehat{L}}{\partial f}(x) d(p+q)(x)\\ &= a - \eta \int \frac{1}{M_\Theta} Wx d(q-p)(x) = a - \frac{\eta}{M_\Theta}W(\mu_2 - \mu_1)\\ W^{(1)} &= W - \eta \int \frac{\partial f}{\partial W}(x)\frac{\partial \widehat{L}}{\partial f}(x) d(p+q)(x)\\ &= W - \eta \int \frac{1}{M_\Theta} ax^\top d(q-p)(x) = W - \frac{\eta}{M_\Theta}a(\mu_2 - \mu_1)^\top\\ b^{(1)} &= b - \eta \int \frac{\partial f}{\partial b}(x)\frac{\partial \widehat{L}}{\partial f}(x) d(p+q)(x)\\ &= b - \eta \int \frac{1}{M_\Theta} a d(q-p)(x) = b - 0 = 0. \end{aligned}$$

This means that after the first gradient descent step, we have

$$\begin{aligned}
f(x;a^{(1)},W^{(1)},0) &= \left(a - \frac{\eta}{M_\Theta} W(\mu_2-\mu_1)\right)^\top \left(\left(W - \frac{\eta}{M_\Theta} a(\mu_2-\mu_1)^\top\right)x\right) \\
&= \frac{\eta}{M_\Theta}\Bigg(\|a\|^2\langle \mu_1-\mu_2, x\rangle + \langle W(\mu_1-\mu_2), Wx\rangle \\
&\quad + \frac{\eta}{M_\Theta}\langle a, W(\mu_2-\mu_1)\rangle\langle \mu_2-\mu_1, x\rangle\Bigg).
\end{aligned}$$

Now notice that if we consider the two-sample test

$$\begin{aligned}
\int f(x;a^{(1)},W^{(1)},0)d(p-q)(x) = \frac{\eta}{M_\Theta}\Bigg(&\|a\|^2\|\mu_1-\mu_2\|^2 + \|W(\mu_1-\mu_2)\| \\
&+ \frac{\eta}{M_\Theta}\langle a, W(\mu_1-\mu_2)\rangle\|\mu_1-\mu_2\|\Bigg).
\end{aligned}$$

In essence, if η is small enough the two-sample test will be positive and the farther μ_1 is away from μ_2, the easier it becomes to detect.

In the case that we have W fixed and M_Θ is large, we can see that a is trying to learn $W(\mu_1-\mu_2)$. From a qualitative point of view, we only really need one row w of W to form a hyperplane that separates μ_1 and μ_2, assuming that $0,\mu_1,\mu_2$ do not all fall on the same line (and μ_1 and μ_2 are not on opposite sides of 0). Moreover, the larger we pick M_Θ, the random matrix W gets a greater probability of producing such a row w. Moreover, producing such a w becomes increasingly more likely when we center the data so that the origin is between the two means.

5.3.2 Relating Finite-sample and Population-level Loss

We now revert to the more general case of neural networks considered and recall the form of $\widehat{L}(\theta)$. Notice that when $n_p, n_q \to \infty$, we get a population-level loss given by

$$\begin{aligned}L(\theta) &= \frac{1}{2}\left(\int_{\mathbb{R}^d}\big(f(x,\theta)-1\big)^2 p(x)dx + \int_{\mathbb{R}^d}\big(f(x,\theta)+1\big)^2 q(x)dx\right)\\ &= \frac{1}{2}\left(\int_{\mathbb{R}^d}\big(f(x,\theta)^2 p(x) - 2f(x,\theta)p(x) + p(x)\right.\\ &\left. + f(x,\theta)^2 q(x) + 2f(x,\theta)q(x) + q(x)\big)dx\right)\\ &= \frac{1}{2}\left(\int_{\mathbb{R}^d}\big(f(x,\theta)^2 - 2f(x,\theta)\underbrace{\frac{p(x)-q(x)}{p(x)+q(x)}}_{f^*((x)} + 1\big)(p(x)+q(x))dx\right).\end{aligned}$$

Here we can notice that

$$\|f(\cdot,\theta) - f^*(\cdot)\|^2_{L^2(p+q)} = \int_{\mathbb{R}^d}\big(f(x,\theta)^2 - 2f(x,\theta)f^*(x) + (f^*(x))^2\big)(p(x)+q(x))dx.$$

If we add the constant

$$C = \frac{1}{2}\int_{\mathbb{R}^d} 4\frac{p(x)q(x)}{p(x)+q(x)}dx,$$

we get that

$$L(\theta) = \frac{1}{2}\|f(\cdot,\theta) - f^*(\cdot)\|^2_{L^2(p+q)} + C.$$

To see this, notice that

$$\begin{aligned}
&\frac{1}{2}\int_{\mathbb{R}^d}\Big((f^*(x))^2(p(x)+q(x))+\frac{4p(x)q(x)}{p(x)+q(x)}\Big)dx\\
=&\frac{1}{2}\int_{\mathbb{R}^d}\Big(\frac{(p(x)-q(x))^2}{(p(x)+q(x))^2}(p(x)+q(x))+\frac{4p(x)q(x)}{p(x)+q(x)}\Big)dx\\
=&\frac{1}{2}\int_{\mathbb{R}^d}\Big(\frac{p(x)^2-2p(x)q(x)+q(x)^2+4p(x)q(x)}{p(x)+q(x)}\Big)dx\\
=&\frac{1}{2}\int_{\mathbb{R}^d}\Big(\frac{(p(x)+q(x))^2}{p(x)+q(x)}\Big)dx=\frac{1}{2}\int_{\mathbb{R}^d}p(x)+q(x)dx.
\end{aligned}$$

This means that minimizing $L(\theta)$ is the same as minimizing $\|f-f^*\|^2_{L^2(p+q)}$ as the constant doesn't depend on θ. Importantly, this means that our target function in the population-level training regimes will be

$$f^*(x) := \frac{p(x)-q(x)}{p(x)+q(x)}.$$

5.3.3 Two-Sample Test

Given probability densities p and q, the two-sample test assesses whether to accept the null hypothesis H_0 or reject it for H_1, where

$$H_0 : p = q, \qquad\qquad H_1 : p \neq q.$$

In words, our test is constructed using the average output of the neural network on measure p minus the average output of the neural network on measure q. This will give either population-level two-sample tests or finite-sample two-sample tests on the datasets X_{test} and Z_{test}. In particular, for the population-level statistic, we can define

$$\mu_p(\theta) = \int_{\mathbb{R}^d} f(x,\theta)dp(x), \quad \mu_q(\theta) = \int_{\mathbb{R}^d} f(x,\theta)dq(x)$$
$$T(\theta; p,q) = (\mu_p(\theta) - \mu_q(\theta)) = \int_{\mathbb{R}^d} f(x,\theta)d(p-q)(x);$$

whereas, for the finite-sample statistic on *test* data, we can define

$$\mu_{\widehat{p}_{test}}(\theta) = \int_{\mathbb{R}^d} f(x,\theta)d\widehat{p}_{test}(x), \quad \mu_{\widehat{q}_{test}}(\theta) = \int_{\mathbb{R}^d} f(x,\theta)d\widehat{q}_{test}(x)$$
$$T(\theta; \widehat{p}_{test}, \widehat{q}_{test}) = (\mu_{\widehat{p}_{test}}(\theta) - \mu_{\widehat{q}_{test}}(\theta)) = \int_{\mathbb{R}^d} f(x,\theta)d(\widehat{p}_{test} - \widehat{q}_{test})(x).$$

Here, we define the neural network two-sample test for a neural network $f(\cdot,\theta)$ by $T(\theta;\widehat{p}_{test},\widehat{q}_{test})$. Given a test threshold $\tau > 0$, we reject the null hypothesis if $|T(\theta;\widehat{p}_{test},\widehat{q}_{test})| > \tau$. Moreover, we control the false discovery of the null by finding the smallest τ such that $\Pr[|T(\theta;\widehat{p}_{test},\widehat{q}_{test})| > \tau | H_0] \leq \alpha$, where $0 < \alpha < 1$ is the significance level. To find τ, we use a permutation test.

In Section 5.5, we will consider different training regimes and each of these training regimes will have different notions of the two-sample test, which change by what the output of the neural network is and which probability measures the two-sample test statistic is computed on. Particularly, the training regime with the zero-time NTK will end up using not the neural network by the function that is trained under zero-time NTK dynamics. The specific notation regarding the two-sample test will be discussed there.

5.4 Balancing time scales and network complexity

In this section, we consider the balancing of time scales for training and the role that the complexity of the neural network plays in training time.

5.4.1 Short-Time Requirements

Throughout this paper, we stress that we will work in the small time regime. The small time scale of the two-sample test is geared towards identifying the case when the alternative hypothesis is correct. Qualitatively, our results relate the time it takes for the neural network two-sample test to produce positive results to the complexity of the zero-time neural tangent kernel (NTK) and how the weighted difference of the densities $\frac{p-q}{p+q}(x)$ projects onto the zero-time NTK's eigenfunctions. We then relate the neural network's population-level dynamics as well as finite-sample dynamics to the zero-time NTK and bound their approximates by time. For the analysis done in this paper, we interplay between training for long enough that the zero-time NTK two-sample test performs well enough yet not too long that the approximations between the zero-time NTK and the other training regimes fail to hold. This interplay ensures that we are in a short-time regime although the details depend on the complexity of the problem and the complexity of the neural network.

To get a better idea of when the two-sample test works well, consider when the densities p and q are vastly different. Now assuming that the zero-time NTK's larger eigenvalue functions correspond to low frequency eigenfunctions, the neural network two-sample test should be able to produce a positive output with little time since $\frac{p-q}{p+q}$ will tend to be more low frequency than high frequency. On the other hand, if the densities p and q are quite close, then $\frac{p-q}{p+q}$ would tend to project onto higher frequency eigenfunctions which would take longer to detect. Detection in this case would require either need more time or a larger neural network to detect the difference between the densities. In the next section, we, moreover, discuss the interplay of the size of the neural network in producing a good two-sample test.

5.4.2 Complexity Scaling

Along with balancing the short-time versus long-time scale of the neural network two-sample test, we need to simultaneously balance the complexity of the neural network. The interplay of neural network complexity comes again in two places. Since we get a two-sample test from the zero-time NTK, we need the complexity of the neural network to be large enough to accurately capture $\frac{p-q}{p+q}$ on the eigenbasis of the NTK. The estimation of the zero-time NTK to the finite sample neural network dynamics, however, is bounded by how large the neural network is. Qualitatively, this means that smaller neural networks are approximated better with the zero-time NTK. This does not arise as much of an issue, however, because the factor of the neural network complexity is multiplied by time. This means that if the small time scale is small enough to counteract the loss in approximation from the size of the neural network, we still get good detection of the alternative hypothesis.

In our empirical results, we consider neural networks with varying parameter-to-sample ratios that range from the severely under-parameterized to highly over-parameterized regime.

5.5 The Three Training Regimes

We will consider the following three different training regimes for our neural network. For all of scenarios, however, we assume the following.

Assumption 5.2. *The neural network is initialized with parameters* θ_0 *such that* $f(x,\theta_0) = 0$ *for all* $x \in \mathbb{R}^d$.

5.5.1 Finite-sample time-varying dynamics

This regime is the most realistic as these are the dynamics that will arise in practice. In this regime, we tend to denote all the associated quantities with a hat. Since we use gradient descent to optimize, we denote the parameters trained from finite-sample data as $\widehat{\theta}(t)$ and will denote the associated neural network's output as $\widehat{u}(x,t) = f(x,\widehat{\theta}(t))$. Now, let us inspect the equation used to optimize the parameters.

$$\begin{aligned}
\widehat{L}(\theta) &= \frac{1}{2}\bigg(\int_{\mathbb{R}^d} \big(f(x,\theta)-1\big)^2 \widehat{p}(x)dx + \int_{\mathbb{R}^d} \big(f(x,\theta)+1\big)^2 \widehat{q}(x)dx \bigg) \\
-\dot{\widehat{\theta}}(t) &= \partial_\theta \widehat{L}(\widehat{\theta}(t)) \\
&= \frac{1}{2}\bigg(\int_{\mathbb{R}^d} \nabla_\theta f(x,\widehat{\theta}(t)) \Big(\big(f(x,\widehat{\theta}(t))-1\big)\widehat{p}(x) + \big(f(x,\widehat{\theta}(t))+1\big)\widehat{q}(x)\Big) dx \bigg) \\
\partial_t \widehat{u}(x,t) &= \langle \nabla_\theta f(x,\widehat{\theta}(t)), \dot{\widehat{\theta}} \rangle_\Theta = -\langle \nabla_\theta f(x,\widehat{\theta}(t)), \partial_\theta \widehat{L}(\widehat{\theta}(t)) \rangle_\Theta \\
&= -\frac{1}{2}\bigg(\int_{\mathbb{R}^d} \langle \nabla_\theta f(x,\widehat{\theta}(t)), \nabla_\theta f(x',\widehat{\theta}(t)) \rangle_\Theta \Big(\big(f(x',\widehat{\theta}(t))-1\big)\widehat{p}(x') \\
&\quad + \big(f(x',\widehat{\theta}(t))+1\big)\widehat{q}(x')\Big) dx' \bigg).
\end{aligned}$$

We define the time-varying finite-sample neural tangent kernel by

$$\widehat{K}_t(x,x') = \langle \nabla_\theta f(x,\widehat{\theta}(t)), \nabla_\theta f(x',\widehat{\theta}(t)) \rangle_\Theta.$$

We can further define density-specific residuals

$$\begin{aligned}
\widehat{e}_p(x,t) &= \big(f(x,\widehat{\theta}(t))-1\big) \\
\widehat{e}_q(x,t) &= \big(f(x,\widehat{\theta}(t))+1\big).
\end{aligned}$$

This means that

$$\begin{aligned}
\partial_t \widehat{u}(x,t) &= -\frac{1}{2}\bigg(\int_{\mathbb{R}^d} \widehat{K}_t(x,x') \Big(\big(f(x',\widehat{\theta}(t))-1\big)\widehat{p}(x') + \big(f(x',\widehat{\theta}(t))+1\big)\widehat{q}(x')\Big) dx' \bigg) \\
&= -\frac{1}{2}\Big(\mathbb{E}_{x'\sim\widehat{p}} \widehat{K}_t(x,x') \widehat{e}_p(x',t) + \mathbb{E}_{x'\sim\widehat{q}} \widehat{K}_t(x,x') \widehat{e}_q(x',t) \Big).
\end{aligned}$$

In the context of the two sample test, since this training regime uses only finite training samples, we will study the two-sample test statistic's behavior evaluated on the training samples, the test samples (*independent* from the training samples), and the population. We denote these different evaluated test statistics by

$$\begin{aligned}
\widehat{T}_{train}(t) &:= T(\widehat{\theta}(t);\widehat{p},\widehat{q}) = \Big(\mathbb{E}_{x\sim\widehat{p}} - \mathbb{E}_{x\sim\widehat{q}}\Big)f(x,\widehat{\theta}(t)) = \Big(\mathbb{E}_{x\sim\widehat{p}} - \mathbb{E}_{x\sim\widehat{q}}\Big)\widehat{u}(x,t),\\
\widehat{T}_{test}(t) &:= T(\widehat{\theta}(t);\widehat{p}_{test},\widehat{q}_{test}) = \Big(\mathbb{E}_{x\sim\widehat{p}_{test}} - \mathbb{E}_{x\sim\widehat{q}_{test}}\Big)f(x,\widehat{\theta}(t))\\
&= \Big(\mathbb{E}_{x\sim\widehat{p}_{test}} - \mathbb{E}_{x\sim\widehat{q}_{test}}\Big)\widehat{u}(x,t),\\
\widehat{T}_{pop}(t) &:= T(\widehat{\theta}(t);p,q) = \Big(\mathbb{E}_{x\sim p} - \mathbb{E}_{x\sim q}\Big)f(x,\widehat{\theta}(t)) = \Big(\mathbb{E}_{x\sim p} - \mathbb{E}_{x\sim q}\Big)\widehat{u}(x,t),
\end{aligned}$$

where $\widehat{T},\widehat{T}_{test},\widehat{T}_{pop}$ denote the evaluation on the training samples, test samples, and population, respectively.

5.5.2 Population-level time-varying dynamics

This regime is essentially what would happen as the number of samples in our datasets grows larger and larger. In this case, we denote the path of the parameters as simply $\theta(t)$ and will denote the associated neural network's output as $u(x,t) = f(x,\theta(t))$. We showed earlier that the population-level loss is equivalent to simply minimizing $\mathcal{L}(\theta) = \frac{1}{2}\|f(\cdot,\theta) - f^*(\cdot)\|_{L^2(p+q)}$. This means that training the population-level neural network is equivalent to running gradient descent on $\mathcal{L}$. Moreover, we can define the population level error function as $e(x,t) = f(x,\theta(t)) - f^*(x)$. Using these facts, we get the following

$$\begin{aligned}
-\dot{\theta}(t) &= \partial_\theta\mathcal{L}(\theta(t)) = \frac{1}{2}\int_{\mathbb{R}^d}\nabla_\theta f(x,\theta(t))\big(f(x,\theta(t)) - f^*(x)\big)(p+q)(x)dx\\
\partial_t u(x,t) &= \langle\nabla_\theta f(x,\theta(t)),\dot{\theta}(t)\rangle_\Theta = -\langle\nabla_\theta f(x,\theta(t)),\partial_\theta\mathcal{L}(\theta(t))\rangle_\Theta\\
&= -\frac{1}{2}\int_{\mathbb{R}^d}\langle\nabla_\theta f(x,\theta(t)),\nabla_\theta f(x',\theta(t))\rangle_\Theta\big(f(x',\theta(t)) - f^*(x')\big)(p+q)(x')dx'.
\end{aligned}$$

Now we define the population level time-varying neural tangent kernel

$$K_t(x,x') = \langle \nabla_\theta f(x,\theta(t)), \nabla_\theta f(x',\theta(t)) \rangle_\Theta.$$

This finally implies that

$$\partial_t u(x,t) = -\frac{1}{2}\mathbb{E}_{x'\sim p+q} K_t(x,x') e(x',t) = \partial_t e(x,t).$$

Contrary to the finite-sample training regime, we only care about the population-level two-sample test statistic in this training regime since this training regime itself uses the population to train on. We denote the two-sample test statistic associated to the population-level time-varying dynamics by

$$T(t) := T(\theta(t); p,q) = \int_{\mathbb{R}^d} u(x,t) d(p-q)(x),$$

where we are integrating the neural network output over the entire densities.

5.5.3 Population-level zero-time kernel dynamics

In this training regime, we will denote related quantities with a bar so that the output of the trained function here becomes $\bar{u}(x,t)$ with $\bar{u}(x,0) = f(x,\theta_0)$. At this point, consider the zero-time NTK of the neural network $f(\cdot,\theta_0)$ by

$$K_0(x,x') = \langle \nabla_\theta f(x,\theta_0), \nabla_\theta f(x',\theta_0) \rangle_\Theta.$$

According to [81, Lemma 4.3], assuming that $\|\nabla_\theta f(x,\theta_0)\|_\Theta$ is squared integrable on $\mathbb{R}^d$ against measure $(p(x)+q(x))dx$, the zero-time kernel can act as a kernel integral operator and admits a spectral decomposition, which we can write as

$$(L_{K_0}g)(x) = \int_{\mathbb{R}^d} K_0(x,x')g(x')(p+q)(x')dx' = \sum_{\ell=1}^{M} \lambda_\ell \langle g, u_\ell \rangle_{L^2(p+q)} u_\ell(x),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$. Although we can find an extended basis for $\ell > M$ for $L^2(p+q)$, the associated eigenvalues of K_0 are 0 on eigenfunctions u_ℓ for $\ell > M$. Since K_0 does not effectively have full basis for $L^2(p+q)$, the quantities that we work with will need to be projected onto the range of the operator L_{K_0}. This motivates the following definition.

Definition 5.3. *Denote the projection operator onto the range of L_{K_0} by Π_{K_0}.*

Now we can define the associated error function

$$\bar{e}(x,t) = \bar{u}(x,t) - \Pi_{K_0}(f^*)(x).$$

Since f^* is fixed, the dynamics of our model that we care about will be given by

$$\partial_t \bar{u}(x,t) = -\frac{1}{2}\mathbb{E}_{x' \sim p+q} K_0(x,x')\bar{e}(x',t) = \partial_t \bar{e}(x,t).$$

With the simplicity in the model dynamics, we can attain better analysis. Using this interpretation, we know that

$$\partial_t \bar{e}(\cdot,t) = -\frac{1}{2}(L_{K_0}\bar{e}(\cdot,t)) = -\sum_{\ell=1}^{M} \lambda_\ell \langle \bar{e}(\cdot,t), u_\ell \rangle_{L^2(p+q)} u_\ell.$$

We formulate an ansatz of what $\bar{e}(\cdot,t)$ would be so that it satisfies this differential equation. In particular, consider

$$\bar{e}(\cdot,t) = \sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot,0) \rangle_{L^2(p+q)} u_\ell.$$

The following proposition ensures that this indeed is a solution of the differential equation of interest.

Proposition 5.4. *The solution*

$$\bar{e}(\cdot,t) = \sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot,0)\rangle u_\ell$$

solves the differential equation

$$\partial_t \bar{e}(\cdot,t) = -\frac{1}{2}(L_{K_0}\bar{e}(\cdot,t)) = -\sum_{\ell=1}^{M} \lambda_\ell \langle \bar{e}(\cdot,t), u_\ell\rangle_{L^2(p+q)} u_\ell.$$

The proof of Proposition 5.4 is given in Section 5.9.

For this training regime, we can now talk about the two-sample test statistic. In particular, recalling that $\bar{u} = \bar{e} + \Pi_{K_0}(f^*)$, we will set

$$\begin{aligned}
\overline{\mu_p}(t) &= \int_{\mathbb{R}^d} \bar{u}(x,t)dp(x) = \int_{\mathbb{R}^d} (\bar{e}(x,t) + \Pi_{K_0}(f^*)(x))p(x)dx, \\
\overline{\mu_p}(t) &= \int_X \bar{u}(x,t)dq(x) = \int_X (\bar{e}(x,t) + \Pi_{K_0}(f^*)(x))q(x)dx.
\end{aligned}$$

Similar to the population-level time-varying dynamics, we only care about the population-level two-sample test statistic for this training regime since we use the entire population for training. For this training regime, we will denote the associated two-sample test by

$$\overline{T}(t) := \overline{\mu_p}(t) - \overline{\mu_q}(t) = \int_{\mathbb{R}^d} \bar{u}(x,t)d(p-q)(x).$$

Note that in this case, $\overline{T}(t)$ is not determined by the parameters of the neural network changing but rather by the output of the NTK trained dynamics. With this in mind, we get the following lemma.

Lemma 5.5. *The population-level zero-time kernel dynamics two-sample test statistic is given by*

$$\overline{T}(t) = \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \sum_{\ell \geq 1} e^{-t\lambda_\ell} \langle u_\ell, \Pi_{K_0}(f^*)\rangle^2_{L^2(p+q)}.$$

The proof of Lemma 5.5 is in Section 5.9.

Note that there is some time-analysis we can undertake at this point. First, you can notice that at $t = 0$,

$$\overline{T}(0) = \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \underbrace{\sum_{\ell \geq 1} \langle u_\ell, \Pi_{K_0}(f^*)\rangle^2_{L^2(p+q)}}_{\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}} = 0,$$

and for any time $t > 0$, we have $\overline{T}(t) > 0$. We will find a theoretical minimum time $t(\varepsilon)$ such that $\overline{T}(t(\varepsilon)) \geq \varepsilon$. Let us define a few quantities before delving into the main result.

Definition 5.6. *Let $S \subseteq \{1, \ldots, M\}$, then we can consider how much of the norm $\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}$ (and hence the norm of f^*) lies on the eigenbasis subset $V_S = \{u_\ell\}_{\ell \in S}$. In particular, we have*

$$\|\Pi_{K_0}(f^*)\|^2_S = \|f^*\|^2_S = \sum_{\ell \in S} \langle u_\ell, f^*\rangle^2_{L^2(p+q)} = x_S \|f^*\|^2_{L^2(p+q)}.$$

Since $\langle u_\ell, \Pi_{K_0}(f^*)\rangle_{L^2(p+q)} = \langle u_\ell, f^*\rangle_{L^2(p+q)}$ for $\ell \leq M$, we can also define these quantities for f^* rather than just $\Pi_{K_0}(f^*)$. Moreover, we can define the minimum and maximum eigenvalue that exist for the eigenvectors that lie in V_S by defining

$$\lambda_{\min}(S) = \min_{\ell \in S} \lambda_\ell, \quad \lambda_{\max}(S) = \max_{\ell \in S} \lambda_\ell.$$

Then we have the following theorem.

Theorem 5.7. *Let* $\varepsilon > 0$ *and assume that there exists a finite subset* $S \subset \{1,\ldots,M\}$ *such that*

$$\|\Pi_{K_0}(f^*)\|_S^2 = \|f^*\|_S^2 = \sum_{\ell \in S} \langle u_\ell, f^* \rangle_{L^2(p+q)}^2 > \varepsilon$$

and $\lambda_{\min}(S) > 0$*. Then*

$$t(\varepsilon) \geq \lambda_{\min}(S) \log\left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_S^2 - \varepsilon}\right) = \lambda_{\min}(S) \log\left(\frac{\|f^*\|_S^2}{\|f^*\|_S^2 - \varepsilon}\right)$$

ensures that $\overline{T}(t(\varepsilon)) \geq \varepsilon$*.*

The proof of Theorem 5.7 is in Section 5.9.

Remark 5.8. *Let us analyze the function*

$$g(\varepsilon, S) = \min_{S \in \mathcal{S}_1(\varepsilon)} \lambda_{\min}(S) \log\left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_S^2 - \varepsilon}\right).$$

Notice first that the largest that ε *can be is* $\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2$ *because as* $t \to \infty$*, we get that* $\overline{T}(t) \to \|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2$ *and it is easy to see that* $\overline{T}(t)$ *is monotonic in* t*. Now notice that as* ε *gets larger, we need S to satisfy* $\|f^*\|_S^2 = \|\Pi_{K_0}(f^*)\|_S^2 > \varepsilon$ *to make sure that* $g(\varepsilon, S)$ *is well-defined. Moreover, we need* $\lambda_{\min}(S) > 0$ *otherwise we find that* $g(\varepsilon, S) = 0$ *which is the trivial bound. In particular, we will want the fraction* $\varepsilon / \|\Pi_{K_0}(f^*)\|_S^2$ *to be as small as possible to give the smallest possible non-trivial time.*

We have an analogous statement for when we want our test statistic to be less than ε. In particular, we get that

Theorem 5.9. *Let* $\varepsilon > 0$ *and assume that there exists a finite subset* $S \subset \mathbb{N}$ *such that*

$$\frac{\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 - \varepsilon}{\|\Pi_{K_0}(f^*)\|_S^2} > 0$$

and $\lambda_{\max}(S) > 0$. *Then*

$$t(\varepsilon) \leq \lambda_{\max}(S) \log \left\{ \frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 - \varepsilon} \right\}$$

ensures that $\overline{T}(t(\varepsilon)) \leq \varepsilon$.

The proof of Theorem 5.9 is in Section 5.9. Now if we optimize over all such subsets S, we get the following corollary.

Corollary 5.10. *Let* $0 < \varepsilon < \|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2$. *Assume that the set*

$$\begin{aligned} &\mathcal{S}_1(\varepsilon) = \{S \subset \mathbb{N} : \|\Pi_{K_0}(f^*)\|_S^2 > \varepsilon, \lambda_{\min}(S) > 0\} \neq \emptyset \\ &\mathcal{S}_2(\varepsilon) = \{S \subset \mathbb{N} : (\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 - \varepsilon)/\|\Pi_{K_0}(f^*)\|_S^2 > 0, \lambda_{\max}(S) > 0\} \neq \emptyset. \end{aligned}$$

Then

$$t \geq t_1^*(\varepsilon) := \min_{S \in \mathcal{S}_1(\varepsilon)} \lambda_{\min}(S) \log \left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_S^2 - \varepsilon} \right)$$

ensures that $\overline{T}(t) \geq \varepsilon$ *whilst*

$$t \leq t_2^*(\varepsilon) := \max_{S \in \mathcal{S}_2(\varepsilon)} \lambda_{\max}(S) \log \left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 - \varepsilon} \right)$$

ensures that $\overline{T}(t) \leq \varepsilon$.

Here let us remark what occurs in the case when our null hypothesis is correct versus when the alternative is correct.

Remark 5.11. *If* H_0 *is true (so that* $p = q$*), then* $f^* = 0$. *We can't apply the theorem above then since the assumption is not satisfied; however, we note by inspection that* $\overline{T}(t) = 0$ *for all* t. *If* $\|\Pi_{K_0}(f^*)\|_{L^2(p+q)} < \delta$ *for small* $\delta > 0$, *then note that both Theorem 5.7 and Theorem 5.9 limit*

$\varepsilon < \delta$. *This means that* $\overline{T}(t)$ *can only detect small changes. On the other hand, if we are under* H_1 *(so that* $p \neq q$*) and we assume that* $\|\Pi_{K_0}(f^*)\|_{L^2(p+q)} > \delta$ *for some larger* $\delta > 0$*, then* ε *can be made much larger and should be more easy to detect.*

5.6 Analysis of u with $\bar{u}$

Let B_R denote the open ball of radius R with center θ_0 and assume that $u(x,0) = \bar{u}(x,0) = f(x,\theta_0) = 0$ for all x. For much of the analysis going forward, we will use the following lemma heavily.

Lemma 5.12. $\mathcal{L}(\theta(0)) = \|u(\cdot,0) - f^*\|^2_{L^2(p+q)} = \|\bar{u}(\cdot,0) - f^*\|^2_{L^2(p+q)} = \|f^*\|^2_{L^2(p+q)}$.

Proof. Notice that since $u(\cdot,0) = \bar{u}(\cdot,0) = f(x,\theta_0) = 0$, we have the result. □

To continue, we will need to assume the following assumptions

Assumption 5.13. *There exists positive constants* R, L_1*, and* L_2 *such that*

1. *(Boundedness) For any* $\theta \in B_R$*,* $\sup_{x \in \text{supp}(p+q)} \|\nabla_\theta f(x,\theta)\| \leq L_1$.

2. *(Lipschitz) For any* $\theta_1, \theta_2 \in B_R$*,* $\sup_{x \in \text{supp}(p+q)} \|\nabla_\theta f(x,\theta_1) - \nabla_\theta f(x,\theta_2)\| \leq L_2\|\theta_1 - \theta_2\|$.

5.6.1 Approximation

Next we apply these assumptions to gain the following proposition.

Proposition 5.14. *Assume that* $u(x,0) = \bar{u}(x,0) = 0$*, then*

$$\|\theta(t) - \theta(0)\| \leq \sqrt{t}\|f^*\|_{L^2(p+q)}.$$

Moreover, if

$$t \leq \left(\frac{R}{\|f^*\|_{L^2(p+q)}} \right)^2,$$

then $\theta(t) \in B_R$*.*

The proof of Proposition 5.14 is contained in Section 5.10. Now we can further bound the operator norm of the difference $K_t - K_0$ with the following lemma.

Proposition 5.15. *Let* $\theta(t) \in B_R$*, then under Assumption 5.13, we have*

$$\|K_t - K_0\|_{L^2(p+q)} \leq 2L_1L_2\sqrt{t}\|f^*\|_{L^2(p+q)}.$$

The proof of Proposition 5.15 is contained in Section 5.10. Now we use this result for bounding the difference $\|u - \bar{u}\|_{L^2(p+q)}$. In particular, we have the following proposition.

Proposition 5.16. *Under Assumption 5.13 and*

$$t \leq \left(\frac{R}{\|f^*\|_{L^2(p+q)}} \right)^2,$$

we get

$$\|(u - \bar{u})(\cdot, t)\|_{L^2(p+q)} = \|(e - \bar{e})(\cdot, t)\|_{L^2(p+q)} \leq \frac{8}{3}L_1L_2\|f^*\|^2_{L^2(p+q)}(t)^{3/2}.$$

The proof of Proposition 5.16 is contained in Section 5.10. Now let us extend our zero-time NTK two-sample test results to the population-level time-varying kernel two-sample test. We first notice the following corollary.

Corollary 5.17. *Under Assumption 5.13 and*

$$t \leq \left(\frac{R}{\|f^*\|_{L^2(p+q)}} \right)^2,$$

we have

$$\left| T(t) - \overline{T}(t) \right| \leq \sqrt{2}\frac{8}{3} L_1 L_2 \|f^*\|^2_{L^2(p+q)} (t)^{3/2}.$$

The proof of Corollary 5.17 is contained in Section 5.10. For further time analysis in the alternative hypothesis case below, we will need to show that this two-sample test $T(t)$ is monotonically increasing. We will be able to show this if our population-level neural network has increasing norm. This assumption is not unsupported since we initialize as $u(x,0) = f(x,\theta(0)) = 0$ and our target function $f^*(x) = \frac{p-q}{p+q}(x)$ has non-zero norm. Using this assumption, we get the following theorem (with proof contained in Section 5.10).

Theorem 5.18. *Assume that* $\|u(x,t)\|_{L^2(p+q)}$ *is monotonically increasing on the interval* $[0,\tau]$, *then* $T(t)$ *is monotonically increasing on* $[0,\tau]$.

5.6.2 Time analysis of u

Given all the approximations done before, we can use the time analysis done with $\bar{u}$ and apply it to u with the correct approximations. In essence, we can use Corollary 5.17 along with Corollary 5.10 to get the following two theorems.

In order to counteract the time-dependent estimation error shown in Corollary 5.17, the next theorem, which is geared towards discovering the alternative hypothesis, necessarily assumes first that the minimum time needed to detect an error ε in Corollary 5.10 is smaller than ε and second that the time scale we work on is valid for detection. Note that as the size of the neural network grows, the minimum time needed for detection decreases but L_1 and L_2 below increase;

thus, there is an interplay of making sure your neural network is large but not too large.

For this next theorem, recall from Corollary 5.10 that

$$
\begin{aligned}
t_1^*(\varepsilon) &:= \min_{S \in \mathcal{S}_1(\varepsilon)} \lambda_{\min}(S) \log\left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_S^2 - \varepsilon} \right), \\
t_2^*(\varepsilon) &:= \max_{S \in \mathcal{S}_2(\varepsilon)} \lambda_{\max}(S) \log\left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 - \varepsilon} \right).
\end{aligned}
$$

Theorem 5.19. *Let $\varepsilon > 0$ and assume that $\|u(x,t)\|_{L^2(p+q)}^2$ is monotonically increasing on $[0,\tau]$ and that*

$$
\frac{8\sqrt{2}}{3} \|f^*\|_{L^2(p+q)}^2 L_1 L_2 \big(t_1^*(\varepsilon)\big)^{3/2} < \varepsilon
$$

as well as

$$
\min\left(\tau, \Big(\frac{R}{\|f^*\|_{L^2(p+q)}}\Big)^2\right) \geq t_1^*(\varepsilon)
$$

where $t_1^*(\varepsilon)$ is defined in Corollary 5.10. Then under Assumption 5.13 and for

$$
\min\left(\tau, \Big(\frac{R}{\|f^*\|_{L^2(p+q)}}\Big)^2\right) \geq t \geq t_1^*(\varepsilon),
$$

we get

$$
|T(t)| \geq \varepsilon - \frac{8\sqrt{2}}{3} \|f^*\|_{L^2(p+q)}^2 L_1 L_2 \big(t_1^*(\varepsilon)\big)^{3/2}.
$$

We prove Theorem 5.19 in Section 5.10. Now, the following theorem is useful in showing the null hypothesis and necessarily needs the time to be smaller the maximum time needed to detect ε as well as the time needed to stay in B_R (so that Proposition 5.16 holds).

Theorem 5.20. *Let* $\varepsilon > 0$*. Under Assumption 5.13 and for*

$$t \leq \min\left\{ \Big(\frac{R}{\|f^*\|_{L^2(p+q)}} \Big)^2, t_2^*(\varepsilon) \right\},$$

where $t_2^*(\varepsilon)$ *is defined in Corollary 5.10, we have*

$$|T(t)| \leq \varepsilon + \frac{8\sqrt{2}}{3} \|f^*\|_{L^2(p+q)}^2 L_1 L_2 \big(t_2^*(\varepsilon)\big)^{3/2}$$

The proof of Theorem 5.20 is given in Section 5.10. In the next section, we will consider how to bound $\|\bar{u} - \widehat{u}\|_{L^2(p+q)}$ as $\widehat{u}$ represents finite-sample behavior.

5.7 Analysis of $\widehat{u}$ with $\bar{u}$

Since we will be using finite-samples, we will use some concentration inequalities and will need a few extra assumptions. To start off, recall that for our finite-sample we have n_p training samples from density p and n_q samples from density q. Moreover, recall that our finite-sample loss function is given by

$$\widehat{L}(\theta) = \frac{1}{2}\bigg(\int_{\mathbb{R}^d} \big(f(x,\theta) - 1\big)^2 \widehat{p}(x) dx + \int_{\mathbb{R}^d} \big(f(x,\theta) + 1\big)^2 \widehat{q}(x) dx \bigg)$$

We first show approximation of the raw dynamics and then approximation with the time-analysis.

5.7.1 Approximation

Let us bound $\|\widehat{\theta}(t) - \widehat{\theta}(0)\|_\Theta$. To this end, we get the following lemma.

Lemma 5.21. *Assume that* $\widehat{\theta}(0) = \theta(0)$ *and that* $f(x, \theta(0)) = 0$*, then*

$$\|\widehat{\theta}(t) - \theta(0)\|_{\Theta} \leq \sqrt{t}.$$

Moreover, if $\theta(0) \in B_R$*, then*

$$t \leq R^2$$

ensures that $\widehat{\theta}(t) \in B_R$*.*

We prove Lemma 5.21 in Section 5.11. Now we will need the following assumption to proceed.

Assumption 5.22. *For* $A > 0$*, consider the function*

$$h(n) = \sqrt{2L_1^2(2L_1^2 + 3/2)\frac{A\log(n) + \log(2M_{\Theta})}{n}}.$$

Assume that n_p *and* n_q *are large enough that* $h(n_p) < \frac{3}{2}$ *and* $h(n_q) < \frac{3}{2}$*.*

Using this assumption, we can apply Theorem 5.40 to get the following lemma to be used later.

Lemma 5.23. *Assume that* $\theta \in B_R$*, then under Assumption 5.13 and Assumption* 5.22 *and let* p *be a probability density, consider the random* M_{Θ}*-by-*M_{Θ} *matrix*

$$X_i = \nabla_{\theta} f(x_i, \theta) \nabla_{\theta} f(x_i, \theta)^{\top} - \mathbb{E}_{x \sim p} \nabla_{\theta} f(x, \theta) \nabla_{\theta} f(x, \theta)^{\top}.$$

If n *is the number of samples from* p*, then with probability greater than* $1 - n^{-A}$*, we have*

$$\|\frac{1}{n}\sum_{i=1}^{n} X_i\| \leq \sqrt{2L_1^2(2L_1^2 + 3/2)\frac{A\log(n) + \log(2M_{\Theta})}{n}}.$$

The proof of Lemma 5.23 is in Section 5.11 and is used in the following proposition.

Proposition 5.24. *Assume that $t \le R^2$ (so that $\theta(t) \in B_R$) as well as Assumption 5.13 and Assumption 5.22, then with probability $\ge 1 - n_p^{-A} - n_q^{-A}$, we have*

$$\|(\widehat{u} - \bar{u})(\cdot, t)\|_{L^2(p+q)} \le C_1 t + C_2 t^{3/2} + C_3 t^2 + C_4 t^{5/2},$$

where the dependence of the constants is given by $C_1 = C(L_1)$, $C_2 = C(L_1, L_2, f^*)$, $C_3 = C(L_1, f^*, n_p, n_q, M_\Theta, A)$, and $C_4 = C(L_1, L_2, f^*, n_p, n_q, M_\Theta, A)$.

Note that the more technical version of Proposition 5.24 is contained in Proposition 5.36 along with its proof. We will use Proposition 5.24 to show that the finite-sample two-sample test statistic and zero-time kernel population-level two-sample test statistic are close for $\widehat{T}_{pop}$, $\widehat{T}_{train}$, and $\widehat{T}_{test}$ (i.e. the evaluation of the finite-sample two-sample test statistic on the population, training samples, and test samples, respectively). For $\widehat{T}_{pop}$, we get the following proposition.

Proposition 5.25. *Assume the conditions of Proposition* 5.24*, then with probability* $\ge 1 - (n_p^{-A} + n_q^{-A})$*, we get the time-approximation error function*

$$\left|\widehat{T}_{pop}(t) - \overline{T}(t)\right| \le C_1 t + C_2 t^{3/2} + C_3 t^2 + C_4 t^{5/2} := \delta_{pop}(t),$$

where C_1, C_2, C_3, C_4 are exactly the constants from Proposition 5.24. Moreover, note that this error function is monotonic.

Proof of Proposition 5.25. Mimic the proof of Corollary 5.17 *mutatis mutandis* applying Proposition 5.24. □

Now for $\widehat{T}_{test}(t)$, the test size sample sizes come into play. Recall that since

$$\widehat{T}_{test}(t) = T(\widehat{\theta}(t); \widehat{p}_{test}, \widehat{q}_{test}) = \left(\mathbb{E}_{x\sim\widehat{p}_{test}} - \mathbb{E}_{x-\widehat{q}_{test}}\right)\widehat{u}(x,t),$$

with test sample sizes m_p and m_q for $\widehat{p}_{test}$ and $\widehat{q}_{test}$, respectively. We also describe the time-approximation error function in the next proposition, which will be used for the theorems and corollaries afterwards.

Proposition 5.26. *Assume the conditions of Proposition 5.24, then with probability* $\geq 1-(m_p^{-A}+m_q^{-A})$*, we have*

$$|\widehat{T}_{test}(t)-\widehat{T}_{pop}(t)| \leq L_1^2 t\sqrt{2}\left(\sqrt{\frac{A\log(m_p)}{m_p}}+\sqrt{\frac{A\log(m_q)}{m_q}}\right).$$

Moreover, with probability $\geq 1-(m_p^{-A}+m_q^{-A}+n_p^{-A}+n_q^{-A})$*, we get the time-approximation error function*

$$\left|\widehat{T}_{test}(t)-\overline{T}(t)\right| \leq \tilde{C}_1 t + C_2 t^{3/2} + C_3 t^2 + C_4 t^{5/2} := \delta(t),$$

where $\tilde{C}_1 = C(L_1, A, m_p, m_q)$ *and* C_2, C_3, C_4 *are exactly the constants from Proposition 5.24. Finally, note that this error function* $\delta(t)$ *is monotonic.*

The proof of this proposition is located in Section 5.11.

Remark 5.27. *We note that Proposition 5.26 works for* $\widehat{T}_{train}$ *if we replace* m_p *and* m_q *with* n_p *and* n_q *respectively. Since the error function* δ *depends on whether we use test samples or training samples, we will regard the error function by* $\delta_{test}(t)$ *and* $\delta_{train}(t)$ *to distinguish these cases. In particular, the only constant that is different in* δ_{train} *and* δ_{test} *is* $\tilde{C}_1$*, where we change* m_p *and* m_q *to* n_p *and* n_q*, respectively. Moreover, using the triangle inequality, we can see that*

$$\begin{aligned}|\widehat{T}_{test}(t)-\widehat{T}_{train}(t)| \leq L_1^2 t\sqrt{2}\Bigg(&\sqrt{\frac{A\log(m_p)}{m_p}}+\sqrt{\frac{A\log(n_p)}{n_p}}\\ &+\sqrt{\frac{A\log(m_q)}{m_q}}+\sqrt{\frac{A\log(n_q)}{n_q}}\Bigg).\end{aligned}$$

Finally, we may also deduce from Proposition 5.26 that

$$|\widehat{T}_{train}(t) - \widehat{T}_{pop}(t)| \leq L_1^2 t\sqrt{2}\left(\sqrt{\frac{A\log(n_p)}{n_p}} + \sqrt{\frac{A\log(n_q)}{n_q}}\right).$$

To do further time-analysis in this finite-sample training case, we will need that $\widehat{T}_{train}(t)$ is monotonic in time. We will then use sampling concentration of $\widehat{T}_{train}$ with $\widehat{T}_{pop}$ and $\widehat{T}_{test}$ to extend the two-sample test statistic to these two different evaluation settings.

Theorem 5.28. *Assume that there is an interval $[0,\widehat{\tau}]$ such that $|\widehat{u}(x,s)| \leq 1$ for $s \in [0,\widehat{\tau}]$. Then $\widehat{T}_{train}(t)$ is monotonically increasing on $[0,\widehat{\tau}]$.*

We include the proof of Theorem 5.28 in Section 5.11.

Remark 5.29. *Note that the assumption $|\widehat{u}(x,s)| \leq 1$ definitely holds for at least small time intervals since the training dynamics are smooth and $\widehat{u}(x,0) = 0$. Moreover, we crucially use the fact that the training loss is decreasing for the proof of Theorem 5.28.*

Now similar to the case of the time-analysis theorems for u and $\bar{u}$, we get the following extensions of the zero-time NTK time-analysis theorems. Again for the next theorem, we must assume that the error detection level is greater than the time-valued approximation error of $\bar{u}$ with $\widehat{u}$. Recall from Corollary 5.10 the ε-detection time thresholds for the zero-time NTK two-sample test given by

$$t_1^*(\varepsilon) := \min_{S \in \mathcal{S}_1(\varepsilon)} \lambda_{\min}(S) \log\left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_S^2 - \varepsilon}\right),$$
$$t_2^*(\varepsilon) := \max_{S \in \mathcal{S}_2(\varepsilon)} \lambda_{\max}(S) \log\left(\frac{\|\Pi_{K_0}(f^*)\|_S^2}{\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 - \varepsilon}\right).$$

Again, making sure that our time scale lies in the correct regimes, we get the following theorems.

Theorem 5.30. *Let $\varepsilon > 0$. Along with the assumptions of Proposition 5.26 and Theorem 5.28, assume $\max(R^2,\widehat{\tau}) \geq t \geq t_1^*(\varepsilon)$, then*

1. *with probability* $\geq 1 - 2(n_p^{-A} + n_q^{-A})$,

$$|\widehat{T}_{train}(t)| \geq \varepsilon - \delta_{train}(t_1^*(\varepsilon)),$$

2. *with probability* $\geq 1 - (n_p^{-A} + n_q^{-A} + m_p^{-A} + m_q^{-A})$,

$$|\widehat{T}_{test}(t)| \geq \varepsilon - \delta_{train}(t_1^*(\varepsilon)) - L_1^2 t\sqrt{2}\Bigg(\sqrt{A\log(m_p)/m_p} + \sqrt{A\log(m_q)/m_q} \\ + \sqrt{A\log(n_p)/n_p} + \sqrt{A\log(n_q)/n_q}\Bigg),$$

3. *with probability* $\geq 1 - (n_p^{-A} + n_q^{-A})$,

$$|\widehat{T}_{pop}(t)| \geq \varepsilon - \delta_{train}(t_1^*(\varepsilon)) - L_1^2 t\sqrt{2}\Bigg(\sqrt{A\log(n_p)/n_p} + \sqrt{A\log(n_q)/n_q}\Bigg),$$

where the approximation error function $\delta_{train}(t)$ *comes from Proposition 5.26 with training samples and* $t_1^*(\varepsilon)$ *from Corollary 5.10. Moreover, if* ε *is not large enough to make the right-hand sides of the inequalities positive, the bounds are vacuous.*

Since we don't need monotonicity for the other case in Corollary 5.10 because we use the regular triangle inequality, the following theorem holds for each of $\widehat{T}_{train}, \widehat{T}_{test}$, and $\widehat{T}_{pop}$ with their respective time-approximation error functions.

Theorem 5.31. *Let* $\varepsilon > 0$ *and assume*

$$t \leq \min\left\{R^2, t_2^*(\varepsilon)\right\},$$

where $t_2^*(\varepsilon)$ *is defined in Corollary 5.10. Then*

1. *with probability* $\geq 1 - 2(n_p^{-A} + n_q^{-A})$*, we have*

$$|\widehat{T}_{train}(t)| \leq \varepsilon + \delta_{train}(t_2^*(\varepsilon)),$$

2. *with probability* $\geq 1 - (n_p^{-A} + n_q^{-A} + m_p^{-A} + m_q^{-A})$*, we have*

$$|\widehat{T}_{test}(t)| \leq \varepsilon + \delta_{test}(t_2^*(\varepsilon)),$$

3. *with probability* $\geq 1 - (n_p^{-A} + n_q^{-A})$*, we have*

$$|\widehat{T}_{pop}(t)| \leq \varepsilon + \delta_{pop}(t_2^*(\varepsilon)).$$

We include the proofs of both Theorem 5.30 and Theorem 5.31 in Section 5.11.

Now, given the more concrete setting of f^* lying on the first k eigenfunctions of K_0, we want to see if the time it takes to detect a desired deviation level $\varepsilon > 0$ is larger whether we are in the null hypothesis or in the first k eigenfunction assumption. The problem becomes slightly complex since there is a time-approximate error term in the deviation that comes from Proposition 5.26. Since this assumption is not exactly the logical complement of the null, let us define the setting more concretely.

Definition 5.32. *If $\Pi_{K_0}(f^*)$ nontrivially projects* ***only*** *onto the first k eigenfunctions of K_0 holds true, we denote the projected target function on the first k eigenfunctions as $\Pi_{K_0}(f^*) = f_k^*$. We denote the test statistic when $f_k^* \neq 0$ by $\widehat{T}_{train,k}(t), \widehat{T}_{test,k}(t), \widehat{T}_{pop,k}(t)$ evaluated on the training set, test set, and population, respectively, and when the evaluation set is understood from the context, we use $\widehat{T}_k(t)$. If $p = q$, we say the null hypothesis holds. We denote the test statistic under this null hypothesis by $\widehat{T}_{train,null}(t), \widehat{T}_{test,null}(t)$, and $\widehat{T}_{pop,null}(t)$ depending on the evaluation set, and when the evaluation set is understood from the context, we use $\widehat{T}_{null}(t)$.*

In this definition, note f^* is not supported on just the first k eigenfunctions, but rather only the projection via the zero-time kernel $\Pi_{K_0}(f^*)$ is supported on the first k eigenfunctions. This means that f^* may have a nonzero component that is orthogonal to $\Pi_{K_0}(f^*)$. Note that we have three two-sample test situations since the two-sample test depends on which dataset it is evaluated on. In particular, we will combine the results for $\widehat{T}_{pop}, \widehat{T}_{test}$, and $\widehat{T}_{train}$ into the following corollary since the only difference is given by a difference in constants.

Corollary 5.33. *Let* $\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}/2 > \varepsilon > 0$ *be a detection level. Now let*

$$C^+ = \begin{cases} \sqrt{2}L_1^2 4 & \widehat{T}_{pop}(t) \text{ evaluation} \\ \sqrt{2}L_1^2\left(4+\sqrt{A\frac{\log(n_p)}{n_p}}+\sqrt{A\frac{\log(n_q)}{n_q}}\right) & \widehat{T}_{train}(t) \text{ evaluation} \\ \sqrt{2}L_1^2\left(4+\sqrt{A\frac{\log(m_p)}{m_p}}+\sqrt{A\frac{\log(m_q)}{m_q}}\right) & \widehat{T}_{test}(t) \text{ evaluation} \end{cases},$$

coming from Proposition 5.26 and Proposition 5.25 and consider a time separation level $\varepsilon/C^+ \geq \gamma > 0$*. Let* $t^-(\varepsilon)$ *be such that for* $t \geq t^-(\varepsilon)$*, we have* $\widehat{T}_k(t) \geq \varepsilon$ *(for our different evaluation settings). Similarly, let* $t^+(\varepsilon)$ *be such that for* $t \geq t^+(\varepsilon)$*, we have* $\widehat{T}_{null}(t) \geq \varepsilon$*. If we assume*

$$\|f_k^*\|^2_{L^2(p+q)} > \max\left\{\frac{2\varepsilon\exp\left((\varepsilon/C^+-\gamma)/\lambda_k\right)}{\exp\left((\varepsilon/C^+-\gamma)/\lambda_k\right)-1}, \max_{a\in\{1,5/2\}}\frac{2\varepsilon\exp\left((\varepsilon/C^-)^{1/a}/\lambda_k\right)}{\exp\left((\varepsilon/C^-)^{1/a}/\lambda_k\right)-1}\right\},$$

where $C^- = C^+ + C_2 + C_3 + C_4$ *and the constants* C_2, C_3, C_4 *coming from Proposition 5.24, then*

$$t^+(\varepsilon) - t^-(\varepsilon) \geq \gamma > 0$$

with probability

$$\geq \begin{cases} 1-(n_p^{-A}+n_q^{-A}) & \widehat{T}_{pop}(t) \\ 1-(n_p^{-A}+n_q^{-A}+m_p^{-A}+m_q^{-A}) & \widehat{T}_{test}(t) \\ 1-2(n_p^{-A}+n_q^{-A}) & \widehat{T}_{train}(t) \end{cases}$$

The proof of Corollary 5.33 is in Section 5.11.

Remark 5.34. *From the proof of Corollary 5.33, we can see that the maximum time separation level is governed by*

$$\frac{\varepsilon}{C^+} - \lambda_k \log\left(\frac{\|f_k^*\|^2_{L^2(p+q)}}{\|f_k^*\|^2_{L^2(p+q)} - 2\varepsilon}\right).$$

Notice that if $\varepsilon = \|f_k^*\|^2_{L^2(p+q)} x$ *for some fraction* $0 < x < 1$*, then we can simplify this expression. In particular, we see that our expression changes to*

$$\begin{aligned}\gamma(x) &= \frac{\|f_k^*\|^2_{L^2(p+q)} x}{C^+} - \lambda_k \log\left(\frac{1}{1-2x}\right) \\ &= \frac{\|f_k^*\|^2_{L^2(p+q)} x}{C^+} - \lambda_k \log\left(\frac{1/2}{1/2-x}\right).\end{aligned}$$

From this, we can see that it is necessary that $0 < x < \frac{1}{2}$*. Note that as* $x \to 0$*, we get* $\gamma(x) \to 0$*; but as* $x \to \frac{1}{2}$*, we get* $\gamma(x) \to -\infty$*. Since* $\gamma(x)$ *is not decreasing, we can find a maximum for the time separation* $\gamma(x)$*. In particular, we see that*

$$\gamma'(x) = \frac{\|f_k^*\|^2_{L^2(p+q)}}{C^+} - \lambda_k \frac{1/2-x}{1/2} \frac{1/2}{(1/2-x)^2} = \frac{\|f_k^*\|^2_{L^2(p+q)}}{C^+} - \lambda_k \frac{1}{1/2-x}.$$

Setting this equal to 0, we see that the extrema is given by

$$x = \frac{1}{2} - \frac{\lambda_k C^+}{\|f_k^*\|^2_{L^2(p+q)}}.$$

Moreover, we note that this is a maximum since

$$\gamma''(x) = -\lambda_k \frac{1}{(1/2-x)^2} < 0.$$

Obviously, this only makes sense if

$$\frac{1}{2} > \frac{\lambda_k C^+}{\|f_k^*\|^2_{L^2(p+q)}} \iff \|f_k^*\|^2_{L^2(p+q)} > 2\lambda_k C^+.$$

This means that as long as f_k^* has large enough norm, our neural network two-sample test should be most trustworthy when we observe deviation $\varepsilon = \frac{\|f_k^*\|^2_{L^2(p+q)}}{2} - \lambda_k C^+$ since that is the deviation level with the maximum time separation between the assumption $\Pi_{K_0}(f^*) = f_k^*$ and the null hypothesis $p = q$.

Remark 5.35. *It is instructive to note what are fixed parameters versus parameters to be chosen in Corollary 5.33. First, notice that the complexity of our neural network determines not only the constants C^+, C^-, and λ_k but also whether or not the assumption $\Pi_{K_0}(f^*) = f_k^*$ holds. Although $f^* = \frac{p-q}{p+q}$ is fixed inherently from the two-sample test problem, we assume that the complexity of the neural network is fixed at initialization, which fixes these constants, the hypothesis, and how large $\|f_k^*\|^2_{L^2(p+q)}$ is. This means that the only choosable parameters are ε and γ (which is upper bounded by ε). Moreover, note that the upper bound $\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}/2 \geq \varepsilon > 0$ is an artifact of side-stepping the time-approximation error from Proposition 5.26. In particular, playing around with the proof of Corollary 5.33, it is possible to get a different bound for $\|f_k^*\|^2_{L^2(p+q)}$ albeit with the deviation level given by $\varepsilon - \delta(t)$ (depending on the evaluation set).*

5.8 Experiments

We run our neural network two-sample test on two different data-generating processes. One of the data-generating processes is a characteristically hard two-sample test problem where the datasets $\widehat{P}$ and $\widehat{Q}$ come from a Gaussian mixture model. The second data-generating process only aims to differentiate two multivariate Gaussians from each other. We scale the neural network complexity in terms of a ratio with respect to the number of samples in the training set. Additionally, we run permutation tests to find the threshold τ at the 95-percentile. We run around 500 different tests and check whether the test statistic is larger than the threshold found from the 95th percentile. We now showcase specifics of the data generating process and how the neural network is constructed.

5.8.1 Data Generating Process

Our hard two-sample testing problem is given by setting P and Q both to be Gaussian mixture models given by

$$P = \sum_{i=1}^{2} \frac{1}{2} \mathcal{N}(\mu_i^h, I_d)$$

$$Q = \sum_{i=1}^{2} \frac{1}{2} \mathcal{N}\left(\mu_i^h, \begin{bmatrix} 1 & \Delta_i^h & 0_{d-2} \\ \Delta_i^h & 1 & 0_{d-2} \\ 0_{d-2}^\top & 0_{d-2}^\top & I_{d-2} \end{bmatrix}\right),$$

where $\mu_1^h = 0_d$, $\mu_2^h = 0.5 * \mathbf{1}_d$, $\Delta_1^h = 0.5$, and $\Delta_2^h = -0.5$. For the purposes of testing, we assume that we have balanced sampling of N from each distribution P and Q so that the total number of samples is $2N$. The number of test samples is typically set to be $M < N$.

5.8.2 Neural Network Architecture

We use a neural network architecture of L layers and a layer-width size of k. We choose the number of parameters in the neural network kL as different ratios of the number of training samples. For example, if we consider the ratio 0.01 and $N = 1000$, then $kL = 10$. To adhere closely to the setup of the theory, we initialize our neural network symmetrically to make sure at time 0 our neural network returns 0. We make sure to initialize the neural network weights with the He initialization introduced in [51] so that the weights are initialized from a random normal distribution with variance $\frac{2}{k}$. This initialization ensures that our neural network training doesn't result in any exploding or vanishing gradients. We train the neural network with a learning rate of 0.1.

5.8.3 Test Results

We have attached below a heatmap of the statistical power as a function of the number of epochs as well as the ratio of parameters to samples.[1] Additionally we attach the evolution of the neural network two-sample test for a particular setting for reference. The hyperparameters for these tests essentially used a learning rate $\eta = 0.1$, 100 permutation tests, dimensionality of $d = 20$, training sample size of $N = 6000$ from each of P and Q, a testing sample size of $M = 1000$ from each of P and Q, and $L = 2$ layers. We train for a maximum of 15 epochs and use a batch size of 50. Moreover, our significance level α is the 95th percentile. We calculate the power of our neural network two-sample test by checking which of our 1000 tests lie past the 95th percentile of their respective permutation test and calculate the power by the ratio of all tests that lie past the 95th percentile divided by the total number of tests 1000. We try this experiment with ratios of parameters to number of training samples to see any double descent type of behavior in how well the statistical power performs.

Observing Figure 5.2, we notice that as the number of epochs increases the statistical

[1] All the code for producing these plots is on Github at this repository.

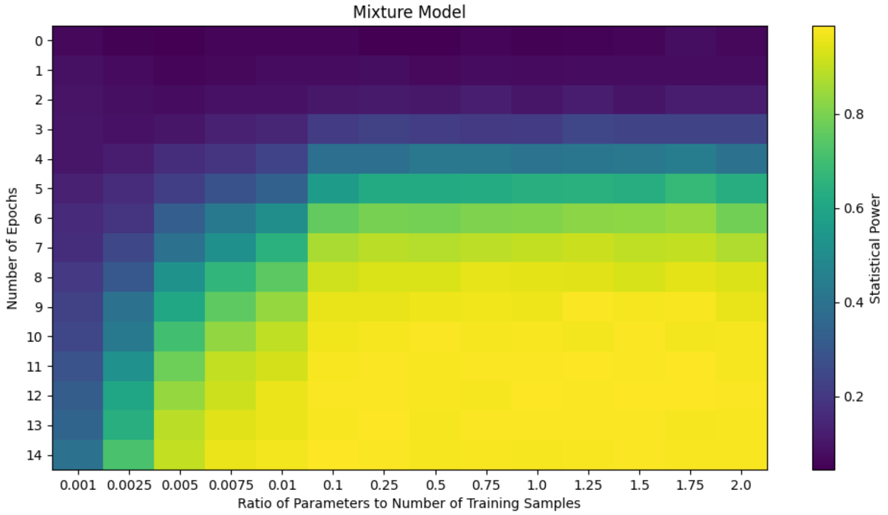


Figure 5.2: Plots statistical power for each epoch and ratio of parameters-to-samples.

power increases as well. On the ratio of parameters to training samples axis, however, we note that the smaller sized neural networks still produce fairly good statistical power with enough neural network training.

5.9 Proofs for Section 5.5.3

Proof of Proposition 5.4. We simply just need to take the derivative of the solution and show that it is exactly the differential equation, then the uniqueness of solutions of differential equations implies that our ansatz is indeed the solution. To see this, let us call the solution

$$e^*(\cdot,t) = \sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} u_\ell.$$

Now notice that

$$\partial_t e^*(\cdot,t) = -\sum_{\ell=1}^{M} \lambda_\ell e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} u_\ell.$$

On the other hand, let us plug in the ansatz e^* into the differential equation and see what we get. In particular,

$$\begin{aligned}
-\sum_{j=1}^{M} \lambda_j \langle u_j, e^*(\cdot,t)\rangle_{L^2(p+q)} u_j &= -\sum_{j=1}^{M} \lambda_j \langle u_j, \sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} u_\ell \rangle_{L^2(p+q)} u_j \\
&= -\sum_{j,\ell=1}^{M} \lambda_j e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} \underbrace{\langle u_j, u_\ell\rangle_{L^2(p+q)}}_{\delta_{j\ell}} u_j \\
&= -\sum_{\ell=1}^{M} \lambda_\ell e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} u_\ell \\
&= \partial_t e^*(\cdot,t).
\end{aligned}$$

This shows the result. □

Proof of Lemma 5.5. Now our two-sample test statistic becomes

$$\begin{aligned}
\overline{T}(t) &= \overline{\mu_P}(t) - \overline{\mu_Q}(t) \\
&= \int_{\mathbb{R}^d} \Pi_{K_0}(f^*)(x)(p-q)(x)dx + \int_{\mathbb{R}^d} \bar{e}(x,t)d(p-q)(x) \\
&= \int_{\mathbb{R}^d} \Pi_{K_0}(f^*)(x)\frac{p-q}{p+q}(x)d(p+q)(x) + \int_{\mathbb{R}^d} \bar{e}(x,t)d(p-q)(x) \\
&= \langle \Pi_{K_0}(f^*), f^*\rangle_{L^2(p+q)} + \int_{\mathbb{R}^d} \bar{e}(x,t)d(p-q)(x).
\end{aligned}$$

Extending the eigenfunctions $\{u_\ell\}_{\ell=1}^{M}$ to a full basis for $L^2(p+q)$ given by $\{u_\ell\}_{\ell=1}^{\infty}$, we can see that the term with f^* reduces to

$$\begin{aligned}
\langle \Pi_{K_0}(f^*), f^* \rangle_{L^2(p+q)} &= \left\langle \sum_{\ell=1}^{\infty} \langle u_\ell, \Pi_{K_0}(f^*) \rangle_{L^2(p+q)} u_\ell, \sum_{\ell'=1}^{\infty} \langle u'_\ell, f^* \rangle_{L^2(p+q)} u'_\ell \right\rangle_{L^2(p+q)} \\
&= \sum_{\ell=1}^{\infty} \langle u_\ell, \Pi_{K_0}(f^*) \rangle_{L^2(p+q)} \langle u_\ell, f^* \rangle_{L^2(p+q)} \\
&= \sum_{\ell=1}^{\infty} \underbrace{\langle \Pi_{K_0}(u_\ell), f^* \rangle_{L^2(p+q)}}_{\ell > M \implies 0} \langle u_\ell, f^* \rangle_{L^2(p+q)} \\
&= \sum_{\ell=1}^{M} \langle \Pi_{K_0}(u_\ell), f^* \rangle_{L^2(p+q)} \langle u_\ell, f^* \rangle_{L^2(p+q)} \\
&= \sum_{\ell=1}^{M} \langle u_\ell, \Pi_{K_0}(f^*) \rangle_{L^2(p+q)} \langle u_\ell, f^* \rangle_{L^2(p+q)}.
\end{aligned}$$

Now since $\langle u_\ell, \Pi_{K_0}(f^*) \rangle_{L^2(p+q)} = \langle u_\ell, f^* \rangle_{L^2(p+q)}$ for $\ell \leq M$, we see that

$$\begin{aligned}
\langle \Pi_{K_0}(f^*), f^* \rangle_{L^2(p+q)} &= \sum_{\ell=1}^{M} \langle u_\ell, \Pi_{K_0}(f^*) \rangle_{L^2(p+q)}^2 \\
&= \langle \Pi_{K_0}(f^*), \Pi_{K_0}(f^*) \rangle_{L^2(p+q)} = \|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2.
\end{aligned}$$

At this point, we plug in our ansatz and get

$$\begin{aligned}
\overline{T}(t) &= \|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 + \int_{\mathbb{R}^d} \sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot, 0) \rangle_{L^2(p+q)} u_\ell(x) d(p-q)(x) \\
&= \|\Pi_{K_0}(f^*)\|_{L^2(p+q)}^2 + \sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, \bar{e}(\cdot, 0) \rangle_{L^2(p+q)} \int_{\mathbb{R}^d} u_\ell(x) d(p-q)(x).
\end{aligned}$$

At this point, recall that we used the initialization θ_0 such that $\bar{e}(\cdot, 0) = \bar{u}(x, 0) - \Pi_{K_0}(f^*) = f(x, \theta_0) - \Pi_{K_0}(f^*) = -\Pi_{K_0}(f^*) = \Pi_{K_0}\left(\frac{q-p}{p+q}\right)(x)$. Along with the fact that $\langle u_\ell, \Pi_{K_0}(f^*) \rangle_{L^2(p+q)} =$

$\langle u_\ell, f^* \rangle_{L^2(p+q)}$ for $\ell \leq M$, we use this fact to see that

$$\begin{aligned}
\langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} &= \langle u_\ell, -\Pi_{K_0}(f^*)\rangle_{L^2(p+q)} \\
&= \int_{\mathbb{R}^d} u_\ell(x)\Pi_{K_0}\Big(\frac{q-p}{p+q}\Big)(x)(p(x)+q(x))dx \\
&= \int_{\mathbb{R}^d} u_\ell(x)\frac{q-p}{p+q}(x)(p(x)+q(x))dx \\
&= \int_{\mathbb{R}^d} u_\ell(x)d(q-p)(x) \\
\implies \langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} &= -\int_{\mathbb{R}^d} u_\ell(x)d(p-q)(x).
\end{aligned}$$

This means that

$$\begin{aligned}
\overline{T}(t) &= \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} + \sum_{\ell=1}^{M} e^{-t\lambda_\ell}\langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)} \underbrace{\int_{\mathbb{R}^d} u_\ell(x)d(p-q)(x)}_{-\langle u_\ell, \bar{e}(\cdot,0)\rangle_{L^2(p+q)}} \\
&= \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \sum_{\ell\geq 1} e^{-t\lambda_\ell}\langle u_\ell, \bar{e}(\cdot,0)\rangle^2_{L^2(p+q)}.
\end{aligned}$$

We get the result by seeing that $\bar{e}(\cdot,0) = -\Pi_{K_0}(f^*)$ and that applying the square gets rid of the negative sign. So we're done. □

Proof of Theorem 5.7. We want to find the smallest time t so that

$$\begin{aligned}
\overline{T}(t) &\geq \varepsilon \\
\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \sum_{\ell=1}^{M} e^{-t\lambda_\ell}\langle u_\ell, \Pi_{K_0}(f^*)\rangle^2_{L^2(p+q)} &\geq \varepsilon \\
\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \varepsilon &\geq \sum_{\ell=1}^{M} e^{-t\lambda_\ell}\langle u_\ell, \Pi_{K_0}(f^*)\rangle^2_{L^2(p+q)}.
\end{aligned}$$

Now using our specific subset $S \subseteq \{1,\ldots,M\}$, so that

$$\langle u_\ell, f^* \rangle_{L^2(p+q)} = \langle u_\ell, \Pi_{K_0}(f^*)\rangle_{L^2(p+q)},$$

allows us to consider the following analysis.

$$\begin{aligned}\sum_{\ell=1}^{M} e^{-t\lambda_\ell}\langle u_\ell, f^*\rangle^2_{L^2(p+q)} &= \sum_{\ell\in S} \underbrace{e^{-t\lambda_\ell}}_{\leq e^{-t\lambda_{\min}(S)}} \langle u_\ell, f^*\rangle^2_{L^2(p+q)} + \sum_{\ell\notin S} \underbrace{e^{-t\lambda_\ell}}_{\leq 1}\langle u_\ell, f^*\rangle^2_{L^2(p+q)} \\ &\leq e^{-t\lambda_{\min}(S)} \underbrace{\sum_{\ell\in S}\langle u_\ell, f^*\rangle_{L^2(p+q)}}_{\|f^*\|_S^2} + \underbrace{\sum_{\ell\notin S}\langle u_\ell, f^*\rangle^2_{L^2(p+q)}}_{\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \|f^*\|_S^2} \\ &= e^{-t\lambda_{\min}(S)}\|f^*\|_S^2 + \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \|f^*\|_S^2.\end{aligned}$$

We want this quantity to still be less than $\|f^*\|^2_{L^2(p+q)} - \varepsilon$ and to ensure this, we get

$$\begin{aligned}e^{-t\lambda_{\min}(S)}\|f^*\|_S^2 + \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \|f^*\|_S^2 &\leq \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \varepsilon \\ e^{-t\lambda_{\min}(S)}\|f^*\|_S^2 - \|f^*\|_S^2 &\leq -\varepsilon \\ e^{-t\lambda_{\min}(S)} - 1 &\leq -\frac{\varepsilon}{\|f^*\|_S^2} \\ e^{-t\lambda_{\min}(S)} &\leq 1 - \frac{\varepsilon}{\|f^*\|_S^2} \\ -t\lambda_{\min}(S) &\leq \log\left(1 - \frac{\varepsilon}{\|f^*\|_S^2}\right) \\ t &\geq \log\left(\left(1 - \frac{\varepsilon}{\|f^*\|_S^2}\right)^{-\lambda_{\min}(S)}\right).\end{aligned}$$

Rearranging the right-hand side and noticing that on S we have $\|\Pi_{K_0}(f^*)\|_S^2 = \|f^*\|_S^2$, we get the result. □

Proof of Theorem 5.9. We want to find the largest time t so that

$$\begin{aligned}
\overline{T}(t) &\leq \varepsilon \\
\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, \Pi_{K_0}(f^*)\rangle^2_{L^2(p+q)} &\leq \varepsilon \\
\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \varepsilon &\leq \sum_{\ell \geq 1} e^{-t\lambda_\ell} \langle u_\ell, \Pi_{K_0}(f^*)\rangle^2_{L^2(p+q)}.
\end{aligned}$$

Now using our specific subset S and that $\langle u_\ell, \Pi_{K_0}(f^*)\rangle_{L^2(p+q)} = \langle u_\ell, f^*\rangle_{L^2(p+q)}$ for $\ell \leq M$ allows us to consider the following analysis.

$$\begin{aligned}
\sum_{\ell=1}^{M} e^{-t\lambda_\ell} \langle u_\ell, f^*\rangle^2_{L^2(p+q)} &= \sum_{\ell \in S} \underbrace{e^{-t\lambda_\ell}}_{\geq e^{-t\lambda_{\max}(S)}} \langle u_\ell, f^*\rangle^2_{L^2(p+q)} + \sum_{\ell \notin S} \underbrace{e^{-t\lambda_\ell}}_{\geq 0} \langle u_\ell, f^*\rangle^2_{L^2(p+q)} \\
&\geq e^{-t\lambda_{\max}(S)} \underbrace{\sum_{\ell \in S} \langle u_\ell, f^*\rangle_{L^2(p+q)}}_{\|\Pi_{K_0}(f^*)\|^2_S} \\
&= e^{-t\lambda_{\max}(S)} \|\Pi_{K_0}(f^*)\|^2_S.
\end{aligned}$$

We want our lower bound found above to still be greater than $\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \varepsilon$ and to ensure this, we get

$$\begin{aligned}
e^{-t\lambda_{\max}(S)} \|\Pi_{K_0}(f^*)\|^2_S &\geq \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)} - \varepsilon \\
e^{-t\lambda_{\max}(S)} &\geq \frac{\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}}{\|\Pi_{K_0}(f^*)\|^2_S} - \frac{\varepsilon}{\|\Pi_{K_0}(f^*)\|^2_S} \\
-t\lambda_{\max}(S) &\geq \log\left(\frac{\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}}{\|\Pi_{K_0}(f^*)\|^2_S} - \frac{\varepsilon}{\|\Pi_{K_0}(f^*)\|^2_S}\right) \\
t &\leq \log\left\{\left(\frac{\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}}{\|\Pi_{K_0}(f^*)\|^2_S} - \frac{\varepsilon}{\|\Pi_{K_0}(f^*)\|^2_S}\right)^{-\lambda_{\max}(S)}\right\}.
\end{aligned}$$

Rearranging the right-hand side, we get the result. □

5.10 Proofs for Section 5.6

Proof of Proposition 5.14. Recall that the dynamics of θ can be written as

$$\dot{\theta}(t) = -\nabla_\theta \mathcal{L}(\theta(t)) = -\mathbb{E}_{x\sim p+q}\big[\nabla_\theta f(x,\theta(t))e(x,t)\big].$$

Moreover, note that we can write

$$\begin{aligned}\|\theta(t)-\theta(0)\| &\leq \int_0^t \|\dot{\theta}(s)\|ds \leq \int_0^t \|\nabla_\theta \mathcal{L}(\theta(s))\|ds \\ &\leq \sqrt{t}\bigg(\int_0^t \|\nabla_\theta \mathcal{L}(\theta(s))\|^2 ds\bigg)^{1/2},\end{aligned}$$

where the last inequality comes from the basic L_p-L_q inclusion inequality. Additionally, we know that

$$\frac{d}{dt}\mathcal{L}(\theta(t)) = \langle \nabla_\theta \mathcal{L}(\theta(t)), \dot{\theta}(t)\rangle = -\|\nabla_\theta \mathcal{L}(\theta(t))\|^2 \leq 0.$$

This not only implies that $\mathcal{L}(\theta(t))$ is decreasing but also allows us to write

$$\begin{aligned}\|\theta(t)-\theta(0)\| &\leq \sqrt{t}\bigg(\int_0^t \|\nabla_\theta \mathcal{L}(\theta(s))\|^2 ds\bigg)^{1/2} \\ &= \sqrt{t}\Big(\mathcal{L}(\theta(0)) - \mathcal{L}(\theta(t))\Big)^{1/2} \\ &\leq \sqrt{t}\sqrt{\mathcal{L}(\theta(0))}.\end{aligned}$$

At this point, we can notice that $\mathcal{L}(\theta(0)) = \|u(x,0)-f^*\|^2_{L^2(p+q)} = \|f^*\|^2_{L^2(p+q)}$. This finally gives us the result

$$\|\theta(t)-\theta(0)\| \leq \sqrt{t}\|f^*\|_{L^2(p+q)}.$$

We need $\theta(t) \in B_R$ and one way to ensure this is

$$\begin{aligned} \|\theta(t) - \theta(0)\| \leq \sqrt{t}\|f^*\|_{L^2(p+q)} \leq R \\ \implies t \leq \left(\frac{R}{\|f^*\|_{L^2(p+q)}}\right)^2 \end{aligned}$$

□

Proof of Proposition 5.15. Let $w \in L^2(p+q)$, then note that for our kernel integral operator K_t, we have $\langle w, \mathbb{E}_{x' \sim p+q} K_t(\cdot, x')w(x')\rangle_{L^2(p+q)}$ equals

$$\begin{aligned} &\mathbb{E}_{x \sim p+q}\mathbb{E}_{x' \sim p+q} w(x)\langle \nabla_\theta f(x, \theta(t)), \nabla_\theta f(x', \theta(t))\rangle_\Theta w(x') \\ &= \langle \mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(t))w(x), \mathbb{E}_{x' \sim p+q} \nabla_\theta f(x', \theta(t))w(x')\rangle_\Theta \\ &= \|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(t))w(x)\|_\Theta^2. \end{aligned}$$

This means that $\langle w, \mathbb{E}_{x' \sim p+q}(K_t - K_0)(\cdot, x')w(x')\rangle_{L^2(p+q)}$ equals

$$\begin{aligned} &\|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(t))w(x)\|_\Theta^2 - \|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(0))w(x)\|_\Theta^2 \\ &= (\|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(t))w(x)\|_\Theta + \|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(0))w(x)\|_\Theta) \\ &\cdot (\|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(t))w(x)\|_\Theta - \|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(0))w(x)\|_\Theta). \end{aligned}$$

Now using Minkowski's integral inequality and Assumption 5.13(1), we get

$$\begin{aligned} \|\mathbb{E}_{x \sim p+q} \nabla_\theta f(x, \theta(t))w(x)\|_\Theta &\leq \mathbb{E}_{x \sim p+q}\|\nabla_\theta f(x, \theta(t))w(x)\|_\Theta \\ &\leq \mathbb{E}_{x \sim p+q}\|\nabla_\theta f(x, \theta(t))\|_\Theta \|w(x)\| \\ &\leq L_1 \mathbb{E}_{x \sim p+q}\|w(x)\| \\ &\leq L_1 \|w\|_{L^2(p+q)}. \end{aligned}$$

Notice that this implies

$$\|\mathbb{E}_{x\sim p+q}\nabla_\theta f(x,\theta(t))w(x)\|_\Theta + \|\mathbb{E}_{x\sim p+q}\nabla_\theta f(x,\theta(0))w(x)\|_\Theta \le 2L_1\|w\|_{L^2(p+q)}.$$

Now using Assumption 5.13(2), we have

$$\begin{aligned}
\|\mathbb{E}_{x\sim p+q}\nabla_\theta f(x,\theta(t))w(x)\|_\Theta - \|\mathbb{E}_{x\sim p+q}\nabla_\theta f(x,\theta(0))w(x)\|_\Theta&\\
\le \|\mathbb{E}_{x\sim p+q}(\nabla_\theta f(x,\theta(t)) - \nabla_\theta f(x,\theta(0)))w(x)\|_\Theta&\\
\le \mathbb{E}_{x\sim p+q}\|\nabla_\theta f(x,\theta(t)) - \nabla_\theta f(x,\theta(0))\|_\Theta\|w(x)\|&\\
\le L_2\|\theta(t)-\theta(0)\|_\Theta\|w\|_{L^2(p+q)}&\\
\le L_2\sqrt{t}\|f^*\|_{L^2(p+q)}\|w\|_{L^2(p+q)}.&
\end{aligned}$$

This means that

$$\langle w, \mathbb{E}_{x'\sim p+q}K_t(\cdot,x')w(x')\rangle_{L^2(p+q)} \le 2L_1L_2\sqrt{t}\|f^*\|_{L^2(p+q)}\|w\|^2_{L^2(p+q)}.$$

Finally this proves that $\|K_t - K_0\|_{L^2(p+q)} \le 2L_1L_2\sqrt{t}\|f^*\|_{L^2(p+q)}$. □

Proof of Proposition 5.16. Note that

$$\begin{aligned}
\partial_t(u-\bar{u})(x,t) = \partial_t(e-\bar{e})(x,t) &= \mathbb{E}_{x'\sim p+q}\big[K_0(x,x')\bar{e}(x',t) - K_t(x,x')e(x',t)\big]\\
&= -\mathbb{E}_{x'\sim p+q}\big[(K_t(x,x')-K_0(x,x'))\bar{e}(x',t) + K_t(x,x')(e-\bar{e})(x',t)\big].
\end{aligned}$$

Notice here that because $K_t(x,x') = \langle\nabla_\theta f(x,\theta(t)),\nabla_\theta f(x',\theta(t))\rangle$, we have that K_t is a positive semi-definite operator (we will use this later). Now if we take an inner product with $e-\bar{e}$ on both sides of the equation, we get

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\|(e-\bar{e})(\cdot,t)\|^2_{L^2(p+q)} &= \langle (e-\bar{e})(\cdot,t), \partial_t(e-\bar{e})(\cdot,t)\rangle_{L^2(p+q)}\\ &= \langle (e-\bar{e})(\cdot,t), -\mathbb{E}_{x'\sim p+q}\big[(K_t(x,x')-K_0(x,x'))\bar{e}(x',t)\\ &+K_t(x,x')(e-\bar{e})(x',t)\big]\rangle_{L^2(p+q)}\\ &\leq |\langle (e-\bar{e})(\cdot,t), -\mathbb{E}_{x'\sim p+q}\big[(K_t(x,x')-K_0(x,x'))\bar{e}(x',t)\big]\rangle_{L^2(p+q)}|\\ &\leq \|(e-\bar{e})(\cdot,t)\|_{L^2(p+q)}\|K_t-K_0\|_{L^2(p+q)}\|\bar{e}(\cdot,t)\|_{L^2(p+q)}\\ &\leq \|(e-\bar{e})(\cdot,t)\|_{L^2(p+q)}\|K_t-K_0\|_{L^2(p+q)}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)},\end{aligned}$$

where the first inequality comes from the fact that K_t is a positive semi-definite operator as well as using absolute values whilst the second inequality comes from using the Cauchy-Schwartz-Bunyakovsky inequality along with the kernel integral operator norm bound of $\|K_t-K_0\|_{L^2(p+q)}$. Now recalling that $\bar{e}(\cdot,0)=\Pi_{K_0}(f^*)$ and using Parseval's identity, the last inequality comes from the fact that

$$\|\bar{e}(\cdot,t)\|^2_{L^2(p+q)} = \sum_{\ell=1}^{M}\underbrace{e^{-2t\lambda_\ell}}_{\leq 1}|\langle u_\ell,\bar{e}(\cdot,0)\rangle|^2 \leq \sum_{\ell=1}^{M}|\langle u_\ell,\Pi_{K_0}(f^*)\rangle|^2 = \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}.$$

From Proposition 5.15, we get that

$$\frac{d}{dt}\frac{1}{2}\|(e-\bar{e})(\cdot,t)\|^2_{L^2(p+q)} \leq 2L_1L_2\sqrt{t}\|f^*\|_{L^2(p+q)}\cdot\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\cdot\|(e-\bar{e})(\cdot,t)\|_{L^2(p+q)}.$$

Now, finally notice that

$$\begin{aligned}\frac{1}{2}\|(e-\bar{e})(\cdot,t)\|^2_{L^2(p+q)} &= \int_0^t\Big(\frac{d}{dt}\frac{1}{2}\|(e-\bar{e})(\cdot,s)\|^2_{L^2(p+q)}\Big)ds\\ &\leq 2L_1L_2\|f^*\|_{L^2(p+q)}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\int_0^t\sqrt{s}\|(e-\bar{e})(\cdot,s)\|_{L^2(p+q)}ds.\end{aligned}$$

Let $t^* \leq t$ be the time such that

$$\sup_{s \in [0,t]} \|(e-\bar{e})(\cdot,s)\|_{L^2(p+q)} = \|(e-\bar{e})(\cdot,t^*)\|_{L^2(p+q)}.$$

Then, we find that

$$\frac{1}{2}\|(e-\bar{e})(\cdot,t^*)\|^2_{L^2(p+q)} \leq 2L_1L_2\|f^*\|_{L^2(p+q)}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\|(e-\bar{e})(\cdot,t^*)\|_{L^2(p+q)}\int_0^{t^*}\sqrt{s}ds,$$

but this implies

$$\begin{aligned}
\frac{1}{2}\|(e-\bar{e})(\cdot,t^*)\|_{L^2(p+q)} &\leq 2L_1L_2\|f^*\|_{L^2(p+q)}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\frac{2}{3}(t^*)^{3/2} \\
\implies \|(e-\bar{e})(\cdot,t)\|_{L^2(p+q)} &\leq 4L_1L_2\|f^*\|_{L^2(p+q)}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\frac{2}{3}(t^*)^{3/2} \\
&\leq (8/3)L_1L_2\|f^*\|_{L^2(p+q)}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}(t)^{3/2},
\end{aligned}$$

where we use the fact that $\|(e-\bar{e})(\cdot,t)\|_{L^2(p+q)} \leq \|(e-\bar{e})(\cdot,t^*)\|_{L^2(p+q)}$ and $t^* \leq t$. Finally using that $\|\Pi_{K_0}(f^*)\|_{L^2(p+q)} \leq \|f^*\|_{L^2(p+q)}$ gives the result, so we're done. □

Proof of Corollary 5.17. Notice that

$$\begin{aligned}
\left|T(t)-\overline{T}(t)\right| &= \left|\int_{\mathbb{R}^d}(u-\bar{u})(x,t)d(p-q)(x)\right| \\
&\leq \left|\int_{\mathbb{R}^d}(u-\bar{u})(x,t)d(p)(x)\right| + \left|\int_{\mathbb{R}^d}(u-\bar{u})(x,t)d(q)(x)\right| \\
&\leq \int_{\mathbb{R}^d}|(u-\bar{u})(x,t)|d(p)(x) + \int_{\mathbb{R}^d}|(u-\bar{u})(x,t)|d(q)(x) \\
&= \int_{\mathbb{R}^d}|(u-\bar{u})(x,t)|d(p+q)(x) \\
&\leq \sqrt{2}\|u-\bar{u}\|_{L^2(p+q)},
\end{aligned}$$

where the last inequality comes from using a basic L_1-L_2 inclusion inequality. Now using

Proposition 5.16, we get the result, and we're done. □

Proof of Theorem 5.18. Note that the loss $\mathcal{L}(\theta(t)) = \|f(\cdot,\theta(t)) - f^*\|^2_{L^2(p+q)}$ is monotonically decreasing since

$$\frac{d}{dt}\mathcal{L}(\theta(t)) = \langle \nabla_\theta \mathcal{L}(\theta(t)), \dot{\theta}(t)\rangle = -\|\nabla_\theta \mathcal{L}(\theta(t))\|^2 \leq 0.$$

So we have that $\mathcal{L}(\theta(s)) \geq \mathcal{L}(\theta(t))$ if $0 \leq s \leq t \leq \tau$. Writing out the loss as $\mathcal{L}(\theta(s)) = \|f(\cdot,\theta(s))\|^2_{L^2(p+q)} - 2\langle f(\cdot,\theta(s)), f^*\rangle_{L^2(p+q)} + \|f^*\|^2_{L^2(p+q)}$, we can see

$$\begin{aligned} \mathcal{L}(\theta(s)) &\geq \mathcal{L}(\theta(t)) \\ \|f(\cdot,\theta(s))\|^2_{L^2(p+q)} - 2\langle f(\cdot,\theta(s)), f^*\rangle_{L^2(p+q)} &\geq \|f(\cdot,\theta(t))\|^2_{L^2(p+q)} - 2\langle f(\cdot,\theta(t)), f^*\rangle_{L^2(p+q)} \\ \langle f(\cdot,\theta(t)) - f(\cdot,\theta(s)), f^*\rangle_{L^2(p+q)} &\geq |f(\cdot,\theta(t))\|^2_{L^2(p+q)} - \|f(\cdot,\theta(s))\|_{L^2(p+q)}. \end{aligned}$$

Notice that because $u(x,t) = f(x,\theta(t))$ and we assume that $\|u(x,t)\|^2_{L^2(p+q)}$ is increasing in time, we know that

$$\begin{aligned} \langle f(\cdot,\theta(t)) - f(\cdot,\theta(s)), f^*\rangle_{L^2(p+q)} &\geq \underbrace{|f(\cdot,\theta(t))\|^2_{L^2(p+q)} - \|f(\cdot,\theta(s))\|_{L^2(p+q)}}_{\geq 0} \\ \int_{\mathbb{R}^d} \big(u(x,t) - u(x,s)\big)\frac{p-q}{p+q}(x)d(p+q)(x) \geq 0 \\ \int_{\mathbb{R}^d} \big(u(x,t) - u(x,s)\big)d(p-q)(x) \geq 0 \\ \int_{\mathbb{R}^d} u(x,t)d(p-q)(x) &\geq \int_{\mathbb{R}^d} u(x,s)d(p-q)(x). \end{aligned}$$

So we see that on the interval $[0,T]$, the two-sample test statistic $T(t)$ is monotonically increasing.

□

Proof of Theorem 5.19. Using the assumption

$$\min\left(\tau, \Big(\frac{R}{\|f^*\|_{L^2(p+q)}}\Big)^2\right) \geq t \geq t_1^*(\varepsilon)$$

allows us to use Corollary 5.17, Corollary 5.10, and Theorem 5.18 simultaneously. Using monotonicity and the reverse triangle inequality shows that

$$\begin{aligned}|T(t)| \geq |T(t_1^*(\varepsilon))| &\geq \left|\left|\overline{T}(t_1^*(\varepsilon))\right| - \left|T(t_1^*(\varepsilon)) - \overline{T}(t_1^*(\varepsilon))\right|\right| \\ &\geq \varepsilon - \frac{8\sqrt{2}}{3}\|f^*\|^2_{L^2(p+q)} L_1 L_2 \big(t_1^*(\varepsilon)\big)^{3/2}\end{aligned}$$

where we can get rid of the absolute values by assumption. So we're done. □

Proof of Theorem 5.20. Because of the assumption on t, we can use both Corollary 5.10 as well as Corollary 5.17. Using the triangle inequality gives us

$$\begin{aligned}|T(t)| \leq \left|\overline{T}(t)\right| + \left|T(t) - \overline{T}(t)\right| &\leq \varepsilon + \frac{8\sqrt{2}}{3}\|f^*\|^2_{L^2(p+q)} L_1 L_2 t \\ &\leq \varepsilon + \frac{8\sqrt{2}}{3}\|f^*\|^2_{L^2(p+q)} L_1 L_2 \big(t_2^*(\varepsilon)\big)^{3/2}.\end{aligned}$$

So we're done. □

5.11 Proofs for Section 5.7

Proof of Lemma 5.21. Similar to the proof of Proposition 5.14, we recall that

$$\begin{aligned}\dot{\widehat{\theta}}(t) &= -\frac{1}{2}\left(\int_{\mathbb{R}^d} \nabla_\theta f(x,\widehat{\theta}(t))\Big(\big(f(x,\widehat{\theta}(t))-1\big)\widehat{p}(x) + \big(f(x,\widehat{\theta}(t))+1\big)\widehat{q}(x)\Big)dx\right) \\ &= -\nabla_\theta \widehat{L}(\widehat{\theta}(t)).\end{aligned}$$

Moreover, recall that

$$\frac{d}{dt}\widehat{L}(\widehat{\theta}(t)) = \langle \nabla_\theta \widehat{L}(\widehat{\theta}(t)), \dot{\widehat{\theta}}(t)\rangle_\Theta = -\|\nabla_\theta \widehat{L}(\widehat{\theta}(t))\|_\Theta \leq 0.$$

This already shows that $\widehat{L}(\widehat{\theta}(t)) \leq \widehat{L}(\theta(0))$. Now, notice that

$$\begin{aligned}\|\widehat{\theta}(t) - \theta(0)\|_\Theta &\leq \int_0^t \|\dot{\widehat{\theta}}(s)\|_\Theta ds \leq \sqrt{t}\left(\int_0^t \|\dot{\widehat{\theta}}(s)\|_\Theta^2 ds\right)^{1/2} \\ &\leq \sqrt{t}\left(\int_0^t \|\nabla_\theta \widehat{L}(\widehat{\theta}(s))\|_\Theta^2 ds\right)^{1/2} \\ &= \sqrt{t}\sqrt{\widehat{L}(\widehat{\theta}(0)) - \widehat{L}(\widehat{\theta}(t))} \leq \sqrt{t}\sqrt{\widehat{L}(\widehat{\theta}(0))}.\end{aligned}$$

Now because $f(\cdot,\theta(0)) = f(\cdot,\widehat{\theta}(0)) = 0$, we get that

$$\begin{aligned}\widehat{L}(\theta(0)) &= \frac{1}{2}\left(\int_{\mathbb{R}^d} \big(f(x,\theta(0)) - 1\big)^2 \widehat{p}(x)dx + \int_{\mathbb{R}^d} \big(f(x,\theta(0)) + 1\big)^2 \widehat{q}(x)dx\right) \\ &= \frac{1}{2}\left(\int_{\mathbb{R}^d} \widehat{p}(x)dx + \int_{\mathbb{R}^d} \widehat{q}(x)dx\right) = 1.\end{aligned}$$

So this implies that

$$\|\widehat{\theta}(t) - \theta(0)\|_\Theta \leq \sqrt{t}.$$

For the second statement, just notice that we want to ensure $\|\widehat{\theta}(t) - \theta(0)\|_\Theta \leq R$. With our bounds, this is ensured if

$$\sqrt{t} \leq R.$$

Readjusting this expression gives us the result. □

Proof of Lemma 5.23. Notice that $\mathbb{E}X_i = 0$ and

$$\|X_i\| \leq \|\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\| + \mathbb{E}_{x\sim p}\|\nabla_\theta f(x,\theta)\nabla_\theta f(x,\theta)^\top\| \leq 3L_1^2.$$

Moreover, notice that

$$\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_iX_i^\top\| \leq \frac{1}{n}\sum_{i=1}^n \|\underbrace{\mathbb{E}X_iX_i^\top}_{I}\|.$$

Simplifying I, we see

$$\begin{aligned} I &= \mathbb{E}_{x_i\sim p}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\|\nabla_\theta f(x_i,\theta)\|^2 \\ &\quad - 2(\mathbb{E}_{x_i\sim p}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top)^2 \\ &\quad + (\mathbb{E}_{x_i\sim p}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top)^2. \end{aligned}$$

This means that

$$I = \mathbb{E}_{x_i\sim p}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\|\nabla_\theta f(x_i,\theta)\|^2 - (\mathbb{E}_{x_i\sim p}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top)^2,$$

which implies

$$\begin{aligned} \|I\| &\leq \mathbb{E}_{x_i\sim p}\underbrace{\|\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\|\|\nabla_\theta f(x_i,\theta)\|^2}_{\leq L_1^4} \\ &\quad + \underbrace{\|\mathbb{E}_{x_i\sim p}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\|^2}_{L_1^4} \\ &\leq 2L_1^4. \end{aligned}$$

Since X_i is symmetric, we know the same bound holds for $X_i^\top X_i$ terms. This means that $\nu = 2L_1^4$.

Finally, using Theorem 5.40 and cleaning some terms, we get that

$$\Pr\Big[\|\frac{1}{n}\sum_{i=1}^{n} X_i\| \geq t\Big] \leq 2M_\Theta \exp\Big\{-\frac{nt^2}{2L_1^2(2L_1^2+t)}\Big\}.$$

Let us consider when

$$t = \sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n)+\log(2M_\Theta)}{n}}.$$

Moreover, we can choose n large enough such that

$$\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n)+\log(2M_\Theta)}{n}} < \frac{3}{2},$$

then we have that

$$\begin{aligned}
\frac{nt^2}{2L_1^2(2L_1^2+t)} &> \frac{nt^2}{2L_1^2(2L_1^2+(3/2))} \\
-\frac{nt^2}{2L_1^2(2L_1^2+t)} &< -\frac{nt^2}{2L_1^2(2L_1^2+(3/2))} \\
\exp\Big\{-\frac{nt^2}{2L_1^2(2L_1^2+t)}\Big\} &< \exp\Big\{-\frac{nt^2}{2L_1^2(2L_1^2+(3/2))}\Big\} \\
\implies 2M_\Theta \exp\Big\{-\frac{nt^2}{2L_1^2(2L_1^2+t)}\Big\} &\leq 2M_\Theta \exp\Big\{-\frac{nt^2}{2L_1^2(2L_1^2+(3/2))}\Big\}.
\end{aligned}$$

So with our choice of t, we actually get that

$$\begin{aligned}
&2M_\Theta \exp\Big\{-\frac{nt^2}{2L_1^2(2L_1^2+(3/2))}\Big\} \\
&= 2M_\Theta \exp\Big\{-\frac{(2L_1^2(2L_1^2+(3/2)))(A\log(n)+\log(2M_\Theta))}{2L_1^2(2L_1^2+(3/2))}\Big\} \\
&= 2M_\Theta \exp\Big\{-(A\log(n)+\log(2M_\Theta))\Big\} \\
&= n^{-A}.
\end{aligned}$$

Taking the compliment of this event, we get that with probability greater than $1-n^{-A}$

$$\|\frac{1}{n}\sum_{i=1}^{n} X_i\| \leq \sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n)+\log(2M_\Theta)}{n}}.$$

So we're done. □

Proposition 5.36. *Assume that $t \leq R^2$ (so that $\theta(t) \in B_R$) as well as Assumption 5.13 and Assumption 5.22, then with probability $\geq 1-n_p^{-A}-n_q^{-A}$, we have $\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}$ is less than or equal to*

$$\begin{aligned}
&4L_1^2t+4L_1L_2t^{3/2}\|f^*\|_{L^2(p+q)}\Big(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}\Big)^{1/2}\\
&\quad+4L_1L_2t^{3/2}\|f^*\|_{L^2(p+q)}\Big(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n_q)+\log(2M_\Theta)}{n_q}}\Big)^{1/2}\\
&\quad+t^2\cdot\sqrt{2}L_1^3\|f^*\|_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}\\
&\quad+t^2\cdot\sqrt{2}L_1^3\|f^*\|_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_q)+\log(2M_\Theta)}{n_p}}.
\end{aligned}$$

So that $\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}$ is $O(t^{5/2})$.

Proof of Proposition 5.36. Inspecting $\partial_t(\widehat{u}-\bar{u})$ more closely, we see that

$$\begin{aligned}
2\partial_t(\widehat{u}-\bar{u})(\cdot,t) &= -\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')\widehat{e}_p(x',t)-\mathbb{E}_{x'\sim\widehat{q}}\widehat{K}_t(\cdot,x')\widehat{e}_q(x',t)\\
&\quad+\mathbb{E}_{x'\sim p}K_0(\cdot,x')\bar{e}(x',t)+\mathbb{E}_{x'\sim q}K_0(\cdot,x')\bar{e}(x',t).
\end{aligned}$$

Notice that

$$\begin{aligned}
&-\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')\widehat{e}_p(x',t)+\mathbb{E}_{x'\sim p}K_0(\cdot,x')\bar{e}(x',t)\\
&=-\Big\{\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')(\widehat{e}_p(x',t)-\bar{e}(x',t))+\mathbb{E}_{x'\sim\widehat{p}}(\widehat{K}_t-K_0)(\cdot,x')\bar{e}(x',t)\\
&\qquad+\Big(\mathbb{E}_{x'\sim\widehat{p}}-\mathbb{E}_{x'\sim p}\Big)K_0(\cdot,x')\bar{e}(x',t)\Big\}.
\end{aligned}$$

For q, we get a similar form

$$\begin{aligned}
&-\mathbb{E}_{x'\sim\widehat{q}}\widehat{K}_t(\cdot,x')\widehat{e}_q(x',t)+\mathbb{E}_{x'\sim q}K_0(\cdot,x')\bar{e}(x',t)\\
&=-\Big\{\mathbb{E}_{x'\sim\widehat{q}}\widehat{K}_t(\cdot,x')(\widehat{e}_q(x',t)-\bar{e}(x',t))+\mathbb{E}_{x'\sim\widehat{q}}(\widehat{K}_t-K_0)(\cdot,x')\bar{e}(x',t)\\
&\qquad+\Big(\mathbb{E}_{x'\sim\widehat{q}}-\mathbb{E}_{x'\sim q}\Big)K_0(\cdot,x')\bar{e}(x',t)\Big\}.
\end{aligned}$$

Putting this together, we get

$$\begin{aligned}
2\partial_t(\widehat{u}-\bar{u})(\cdot,t)&=\underbrace{-\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')(\widehat{e}_p(x',t)-\bar{e}(x',t))}_{I_{1,p}}\\
&+\underbrace{\mathbb{E}_{x'\sim\widehat{p}}(K_0-\widehat{K}_t)(\cdot,x')\bar{e}(x',t)}_{I_{2,p}}+\underbrace{\Big(\mathbb{E}_{x'\sim p}-\mathbb{E}_{x'\sim\widehat{p}}\Big)K_0(\cdot,x')\bar{e}(x',t)}_{I_{3,p}}\\
&\underbrace{-\mathbb{E}_{x'\sim\widehat{q}}\widehat{K}_t(\cdot,x')(\widehat{e}_q(x',t)-\bar{e}(x',t))}_{I_{1,q}}+\underbrace{\mathbb{E}_{x'\sim\widehat{q}}(K_0-\widehat{K}_t)(\cdot,x')\bar{e}(x',t)}_{I_{2,q}}\\
&+\underbrace{\Big(\mathbb{E}_{x'\sim q}-\mathbb{E}_{x'\sim\widehat{q}}\Big)K_0(\cdot,x')\bar{e}(x',t)}_{I_{3,q}}.
\end{aligned}$$

Similar to the proof of Proposition 5.16, we will consider

$$\begin{aligned}\frac{d}{dt}\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}^2 &= \langle(\widehat{u}-\bar{u})(\cdot,t), 2\partial_t(\widehat{u}-\bar{u})(\cdot,t)\rangle_{L^2(p+q)} \\ &= \langle(\widehat{u}-\bar{u})(\cdot,t), I_{1,p}+I_{1,q}\rangle_{L^2(p+q)} + \langle(\widehat{u}-\bar{u})(\cdot,t), I_{2,p}+I_{2,q}\rangle_{L^2(p+q)} \\ &+ \langle(\widehat{u}-\bar{u})(\cdot,t), I_{3,p}+I_{3,q}\rangle_{L^2(p+q)} \\ &\leq \langle(\widehat{u}-\bar{u})(\cdot,t), I_{1,p}+I_{1,q}\rangle_{L^2(p+q)} \\ &+ \|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}\Big(\|I_{2,p}+I_{2,q}\|_{L^2(p+q)} + \|I_{3,p}+I_{3,q}\|_{L^2(p+q)}\Big).\end{aligned}$$

So we'll need to bound $I_{2,p}, I_{2,q}, I_{3,p}$, and $I_{3,q}$ and will deal with the $I_{1,p}$ and $I_{1,q}$ terms at the end.

Before starting, let A_P be the event that

$$\|(\mathbb{E}_{x'\sim p}-\mathbb{E}_{x'\sim\widehat{p}})\nabla_\theta u(x',\theta(0))\nabla_\theta u(x',\theta(0))^\top\| \leq \sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}$$

and let A_Q be the event that

$$\|(\mathbb{E}_{x'\sim q}-\mathbb{E}_{x'\sim\widehat{q}})\nabla_\theta u(x',\theta(0))\nabla_\theta u(x',\theta(0))^\top\| \leq \sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_q)+\log(2M_\Theta)}{n_q}}.$$

Note that from Lemma 5.23, we know that A_P occurs with probability $\geq 1-n_p^{-A}$ and A_Q occurs with probability $\geq 1-n_q^{-A}$. Since these events are disjoint, notice that

$$\Pr\big(A_P\cap A_Q\big) = 1-\Pr\big(A_P^c\cup A_Q^c\big) = 1-n_p^{-A}-n_q^{-A}$$

where A_P^c and A_Q^c are the complements of A_P and A_Q respectively. We work in the regime that both A_P and A_Q occur.

Bounding $I_{3,p}$ and $I_{3,q}$: We will first work with just $I_{3,p}$ and will notice that the method of bounding $I_{3,q}$ is the same. Then using the triangle inequality, we will get our bounds. Notice that

$$
\begin{aligned}
I_{3,p} &= \langle \nabla_\theta u(\cdot,\theta(0)), (\mathbb{E}_{x'\sim p} - \mathbb{E}_{x'\sim \widehat{p}})\nabla_\theta u(x',\theta(0))\bar{e}(x',t)\rangle_\Theta \\
\implies \|I_{3,p}\|_{L^2(p+q)} &\leq \Big\| \underbrace{\|\nabla_\theta u(\cdot,\theta(0))\|_\Theta}_{\leq L_1} \Big\|_{L^2(p+q)} \|(\mathbb{E}_{x'\sim p} - \mathbb{E}_{x'\sim \widehat{p}})\nabla_\theta u(x',\theta(0))\bar{e}(x',t)\|_\Theta \\
&\leq \sqrt{2}L_1 \| \underbrace{(\mathbb{E}_{x'\sim p} - \mathbb{E}_{x'\sim \widehat{p}})\nabla_\theta u(x',\theta(0))\bar{e}(x',t)}_{a_3} \|_\Theta.
\end{aligned}
$$

Now we can use the fact that

$$
\begin{aligned}
\bar{e}(x,t) &= -\int_0^t \mathbb{E}_{y\sim p+q} K_0(x,y)\bar{e}(y,s)ds \\
&= -\int_0^t \langle \nabla_\theta u(x,\theta(0)), \mathbb{E}_{y\sim p+q}\nabla_\theta u(y,\theta(0))\bar{e}(y,s)\rangle_\Theta ds.
\end{aligned}
$$

This means that we can rewrite a_3 as

$$
a_3 = -\int_0^t \left[(\mathbb{E}_{x'\sim p} - \mathbb{E}_{x'\sim \widehat{p}})\nabla_\theta u(x',\theta(0))\nabla_\theta u(x',\theta(0))^\top\right]\mathbb{E}_{y\sim p+q}\nabla_\theta u(y,\theta(0))\bar{e}(y,s)ds.
$$

This would mean that $\|a_3\|_\Theta$ is bounded by

$$
\begin{aligned}
&\int_0^t \|(\mathbb{E}_{x'\sim p} - \mathbb{E}_{x'\sim \widehat{p}})\nabla_\theta u(x',\theta(0))\nabla_\theta u(x',\theta(0))^\top\| \underbrace{\|\nabla_\theta u(y,\theta(0))\|_\Theta}_{\leq L_1} \mathbb{E}_{y\sim p+q}\|\bar{e}(y,s)\|ds \\
&\qquad \leq tL_1\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\|\|(\mathbb{E}_{x'\sim p} - \mathbb{E}_{x'\sim \widehat{p}})\nabla_\theta u(x',\theta(0))\nabla_\theta u(x',\theta(0))^\top\|.
\end{aligned}
$$

Now recalling that $\bar{e}(\cdot,0) = \Pi_{K_0}(f^*)$ and using Parseval's identity, the last inequality comes from the fact that

$$
\|\bar{e}(\cdot,t)\|^2_{L^2(p+q)} = \sum_{\ell=1}^M \underbrace{e^{-2t\lambda_\ell}}_{\leq 1}|\langle u_\ell, \bar{e}(\cdot,0)\rangle|^2 \leq \sum_{\ell=1}^M |\langle u_\ell, \Pi_{K_0}(f^*)\rangle|^2 = \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}.
$$

So we only need to bound the operator norm of

$$\|(\mathbb{E}_{x'\sim p}-\mathbb{E}_{x'\sim\widehat{p}})\nabla_\theta u(x',\theta(0))\nabla_\theta u(x',\theta(0))^\top\|.$$

To this end, since we assume that we are working under event $A_P \cap A_Q$, we can again use Lemma 5.23 and get that with probability greater than $1-{n_p}^{-A}-n_q^{-A}$,

$$\|(\mathbb{E}_{x'\sim p}-\mathbb{E}_{x'\sim\widehat{p}})\nabla_\theta u(x',\theta(0))\nabla_\theta u(x',\theta(0))^\top\| \leq \sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}.$$

Now putting all these bounds together, we get that

$$\|I_{3,p}\|_{L^2(p+q)} \leq t\cdot\sqrt{2}L_1^3\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}.$$

For ease later on, let us define

$$g_3(t,n)=t\cdot\sqrt{2}L_1^3\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n)+\log(2M_\Theta)}{n}}.$$

Note that because we are working under event $A_P \cap A_Q$, we know that with probability greater than $1-n_p^{-A}-{n_q}^{-A}$

$$\|I_{3,q}\|_{L^2(p+q)} \leq t\cdot\sqrt{2}L_1^3\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_q)+\log(2M_\Theta)}{n_q}} = g_3(t,n_q).$$

Now let us bound $I_{2,p}$ and $I_{2,q}$.

Bounding $I_{2,p}$ and $I_{2,q}$: We will again bound for $I_{2,p}$ and essentially use the same logic for bounding $I_{2,q}$. Note that

$$\begin{aligned}\|I_{2,p}\|^2_{L^2(p+q)} &= \mathbb{E}_{x\sim p+q}\Big|\mathbb{E}_{x'\sim\widehat{p}}(\widehat{K}_t - K_0)(x,x')\bar{e}(x',t)\Big|^2 \\ &\leq \mathbb{E}_{x\sim p+q}\Big(\mathbb{E}_{x'\sim\widehat{p}}|(\widehat{K}_t - K_0)(x,x')||\bar{e}(x',t)|\Big)^2.\end{aligned}$$

Using Lemma 5.21, note that

$$\begin{aligned}|\widehat{K}_t(x,x') - K_0(x,x')| &= |\langle\nabla_\theta u(x,\widehat{\theta}(t)),\nabla_\theta u(x',\widehat{\theta}(t))\rangle - \langle\nabla_\theta u(x,\theta(0)),\nabla_\theta u(x',\theta(0))\rangle| \\ &\leq \|\nabla_\theta u(x,\widehat{\theta}(t))\|_\Theta\|\|\nabla_\theta u(x',\widehat{\theta}(t)) - \nabla_\theta u(x',\theta(0))\|_\Theta \\ &+ \|\nabla_\theta u(x,\widehat{\theta}(t)) - \nabla_\theta u(x,\theta(0))\|_\Theta\|\nabla_\theta u(x',\theta(0))\|_\Theta \\ &\leq 2L_1L_2\|\widehat{\theta}(t) - \theta(0)\|_\Theta \leq 2L_1L_2\sqrt{t}.\end{aligned}$$

Now the only thing left to bound is $\mathbb{E}_{x'\sim\widehat{p}}|\bar{e}(x',t)|$. To do this, recalling the time-integrated form of $\bar{e}$, we have that

$$\mathbb{E}_{x\sim\widehat{p}}|\bar{e}(x,t)| = \|\bar{e}(\cdot,t)\|_{L^1(\widehat{p})} \leq \Big(\|\bar{e}(\cdot,t)\|^2_{L^2(\widehat{p})}\Big)^{1/2}.$$

Now notice that

$$\|\bar{e}(\cdot,t)\|^2_{L^2(\widehat{p})} = \mathbb{E}_{x\sim\widehat{p}}|\bar{e}(x,t)|^2 = (\mathbb{E}_{x\sim\widehat{p}} - \mathbb{E}_{x\sim p})|\bar{e}(x,t)|^2 + \mathbb{E}_{x\sim p}|\bar{e}(x,t)|^2.$$

Because integrating a positive function over both p and q is an upper bound of just integrating over p, we know that

$$\mathbb{E}_{x\sim p}|\bar{e}(x,t)|^2 = \|\bar{e}(\cdot,t)\|^2_{L^2(p)} \leq \|\bar{e}(\cdot,0)\|^2_{L^2(p+q)} = \|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)},$$

so we only need to deal with the first term. In particular, using the time-integrated form of

$|\bar{e}(x,t)|^2$, we get that equals

$$\begin{aligned}\left|(\mathbb{E}_{x\sim\widehat{p}}-\mathbb{E}_{x\sim p})|\bar{e}(x,t)|^2\right| &= \bigg|\int_0^t\int_0^t \mathbb{E}_{y_1,y_2\sim p+q}\bar{e}(y_2,s_2)\nabla_\theta u(y_2,\theta(0))^\top \\ &\cdot\Big[(\mathbb{E}_{x\sim\widehat{p}}-\mathbb{E}_{x\sim p})\nabla_\theta u(x,\theta(0)),\nabla_\theta u(x,\theta(0))^\top\Big]\nabla_\theta u(y_1,\theta(0))\bar{e}(y_1,s_1)ds_1ds_2\bigg| \\ &\le L_1^2\|(\mathbb{E}_{x\sim\widehat{p}}-\mathbb{E}_{x\sim p})\nabla_\theta u(x,\theta(0)),\nabla_\theta u(x,\theta(0))^\top\|\underbrace{\left(\int_0^t\mathbb{E}_{x\sim p+q}|\bar{e}(x,s)|ds\right)^2}_{\le(t\sqrt{2}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)})^2} \\ &\le 2L_1^2t^2\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}\left\|(\mathbb{E}_{x\sim\widehat{p}}-\mathbb{E}_{x\sim p})\nabla_\theta u(x,\theta(0)),\nabla_\theta u(x,\theta(0))^\top\right\|.\end{aligned}$$

Again, since we are under event $A_P\cap A_Q$ we can use Lemma 5.23 and get that with probability greater than $1-n_p^{-A}-n_q^{-A}$

$$\left|(\mathbb{E}_{x\sim\widehat{p}}-\mathbb{E}_{x\sim p})|\bar{e}(x,t)|^2\right|\le 2L_1^2t^2\|\Pi_{K_0}(f^*)\|^2_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}.$$

Plugging this back, we get

$$\mathbb{E}_{x\sim\widehat{p}}|\bar{e}(x,t)|\le\sqrt{2}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\left(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}\right)^{1/2}.$$

Plugging back to our original expression for $I_{2,p}$ and using the fact that $\mathbb{E}_{x\sim p+q}1=2$, we get that

$$\|I_{2,p}\|_{L^2(p+q)}\le 4L_1L_2\sqrt{t}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\left(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}\right)^{1/2}$$

with probability $\ge 1-n_p^{-A}-n_q^{-A}$. Similar to before, we define

$$g_2(t,n)=4L_1L_2\sqrt{t}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\left(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n)+\log(2M_\Theta)}{n}}\right)^{1/2}.$$

Using the same logical reasoning of being in the event $A_P\cap A_Q$, we get that with probability

$\geq 1 - n_p^{-A} - n_q^{-A}$

$$\begin{aligned}\|I_{2,q}\|_{L^2(p+q)} &\leq 4L_1L_2\sqrt{t}\|f^*\|_{L^2(p+q)}\Big(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n_q)+\log(2M_\Theta)}{n_q}}\Big)^{1/2}\\ &= g_2(t,n_Q).\end{aligned}$$

Working with the I_1 terms: Let us again work with $I_{1,p}$ and use the same logic for $I_{1,q}$ later. In particular, note that

$$\begin{aligned}\langle(\widehat{u}-\bar{u})(\cdot,t),I_{1,p}\rangle_{L^2(p+q)} &= \langle(\widehat{u}-\bar{u})(\cdot,t),-\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')(\widehat{e}_P(x',t)-\bar{e}(x',t))\rangle_{L^2(p+q)}\\ &= -\langle(\widehat{u}-\bar{u})(\cdot,t),\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')(\widehat{u}(x',t)-\bar{u}(x',t))\rangle_{L^2(p+q)}\\ &-\langle(\widehat{u}-\bar{u})(\cdot,t),\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')(f^*(x')-1)\rangle_{L^2(p+q)}\\ &\leq -\langle(\widehat{u}-\bar{u})(\cdot,t),\mathbb{E}_{x'\sim\widehat{p+q}}\widehat{K}_t(\cdot,x')(f^*(x')-1)\rangle_{L^2(p+q)},\end{aligned}$$

where we get the inequality because $\widehat{K}_t$ is a positive semi-definite operator so the first term is less than 0. Now we can bound by the following

$$|\langle(\widehat{u}-\bar{u})(\cdot,t),I_{1,p}\rangle_{L^2(p+q)}| \leq \|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}\|\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')(f^*(x')-1)\|_{L^2(p+q)}.$$

Here note that

$$\begin{aligned}
\|\mathbb{E}_{x'\sim\widehat{p}}\widehat{K}_t(\cdot,x')(f^*(x')-1)\|_{L^2(p+q)} &= \|\mathbb{E}_{x'\sim\widehat{p}}\langle\nabla_\theta u(\cdot,\widehat{\theta}(t)),\nabla_\theta u(x',\widehat{\theta}(t))\rangle_\Theta(f^*(x')-1)\|_{L^2(p+q)} \\
&\leq L_1^2\mathbb{E}_{x'\sim\widehat{p}}|f^*(x')-1| \\
&\leq L_1^2\int_{\mathbb{R}^d}\left|\frac{p-q}{p+q}(x)-1\right|d\widehat{p}(x) \\
&= L_1^2\int_{\mathbb{R}^d}\left|\frac{p(x)-q(x)-p(x)-q(x)}{p(x)+q(x)}\right|d\widehat{p}(x) \\
&= L_1^2\int_{\mathbb{R}^d}2\left|\frac{q(x)}{p(x)+q(x)}\right|d\widehat{p}(x) \\
&= 2L_1^2\frac{1}{n_p}\sum_{i=1}^{n_p}\underbrace{\left|\frac{q(x_i)}{p(x_i)+q(x_i)}\right|}_{\leq 1} \\
&\leq 2L_1^2\frac{1}{n_p}\sum_{i=1}^{n_p}1 = 2L_1^2.
\end{aligned}$$

This means that

$$|\langle(\widehat{u}-\bar{u})(\cdot,t),I_{1,p}\rangle_{L^2(p+q)}| \leq 2L_1^2\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}.$$

Using the same logic (but with the term $\frac{p(x)}{p(x)+q(x)}$), we can show that

$$|\langle(\widehat{u}-\bar{u})(\cdot,t),I_{1,q}\rangle_{L^2(p+q)}| \leq 2L_1^2\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}.$$

Putting this altogether, we see that

$$\begin{aligned}\frac{d}{dt}\|(\widehat{u}-\bar{u})(\cdot,t)\|^2_{L^2(p+q)} &\leq |\langle(\widehat{u}-\bar{u})(\cdot,t),I_{1,p}\rangle_{L^2(p+q)}| + |\langle(\widehat{u}-\bar{u})(\cdot,t),I_{1,q}\rangle_{L^2(p+q)}| \\ &+ \|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}\Big(\|I_{2,p}\|_{L^2(p+q)} + \|I_{2,q}\|_{L^2(p+q)} \\ &+ \|I_{3,p}\|_{L^2(p+q)} + \|I_{3,q}\|_{L^2(p+q)}\Big) \\ &= \Bigg(4L_1^2 + g_2(t,n_p) + g_3(t,n_p) \\ &+ g_2(t,n_q) + g_3(t,n_q)\Bigg)\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)}.\end{aligned}$$

Using the same argument in Proposition 5.16, let $t^* \in [0,t]$ be such that

$$\sup_{s\in[0,t]} \|(\widehat{u}-\bar{u})(\cdot,s)\|_{L^2(p+q)} = \|(\widehat{u}-\bar{u})(\cdot,t^*)\|_{L^2(p+q)},$$

then we know that

$$\begin{aligned}\|(\widehat{u}-\bar{u})(\cdot,t^*)\|^2_{L^2(p+q)} &\leq \int_0^{t^*}\Bigg(4L_1^2 + g_2(s,n_p) + g_3(s,n_p) \\ &+ g_2(s,n_q) + g_3(s,n_q)\Bigg)\|(\widehat{u}-\bar{u})(\cdot,s)\|_{L^2(p+q)}ds \\ &\leq \int_0^{t^*}\Bigg(4L_1^2 + g_2(s,n_p) + g_3(s,n_p) \\ &+ g_2(s,n_q) + g_3(s,n_q)\Bigg)\|(\widehat{u}-\bar{u})(\cdot,t^*)\|_{L^2(p+q)}ds \\ &\leq \|(\widehat{u}-\bar{u})(\cdot,t^*)\|_{L^2(p+q)}\int_0^{t^*}\Bigg(4L_1^2 + g_2(s,n_p) + g_3(s,n_p) \\ &+ g_2(s,n_q) + g_3(s,n_q)\Bigg)ds \\ \|(\widehat{u}-\bar{u})(\cdot,t^*)\|_{L^2(p+q)} &\leq \int_0^{t^*}\Bigg(4L_1^2 + g_2(s,n_p) + g_3(s,n_p) + g_2(s,n_q) + g_3(s,n_q)\Bigg)ds.\end{aligned}$$

Now since $t^* \leq t$ and

$$\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)} \leq \|(\widehat{u}-\bar{u})(\cdot,t^*)\|_{L^2(p+q)},$$

we know that

$$\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)} \leq \int_0^{t^*} \bigg(4L_1^2 + g_2(s,n_p) + g_3(s,n_p) + g_2(s,n_q) + g_3(s,n_q)\bigg) ds.$$

Moreover, by inspection, we can see that

$$4L_1^2 + g_2(s,n_p) + g_3(s,n_p) + g_2(s,n_q) + g_3(s,n_q)$$

is monotone in s, which means that

$$\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)} \leq \bigg(4L_1^2 + g_2(s,n_p) + g_3(s,n_p) + g_2(s,n_q) + g_3(s,n_q)\bigg) t.$$

Putting this altogether and using the fact that we are working under the regime of event $A_P \cap A_Q$, we can use Lemma 5.23 for the zero-time NTK for samples from p and q to get that with probability $\geq 1 - n_p^{-A} - n_q^{-A}$, we have

$$\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)} \leq 4L_1^2 t + g_2(s,n_p)t + g_3(s,n_p)t + g_2(s,n_q)t + g_3(s,n_q)t,$$

but the right-hand side of the inequality is just

$$\begin{aligned}
&4L_1^2 t + 4L_1L_2t^{3/2}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\Big(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}\Big)^{1/2}\\
&\quad+4L_1L_2t^{3/2}\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\Big(1+2L_1^3t^2\sqrt{2(2L_1^2+3/2)\frac{A\log(n_q)+\log(2M_\Theta)}{n_q}}\Big)^{1/2}\\
&\quad+t^2\cdot\sqrt{2}L_1^3\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_p)+\log(2M_\Theta)}{n_p}}\\
&\quad+t^2\cdot\sqrt{2}L_1^3\|\Pi_{K_0}(f^*)\|_{L^2(p+q)}\sqrt{2L_1^2(2L_1^2+3/2)\frac{A\log(n_q)+\log(2M_\Theta)}{n_p}}.
\end{aligned}$$

This means that

$$\|(\widehat{u}-\bar{u})(\cdot,t)\|_{L^2(p+q)} = O(t^{5/2}).$$

Moreover, we get the result using the fact that

$$\|\Pi_{K_0}(f^*)\|_{L^2(p+q)} \le \|f^*\|_{L^2(p+q)}. \qquad \square$$

Proof of Proposition 5.26. Consider following calculation

$$\begin{aligned}
\left|\widehat{T}_{test}(t)-\overline{T}(t)\right| &= \left|\int_{\mathbb{R}^d}\widehat{u}(x,t)d(\widehat{p}_{test}-\widehat{q}_{test})(x) - \int_{\mathbb{R}^d}\bar{u}(x,t)d(p-q)(x)\right|\\
&= \bigg|\int_{\mathbb{R}^d}\widehat{u}(x,t)d(\widehat{p}_{test}-\widehat{q}_{test})(x) - \int_{\mathbb{R}^d}\widehat{u}(x,t)d(p-q)(x)\\
&\quad+\int_{\mathbb{R}^d}\widehat{u}(x,t)d(p-q)(x) - \int_{\mathbb{R}^d}\bar{u}(x,t)d(p-q)(x)\bigg|\\
&\le \underbrace{\left|\int_{\mathbb{R}^d}\widehat{u}(x,t)d\big[(\widehat{p}_{test}-p)+(q-\widehat{q}_{test})\big](x)\right|}_{A_1}\\
&\quad+\underbrace{\left|\int_{\mathbb{R}^d}(\widehat{u}-\bar{u})(x,t)d(p-q)(x)\right|}_{A_2}.
\end{aligned}$$

Let us deal with A_2 first and then with A_1. Note that

$$\begin{aligned} A_2 &= \left| \int_{\mathbb{R}^d} (\widehat{u}-\bar{u})(x,t)dp(x) - \int_{\mathbb{R}^d} (\widehat{u}-\bar{u})(x,t)dq(x) \right| \\ &\le \left| \int_{\mathbb{R}^d} (\widehat{u}-\bar{u})(x,t)dp(x) \right| + \left| \int_{\mathbb{R}^d} (\widehat{u}-\bar{u})(x,t)dq(x) \right| \\ &\le \int_{\mathbb{R}^d} |(\widehat{u}-\bar{u})(x,t)| d(p+q)(x) \\ &\le \sqrt{2} \|\widehat{u}-\bar{u}(\cdot,t)\|_{L^2(p+q)}. \end{aligned}$$

So we can use Proposition 5.36 for A_2 and will use this as part of the final bound. Notice that A_1 is actually bounds $|\widehat{T}_{test}(t) - \widehat{T}_{pop}(t)|$. Now let us bound A_1. First note that

$$A_1 \le \underbrace{\left| \int_{\mathbb{R}^d} \widehat{u}(x,t) d(\widehat{p}_{test} - p) \right|}_{A_{1,p}} + \underbrace{\left| \int_{\mathbb{R}^d} \widehat{u}(x,t) d(q - \widehat{q}_{test}) \right|}_{A_{1,q}}.$$

To bound $A_{1,p}$ and $A_{1,q}$, we will aim to use Hoeffding's inequality, but we must first show that $|\widehat{u}(x,t)|$ is bounded. To this end, consider the time-integrated form of $\widehat{u}(x,t) = f(x, \widehat{\theta}(t))$. Recalling the density-specific residuals

$$\begin{aligned} \widehat{e}_p(x,t) &= \big(f(x', \widehat{\theta}(t)) - 1\big) \\ \widehat{e}_q(x,t) &= \big(f(x', \widehat{\theta}(t)) + 1\big), \end{aligned}$$

and using Assumption 5.13, we have

$$\begin{aligned} |f(x, \widehat{\theta}(t))| = \Bigg| -\frac{1}{2} \int_0^t \Bigg(& \mathbb{E}_{x' \sim \widehat{p}} \langle \nabla_\theta f(x, \widehat{\theta}(s)), \nabla_\theta f(x', \widehat{\theta}(s)) \rangle_\Theta \widehat{e}_p(x', s) \\ & + \mathbb{E}_{x' \sim \widehat{q}} \langle \nabla_\theta f(x, \widehat{\theta}(s)), \nabla_\theta f(x', \widehat{\theta}(s)) \rangle_\Theta \widehat{e}_q(x', s) \Bigg) ds \Bigg|. \end{aligned}$$

It is important to note that in the equation above, $\widehat{p}$ and $\widehat{q}$ are training datasets (**not** $\widehat{p}_{test}$ and $\widehat{q}_{test}$), and with this in mind, we continue as

$$
\begin{aligned}
|f(x,\widehat{\theta}(t))| \leq \frac{1}{2}\int_0^t \mathbb{E}_{x'\sim\widehat{p}} & \underbrace{|\langle \nabla_\theta f(x,\widehat{\theta}(s)), \nabla_\theta f(x',\widehat{\theta}(s))\rangle_\Theta|}_{\leq L_1^2} |\widehat{e}_p(x',s)| \\
& + \mathbb{E}_{x'\sim\widehat{q}} \underbrace{|\langle \nabla_\theta f(x,\widehat{\theta}(s)), \nabla_\theta f(x',\widehat{\theta}(s))\rangle_\Theta|}_{\leq L_1^2} |\widehat{e}_q(x',s)| ds \\
& \leq \frac{1}{2} L_1^2 \int_0^t \mathbb{E}_{x'\sim\widehat{p}} |\widehat{e}_p(x',s)| + \mathbb{E}_{x'\sim\widehat{q}} |\widehat{e}_q(x',s)| ds \\
& = \frac{1}{2} L_1^2 \int_0^t \left(\int_{\mathbb{R}^d} |f(x,\widehat{\theta}(s)) - 1| d\widehat{p}(x) + \int_{\mathbb{R}^d} |f(x,\widehat{\theta}(s)) + 1| d\widehat{q}(x) \right) ds.
\end{aligned}
$$

Using Lemma 5.37 with $a(x,t) = f(x,\widehat{\theta}(s)) - 1$ and $b(x,t) = f(x,\widehat{\theta}(s)) + 1$, we know that the right hand side of the equation above is decreasing if $\widehat{L}(\widehat{\theta}(s))$ is decreasing. Indeed, recall

$$\frac{d}{dt}\widehat{L}(\widehat{\theta}(t)) = \langle \nabla_\theta \widehat{L}(\widehat{\theta}(t)), \dot{\widehat{\theta}}(t)\rangle_\Theta = -\|\nabla_\theta \widehat{L}(\widehat{\theta}(t))\|_\Theta \leq 0.$$

This means that

$$
\begin{aligned}
& \int_{\mathbb{R}^d} |f(x,\widehat{\theta}(s)) - 1| d\widehat{p}(x) + \int_{\mathbb{R}^d} |f(x,\widehat{\theta}(s)) + 1| d\widehat{q}(x) \\
\leq & \int_{\mathbb{R}^d} |f(x,\widehat{\theta}(0)) - 1| d\widehat{p}(x) + \int_{\mathbb{R}^d} |f(x,\widehat{\theta}(0)) + 1| d\widehat{q}(x) = 2.
\end{aligned}
$$

Plugging this back in, we get that

$$|f(x,\widehat{\theta}(t))| \leq L_1^2 t.$$

Because we have boundedness, we can use Theorem 5.38. Reworking the probability and lower bound in Hoeffding's inequality, we see that

$$A_{1,p} \leq L_1^2 t \sqrt{2\frac{A\log(m_p)}{m_p}}$$

with probability $\geq 1 - m_p^{-A}$. Similarly, we get that

$$A_{1,q} \leq L_1^2 t \sqrt{2 \frac{A \log(m_q)}{m_q}}$$

with probability $\geq 1 - m_q^{-A}$. So for both these events to occur together, we can use a probability intersection bound to get that

$$A_1 \leq L_1^2 t \sqrt{2} \left(\sqrt{\frac{A \log(m_p)}{m_p}} + \sqrt{\frac{A \log(m_q)}{m_q}} \right)$$

with probability $\geq 1 - m_p^{-A} - m_q^{-A}$. Coming back to A_2, we know the bound from Proposition 5.36 occurs with probability $\geq 1 - n_p^{-A} - n_q^{-A}$ (the finite-sample training dataset size); thus, to have the bound for A_1 and A_2 simultaneously, we again use an intersection probability bound to get that both events occur simultaneously with probability $\geq 1 - (m_p^{-A} + m_q^{-A} + n_p^{-A} + n_q^{-A})$. Putting this altogether, we see that with probability $\geq 1 - (m_p^{-A} + m_q^{-A} + n_p^{-A} + n_q^{-A})$ we have

$$\begin{aligned} \left| \widehat{T}_{test}(t) - \overline{T}(t) \right| &\leq C_{1 L_1, A, m_p, m_q} t + C_{L_1, L_2, f^*} t^{3/2} \\ &\quad + C_{L_1, f^*, n_p, n_q, M_\Theta, A} t^2 \\ &\quad + C_{L_1, L_2, f^*, n_p, n_q, M_\Theta, A} t^{5/2}, \end{aligned}$$

where the constants can be recovered by putting the bound for A_1 together with Proposition 5.36. So we're done. □

Proof of Theorem 5.28. Recall that the loss $\widehat{L}(\widehat{\theta}(s))$ is monotonically decreasing because

$$\frac{d}{dt} \widehat{L}(\widehat{\theta}(t)) = \langle \nabla_\theta \widehat{L}(\widehat{\theta}(t)), \widehat{\theta}(t) \rangle_\Theta = -\|\nabla_\theta \widehat{L}(\widehat{\theta}(t))\|_\Theta \leq 0.$$

Now since the loss

$$\widehat{L}(\widehat{\theta}(s)) = \int_{\mathbb{R}^d} |\widehat{u}(x,s) - 1|^2 d\widehat{p}(x) + \int_{\mathbb{R}^d} |\widehat{u}(x,s) + 1|^2 d\widehat{q}(x)$$

is decreasing, we can use Lemma 5.37 applied to $\widehat{L}(\widehat{\theta}(s))$ to see that

$$\int_{\mathbb{R}^d} |\widehat{u}(x,s) - 1| d\widehat{p}(x) + \int_{\mathbb{R}^d} |\widehat{u}(x,s) + 1| d\widehat{q}(x)$$

is actually monotonically decreasing. Notice that because $|\widehat{u}(x,s)| \leq 1$ on $[0,\widehat{\tau}]$, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} |\widehat{u}(x,s) - 1| d\widehat{p}(x) &= \int_{\mathbb{R}^d} (1 - \widehat{u}(x,s)) \\
\int_{\mathbb{R}^d} |\widehat{u}(x,s) + 1| d\widehat{q}(x) &= \int_{\mathbb{R}^d} (\widehat{u}(x,s) + 1) d\widehat{q}(x).
\end{aligned}$$

So putting this back into the definition of monotonically decreasing loss, we see that

$$\begin{aligned}
&\int_{\mathbb{R}^d} 1 - \widehat{u}(x,s) d\widehat{p}(x) + \int_{\mathbb{R}^d} \widehat{u}(x,s) + 1 d\widehat{q}(x) \\
&\geq \int_{\mathbb{R}^d} 1 - \widehat{u}(x,t) d\widehat{p}(x) + \int_{\mathbb{R}^d} \widehat{u}(x,t) + 1 d\widehat{q}(x) \\
\implies &\int_{\mathbb{R}^d} \widehat{u}(x,t) d(\widehat{p} - \widehat{q})(x) \geq \int_{\mathbb{R}^d} \widehat{u}(x,s) d(\widehat{p} - \widehat{q})(x).
\end{aligned}$$

This implies that $\widehat{T}_{train}(t)$ is monotonically increasing. So we're done. □

Proof of Theorem 5.30. The proof is identical to the case with u. In particular, because we have

$$\max(R^2, \widehat{\tau}) \geq t \geq t_1^*(\varepsilon),$$

we can use Proposition 5.26, Corollary 5.10, and Theorem 5.28 simultaneously. With probability

$\geq 1 - 2(n_p^{-A} + n_q^{-A})$, using the reverse triangle inequality and montonicity gives us

$$\begin{aligned}|\widehat{T}_{train}(t)| &\geq |\widehat{T}_{train}(t_1^*(\varepsilon))| \\ &\geq \left| \left|\overline{T}(t_1^*(\varepsilon))\right| - \left|\widehat{T}_{train}(t_1^*(\varepsilon)) - \overline{T}(t_1^*(\varepsilon))\right| \right| \\ &\geq \varepsilon - \delta_{train}(t_1^*(\varepsilon))\end{aligned}$$

where we can rid of the absolute values by assumption. Now, note that if we assumed that

$$\varepsilon > \delta_{train}(t_1^*(\varepsilon)) + L_1^2 t\sqrt{2}\Bigg(\sqrt{A\log(m_p)/m_p} + \sqrt{A\log(m_q)/m_q} \\ \sqrt{A\log(n_p)/n_p} + \sqrt{A\log(n_q)/n_q}\Bigg),$$

then we would have

$$\begin{aligned}|\widehat{T}_{test}(t)| &\geq \left| |\widehat{T}_{train}(t)| - |\widehat{T}_{test}(t) - \widehat{T}_{train}(t)| \right| \\ &\geq \varepsilon - \delta_{train}(t_1^*(\varepsilon)) - L_1^2 t\sqrt{2}\Bigg(\sqrt{A\log(m_p)/m_p} + \sqrt{A\log(m_q)/m_q} \\ &+ \sqrt{A\log(n_p)/n_p} + \sqrt{A\log(n_q)/n_q}\Bigg)\end{aligned}$$

Similarly, if we assume that

$$\varepsilon > \delta_{train}(t_1^*(\varepsilon)) + L_1^2 t\sqrt{2}\Bigg(\sqrt{A\log(n_p)/n_p} + \sqrt{A\log(n_q)/n_q}\Bigg),$$

then we have

$$\begin{aligned}|\widehat{T}_{pop}(t)| &\geq \left| |\widehat{T}_{train}(t)| - |\widehat{T}_{pop}(t) - \widehat{T}_{train}(t)| \right| \\ &\geq \varepsilon - \delta_{train}(t_1^*(\varepsilon)) - L_1^2 t\sqrt{2}\Bigg(\sqrt{A\log(n_p)/n_p} + \sqrt{A\log(n_q)/n_q}\Bigg).\end{aligned}$$

So we're done. □

Proof of Theorem 5.31. Because of the conditions on t, we can use all of Corollary 5.10, Proposition 5.26, and Proposition 5.25 simultaneously. So essentially, we can use the triangle inequality to get

$$|\widehat{T}(t)| \leq \left|\overline{T}(t)\right| + \left|\widehat{T}(t) - \overline{T}(t)\right| \leq \varepsilon + \delta(t) \leq \varepsilon + \delta(t_2^*(\varepsilon)),$$

where, in general, $\widehat{T}(t)$ and $\delta(t)$ can be replaced by $\widehat{T}_{train}, \widehat{T}_{test}(t), \widehat{T}_{pop}(t)$ and $\delta_{train}(t), \delta_{test}(t), \delta_{pop}(t)$, respectively. These situations happen with probability $\geq 1 - 2(n_p^{-A} + n_q^{-A})$, $\geq 1 - (n_p^{-A} + n_q^{-A} + m_p^{-A} + m_q^{-A})$, and $\geq 1 - (n_p^{-A} + n_q^{-A})$, respectively. So we're done. □

Proof of Corollary 5.33. We will first work with the time associated with detecting deviation ε under the null hypothesis, and then we consider time associated with detecting ε under the assumption that f^* lies on the first k eigenfunctions of K_0. After both these detection times are studied, we study when they are well-separated.

Null Hypothesis: We first note that if we are in the null hypothesis so that $p = q$, then $f^* = 0$, which implies that $\|f^*\|_{L^2(p+q)} = \|\Pi_{K_0}(f^*)\|_{L^2(p+q)} = 0$. Looking into the proof of Proposition 5.26 and Proposition 5.25, we see that the only term that does not depend on f^* is of the form C^+t but C^+ changes depending on which dataset the two-sample test is evaluated on. In particular, we specify

$$C^+ = \begin{cases} \sqrt{2}L_1^2 4 & \widehat{T}_{pop}(t) \text{ evaluation} \\ \sqrt{2}L_1^2\left(4 + \sqrt{A\frac{\log(n_p)}{n_p}} + \sqrt{A\frac{\log(n_q)}{n_q}}\right) & \widehat{T}_{train}(t) \text{ evaluation} \\ \sqrt{2}L_1^2\left(4 + \sqrt{A\frac{\log(m_p)}{m_p}} + \sqrt{A\frac{\log(m_q)}{m_q}}\right) & \widehat{T}_{test}(t) \text{ evaluation.} \end{cases}$$

This means that under the null hypothesis $p = q$ and with either $\widehat{T}_{pop}, \widehat{T}_{test}$, or $\widehat{T}_{train}$ determining

C^+, if

$$t^+(\varepsilon) \geq \frac{\varepsilon}{C^+},$$

then we cannot trust the neural network two-sample test statistic past the time threshold $t^+(\varepsilon)$. Note that as $n_p, n_q, m_p, m_q \to \infty$, the threshold for t to cross becomes $\frac{\varepsilon}{4\sqrt{2}L_1^2}$ and reverts back to the constant C^+ in the case we use $\widehat{T}_{pop}(t)$.

Assumption $\Pi_{K_0}(f^*) = f_k^*$: Recall that we are dealing with the case that $\Pi_{K_0}(f^*) = f_k^*$ so that $\Pi_{K_0}(f^*)$ nontrivially projects onto *only* the first k eigenfunctions. To deal with the time-approximation error $\delta(t)$, we will consider the detection time needed for 2ε and conduct analysis for this case. If we are in the assumption $\Pi_{K_0}(f^*) = f_k^*$, notice that the minimum time needed for the zero-time NTK dynamics to detect a deviation 2ε from Corollary 5.10 is given by

$$t_1^*(2\varepsilon) = \min_{S \in \mathcal{S}_1(\varepsilon)} \lambda_{\min}(S) \log\left(\frac{\|f_k^*\|_S^2}{\|f_k^*\|_S^2 - 2\varepsilon}\right).$$

Importantly, if we want to counteract the approximation error from Proposition 5.26 and Proposition 5.25, we simply need to make sure $\delta(t_1^*(2\varepsilon)) < \varepsilon$ so that the total detection will be $2\varepsilon - \delta(t_1^*(2\varepsilon)) > \varepsilon$, where δ will be $\delta_{pop}, \delta_{test}$, or δ_{train}. Notice from the form of time-approximation error function $\delta(t)$, we have

$$C^- \min\{t, t^{5/2}\} \leq \delta(t) \leq C^- \max\{t, t^{5/2}\},$$

where $C^- = C^+ + C_2 + C_3 + C_4$ with the constants coming from Proposition 5.26, Proposition 5.25, and C^+ defined above. Thus, note that C^- depends on whether we use the two-sample test $\widehat{T}_{pop}, \widehat{T}_{test}$, or $\widehat{T}_{train}$. Assuming the specific assumption that f^* nontrivially projects only on the first k eigenfunctions of K_0 so that $\Pi_{K_0}(f^*) = f_k^*$, notice that

$$t_1^*(2\varepsilon) = \min_{S \in \mathcal{S}_1(\varepsilon)} \lambda_{\min}(S) \log\left(\frac{\|f_k^*\|_S^2}{\|f_k^*\|_S^2 - 2\varepsilon}\right)$$
$$\leq \lambda_k \log\left(\frac{\|f_k^*\|_{L^2(p+q)}^2}{\|f_k^*\|_{L^2(p+q)}^2 - 2\varepsilon}\right) := t_k^-(2\varepsilon).$$

With this in mind, notice that

$$\delta(t_1^*(2\varepsilon)) \leq C^- \max\{t_1^*(2\varepsilon), (t_1^*(2\varepsilon))^{5/2}\} \leq C^- \max\{t_k^-(2\varepsilon), (t_k^-(2\varepsilon))^{5/2}\}$$

so we only need to ensure

$$C^- \max\{t_k^-(2\varepsilon), (t_k^-(2\varepsilon))^{5/2}\} \leq \varepsilon.$$

Rearranging this formula and plugging in the expression for $t_k^-(2\varepsilon)$, we see that our condition above is ensured if

$$\|f_k^*\|_{L^2(p+q)}^2 \geq \max_{a \in \{1,5/2\}} \frac{2\varepsilon \exp\left((\varepsilon/C^-)^{1/a}/\lambda_k\right)}{\exp\left((\varepsilon/C^-)^{1/a}/\lambda_k\right) - 1},$$

which is our assumption.

Separation of null and assumption $\Pi_{K_0}(f^*) = f_k^*$ times: Finally, we want to ensure that the time needed $t^+(\varepsilon) - t^-(\varepsilon) \geq \gamma > 0$ for some. Noting the lower and upper bounds on $t^+(\varepsilon)$ and $t^-(\varepsilon)$, respectively, we find that our condition will be satisfied if

$$t^+(\varepsilon) - t^-(\varepsilon) \geq \frac{\varepsilon}{C^+} - \lambda_k \log\left(\frac{\|f_k^*\|_{L^2(p+q)}^2}{\|f_k^*\|_{L^2(p+q)}^2 - 2\varepsilon}\right) \geq \gamma.$$

Rewriting this inequality, we see that it is satisfied when

$$\|f_k^*\|_{L^2(p+q)}^2 \geq \frac{2\varepsilon \exp\Big((\varepsilon/C^+ - \gamma)/\lambda_k\Big)}{\exp\Big((\varepsilon/C^+ - \gamma)/\lambda_k\Big) - 1}.$$

As this is an assumption, we see that we are done. □

5.12 Helper Lemmas

Lemma 5.37. *Let $a(x,t), b(x,t) : \mathbb{R}^d \times [0,\infty) \to \mathbb{R}$ be differentiable functions in t and let $d\widehat{p}(x)$ and $d\widehat{q}(x)$ be discrete probability measures supported only on a finite number of Dirac masses. Then*

$$g(t) = \int_{\mathbb{R}^d} |a(x,t)|^2 d\widehat{p}(x) + \int_{\mathbb{R}^d} |b(x,t)|^2 d\widehat{q}(x)$$

is decreasing if and only if

$$h(t) = \int_{\mathbb{R}^d} |a(x,t)| d\widehat{p}(x) + \int_{\mathbb{R}^d} |b(x,t)| d\widehat{q}(x)$$

is decreasing.

Proof. We will take the derivatives of both $g(t)$ and $h(t)$ with respect to time and compare them, but we will restrict the integrals to $\text{supp}(\widehat{p})_+ = \{x \in \text{supp}(\widehat{p}) : |a(x,t)| > 0\}$ and $\text{supp}(\widehat{q})_+ = \{x \in$

$\mathrm{supp}(\widehat{q}) : |b(x,t)| > 0\}$. In particular, consider

$$\begin{aligned}
\frac{d}{dt}g(t) &= \frac{d}{dt}\left(\int_{\mathbb{R}^d}|a(x,t)|^2 d\widehat{p}(x) + \int_{\mathbb{R}^d}|b(x,t)|^2 d\widehat{q}(x)\right)\\
&= \int_{\mathrm{supp}(\widehat{p})_+}\partial_t|a(x,t)|^2 d\widehat{p}(x) + \int_{\mathrm{supp}(\widehat{q})_+}\partial_t|b(x,t)|^2 d\widehat{q}(x)\\
&= \int_{\mathrm{supp}(\widehat{p})_+} 2|a(x,t)|\mathrm{sgn}(a(x,t))\partial_t a(x,t) d\widehat{p}(x)\\
&+ \int_{\mathrm{supp}(\widehat{q})_+} 2|b(x,t)|\mathrm{sgn}(b(x,t))\partial_t b(x,t) d\widehat{q}(x).
\end{aligned}$$

For $h(t)$, we get

$$\begin{aligned}
\frac{d}{dt}h(t) &= \frac{d}{dt}\left(\int_{\mathbb{R}^d}|a(x,t)|d\widehat{p}(x) + \int_{\mathbb{R}^d}|b(x,t)|d\widehat{q}(x)\right)\\
&= \int_{\mathrm{supp}(\widehat{p})_+}\partial_t|a(x,t)|d\widehat{p}(x) + \int_{\mathrm{supp}(\widehat{q})_+}\partial_t|b(x,t)|d\widehat{q}(x)\\
&= \int_{\mathrm{supp}(\widehat{p})_+}\mathrm{sgn}(a(x,t))\partial_t a(x,t) d\widehat{p}(x) + \int_{\mathrm{supp}(\widehat{q})_+}\mathrm{sgn}(b(x,t))\partial_t b(x,t) d\widehat{q}(x).
\end{aligned}$$

Because we are using points only in $\mathrm{supp}(\widehat{p})_+$ and $\mathrm{supp}(\widehat{q})_+$ and since the supports of $d\widehat{p}$ and $d\widehat{q}$ are discrete measures, we can define

$$\begin{aligned}
C(t) &= 2\max\left\{\max_{x\in\mathrm{supp}(\widehat{p})_+}|a(x,t)|, \max_{x\in\mathrm{supp}(\widehat{q})_+}|b(x,t)|\right\} > 0\\
c(t) &= 2\min\left\{\min_{x\in\mathrm{supp}(\widehat{p})_+}|a(x,t)|, \min_{x\in\mathrm{supp}(\widehat{q})_+}|b(x,t)|\right\} > 0.
\end{aligned}$$

Notice that the assumption that $c(t) > 0$ heavily depends on that the measures $d\widehat{p}$ and $d\widehat{q}$ are composed of a finite number of Dirac measures. Now, notice that

$$\begin{aligned}\frac{d}{dt}g(t) &= \int_{\text{supp}(\widehat{p})_+} \underbrace{2|a(x,t)|}_{\substack{\le C(t)\\ \ge c(t)}} \text{sgn}(a(x,t))\partial_t a(x,t) d\widehat{p}(x)\\ &+ \int_{\text{supp}(\widehat{q})_+} \underbrace{2|b(x,t)|}_{\substack{\le C(t)\\ \ge c(t)}} \text{sgn}(b(x,t))\partial_t b(x,t) d\widehat{q}(x)\\ \implies c(t)\frac{d}{dt}h(t) &\le \frac{d}{dt}g(t) \le C(t)\frac{d}{dt}h(t).\end{aligned}$$

Since $\frac{d}{dt}h(t)$ and $\frac{d}{dt}g(t)$ are off by positive factors, we see that if one is decreasing, the other must also be decreasing. This proves the lemma. □

5.12.1 Concentration Inequalities

Theorem 5.38 (Hoeffding's Inequality)**.** *Suppose $\{X_i\}_{i=1}^n$ are independent random variables with $|X_i| \le L$, then for all $t \ge 0$*

$$Pr\bigg\{\Big|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\Big| \ge t\bigg\} \le 2\exp\left(-\frac{nt^2}{2L^2}\right).$$

Theorem 5.39 (Hoeffding's Subgaussian Inequality)**.** *Suppose $\{X_i\}_{i=1}^n$ are independent σ_i-subgaussian random variables with X_i having mean μ, then for all $t \ge 0$*

$$Pr\bigg\{\Big|\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i)\Big| \ge t\bigg\} \le 2\exp\left(-\frac{t}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Theorem 5.40 (Matrix Bernstein)**.** *Let X_i be a sequence of n independent, random, real-valued matrices of size d_1-by-d_2. Assume that $\mathbb{E}X_i = 0$ and $\|X_i\| \le L$ for each i and $\nu > 0$ be such that*

$$\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_iX_i^\top\|, \|\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i^\top X_i\| \le \nu.$$

Then for any $t \geq 0$,

$$Pr\Big[\|\frac{1}{n}\sum_{i=1}^{n} X_i\| \geq t\Big] \leq (d_1 + d_2)\exp\Big\{-\frac{nt^2}{2(\nu + Lt/3)}\Big\}.$$

5.13 Acknowledgements

This chapter, in full, is a reprint of *A Neural Network Two-Sample Test* (2024) by V. Khurana, A. Cloninger, and X. Cheng as it is due to appear. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] A. Aldroubi, S. Li, and G. K. Rohde. Partitioning signal classes using transport transforms for data analysis and machine learning. *Sampl. Theory Signal Process. Data Anal.*, 19(6), 2021.

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

[3] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in Neural Information Processing Systems*, 2017-December:1965–1975, 2017.

[4] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1961–1971, Red Hook, NY, USA, 2017. Curran Associates Inc.

[5] L. Ambrosio and N. Gigli. *A User's Guide to Optimal Transport*, pages 1–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[6] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[7] L. Ambrosio, N. Gigli, and S. Giuseppe. *Gradient flows in metric spaces and in the space of probability measures*. BirkhaHERE!HERE!user, 2005.

[8] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[9] E. Arias-Castro, A. Javanmard, and B. Pelletier. Perturbation bounds for procrustes, classical scaling, and trilateration, with applications to manifold learning. *Journal of machine learning research*, 21, 2020.

[10] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of Machine Learning Research*, volume 70, pages 214–223. PMLR, 2017.

[11] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman. *Asymptotic Geometric Analysis, Part I*, volume 202. American Mathematical Soc., 2015.

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

[13] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[14] O. Bencheikh and B. Jourdain. Approximation rate in wasserstein distance of probability measures on the real line by deterministic empirical measures. *Journal of Approximation Theory*, 274:105684, 2022.

[15] R. Berman. Convergence rates for discretized Monge–Ampère equations and quantitative stability of optimal transport. *Found Comput Math*, 21:1099–1140, 2021.

[16] G. L. Bradley, Z. Babutsidze, A. Chai, and J. P. Reser. The role of climate change risk perception, response efficacy, and psychological adaptation in pro-environmental behavior: A two nation study. *Journal of Environmental Psychology*, 68:101410, 2020.

[17] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

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

[19] R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, and G. P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, 2014.

[20] K. Bucci, M. Tulio, and C. Rochman. What is known and unknown about the effects of plastic pollution: A meta-analysis and systematic review. *Ecological Applications*, 30(2):e02044, 2020.

[21] J. Bucklew and G. Wise. Multidimensional asymptotic quantization theory with r-th power distortion measures. *IEEE Transactions on Information Theory*, 28(2):239–247, 1982.

[22] A. C. Burns and A. Veeck. *Marketing research*. Pearson, 2020.

[23] L. A. Caffarelli. Boundary regularity of maps with convex potentials. *Communications on Pure and Applied Mathematics*, 45(9):1141–1151, 1992.

[24] L. A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[25] L. A. Caffarelli. Boundary regularity of maps with convex potentials–II. *Annals of Mathematics*, 144(3):453–496, 1996.

[26] G. D. Cañas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In *NIPS*, 2012.

[27] Y. Chen, F. D. Cruz, R. Sandhu, A. L. Kung, P. Mundi, J. O. Deasy, and A. Tannenbaum. Pediatric sarcoma data forms a unique cluster measured via the earth mover's distance. *Scientific Reports*, 7(1):7035, 2017.

[28] X. Cheng and A. Cloninger. Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory*, 68(10):6631–6662, Oct. 2022.

[29] X. Cheng, A. Cloninger, and R. R. Coifman. Two-sample statistics based on anisotropic kernels. *Information and Inference: A Journal of the IMA*, 9(3):677–719, 2019.

[30] X. Cheng and Y. Xie. Neural tangent kernel maximum mean discrepancy. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6658–6670. Curran Associates, Inc., 2021.

[31] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming, 2020.

[32] A. Cloninger, K. Hamm, V. Khurana, and C. Moosmüller. Linearized wasserstein dimensionality reduction with approximation guarantees, 2023. arXiv:2302.07373.

[33] A. Cloninger, B. Roy, C. Riley, and H. M. Krumholz. People mover's distance: Class level geometry using fast pairwise data adaptive transportation costs. *Applied and Computational Harmonic Analysis*, 47(1):248–257, 2019.

[34] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[35] M. A. Cox and T. F. Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.

[36] M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *NIPS*, volume 2, page 4, 2013.

[37] S. Dara, S. Dhamercherla, S. Jadav, et al. Machine learning in drug discovery: A review. *Artificial Intelligence Review*, 55:1947–1999, 2022.

[38] N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.

[39] A. Delalande and Q. Mérigot. Quantitative stability of optimal transport maps under variations of the target measure. *arXiv preprint arXiv:2103.05934*, 2021.

[40] L. Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[41] S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 1183–1203, 2013.

[42] V. Divol. A short proof on the rate of convergence of the empirical measure for the wasserstein distance, 2021.

[43] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[44] N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure, 2013.

[45] N. Gigli. On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54(2):401–409, 2011.

[46] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer, 2007.

[47] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773, mar 2012.

[48] K. Hamm, N. Henscheid, and S. Kang. Wassmap: Wasserstein isometric mapping for image manifold learning. *arXiv preprint arXiv:2204.06645*, 2022.

[49] D. Hardin, E. B. Saff, and O. Vlasiuk. Asymptotic properties of short-range interaction functionals. *arXiv preprint arXiv:2010.11937*, 2020.

[50] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009.

[51] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[52] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms II: Advanced theory and bundle methods*. Springer Verlag, 1996.

[53] R. A. Horn and C. R. Johnson. *Matrix Analysis, 2nd Ed.* Cambridge University Press, 2012.

[54] J. M. Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011.

[55] L. Kantorovich. On the transfer of masses. *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[56] M. Khoury, Y. Hu, S. Krishnan, and C. Scheidegger. Drawing large graphs by low-rank stress majorization. In *Computer Graphics Forum*, volume 31, pages 975–984. Wiley Online Library, 2012.

[57] V. Khurana, H. Kannan, A. Cloninger, and C. Moosmüller. Supervised learning of sheared distributions using linearized optimal transport. *Sampling Theory, Signal Processing, and Data Analysis*, 21(1), 2023.

[58] M. Kirchler, S. Khorasani, M. Kloft, and C. Lippert. Two-sample testing using deep learning, 2020.

[59] S. Kolouri, S. R. Park, and G. K. Rohde. The radon cumulative distribution transform and its application to image classification. *IEEE Transactions on Image Processing*, 25(2):920–934, 2016.

[60] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[61] X. Liu, Y. Bai, Y. Lu, A. Soltoggio, and S. Kolouri. Wasserstein task embedding for measuring task similarities. *arXiv preprint arXiv:2208.11726*, 2022.

[62] D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests, 2018.

[63] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[64] K. V. Mardia. Multivariate analysis. Technical report, 1979.

[65] J. Mathews, M. Pouryahya, C. Moosmüller, I. G. Kevrekidis, J. O. Deasy, and A. Tannenbaum. Molecular phenotyping using networks, diffusion, and topology: soft-tissue sarcoma. *Scientific Reports*, 9, 2019. Article number: 13982.

[66] J. C. Mathews, M. Pouryahya, C. Moosmüller, Y. G. Kevrekidis, J. O. Deasy, and A. Tannenbaum. Molecular phenotyping using networks, diffusion, and topology: soft tissue sarcoma. *Scientific reports*, 9(1):13982, 2019.

[67] R. J. McCann. Polar factorization of maps on Riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11(3):589–608, 2001.

[68] Q. Mérigot, A. Delalande, and F. Chazal. Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3186–3196. PMLR, 26–28 Aug 2020.

[69] J. Miller, V. Huroyan, and S. Kobourov. Spherical graph drawing by multi-dimensional scaling. *arXiv preprint arXiv:2209.00191*, 2022.

[70] G. Mishne, R. Talmon, R. Meir, J. Schiller, M. Lavzin, U. Dubin, and R. R. Coifman. Hierarchical coupled-geometry analysis for neuronal structure and activity pattern discovery. *IEEE Journal of Selected Topics in Signal Processing*, 10(7):1238–1253, 2016.

[71] C. Moosmüller and A. Cloninger. Linear Optimal Transport Embedding: Provable Wasserstein classification for certain rigid transformations and perturbations. To appear in: Information and Inference: A Journal of the IMA, 2022.

[72] C. Moosmüller and A. Cloninger. Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023.

[73] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. *Kernel Mean Embedding of Distributions: A Review and Beyond.* Now Foundations and Trends, 2017.

[74] M. Mueller, S. Aeron, J. M. Murphy, and A. Tasissa. Geometric sparse coding in Wasserstein space. *arXiv preprint arXiv:2210.12135*, 2022.

[75] E. Negrini and L. Nurbekyan. Applications of no-collision transportation maps in manifold learning. *arXiv preprint arXiv:2304.00199*, 2023.

[76] W. K. Newey and K. D. West. Hypothesis testing with efficient method of moments estimation. *International Economic Review*, pages 777–787, 1987.

[77] V. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer International Publishing, 2020.

[78] S. R. Park, S. Kolouri, S. Kundu, and G. K. Rohde. The cumulative distribution transform and linear pattern classification. *Applied and Computational Harmonic Analysis*, 45(3):616 – 641, 2018.

[79] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[80] A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *arXiv:2109.12004*, 2021.

[81] M. Repasky, X. Cheng, and Y. Xie. Neural stein critics with staged l^2-regularization, 2023.

[82] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[83] S. Singh and B. Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

[84] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348, 1967.

[85] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pages 306–314, 2014.

[86] G. Stockman and L. G. Shapiro. *Computer vision*. Prentice Hall PTR, 2001.

[87] A. Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005 – 1026, 2011.

[88] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[89] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

[90] C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[91] C. Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009.

[92] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, and G. K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE transactions on medical imaging*, 30(3):621–631, 2010.

[93] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, and G. K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE Trans Med Imaging*, 30(3):621–631, 2011.

[94] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int J Comput Vis*, 101:254–269, 2013.

[95] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4 A):2620–2648, 2019.

[96] M. E. Werenski, R. Jiang, A. Tasissa, S. Aeron, and J. M. Murphy. Measure estimation in the barycentric coding model. In *Proceedings of the 39 th International Conference on Machine Learning*, pages 23781–23803. PMLR, 2022.

[97] C. Xu and A. Berger. Best finite constrained approximations of one-dimensional probabilities. *Journal of approximation theory*, 244:1–36, 2019.

[98] G. Young and A. S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

[99] N. Zelesko, A. Moscovich, J. Kileel, and A. Singer. Earthmover-based manifold learning for analyzing molecular conformation spaces. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1715–1719, 2020.

[100] Y. Zhang, R. Jin, and Z.-H. Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

[101] J. Zhao, A. Jaffe, H. Li, O. Lindenbaum, E. Sefik, R. Jackson, X. Cheng, R. A. Flavell, and Y. Kluger. Detection of differentially abundant cell subpopulations in scrna-seq data. *Proceedings of the National Academy of Sciences*, 118(22), 2021.