Information and Inference: A Journal of the IMA (2022) **00**, 1–27
https://doi.org/10.1093/imaiai/iaac023

Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations

Caroline Moosmüller[†]
Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[†]Corresponding author. Email: cmoosm@unc.edu
and
Alexander Cloninger
Department of Mathematics and Halicioğlu Data Science Institute, University of California, San Diego, La Jolla, CA 92093, USA

[Received on 29 May 2021; revised on 8 July 2022; accepted on 15 July 2022]

Discriminating between distributions is an important problem in a number of scientific fields. This motivated the introduction of Linear Optimal Transportation (LOT), which embeds the space of distributions into an L^2-space. The transform is defined by computing the optimal transport of each distribution to a fixed reference distribution and has a number of benefits when it comes to speed of computation and to determining classification boundaries. In this paper, we characterize a number of settings in which LOT embeds families of distributions into a space in which they are linearly separable. This is true in arbitrary dimension, and for families of distributions generated through perturbations of shifts and scalings of a fixed distribution. We also prove conditions under which the L^2 distance of the LOT embedding between two distributions in arbitrary dimension is nearly isometric to Wasserstein-2 distance between those distributions. This is of significant computational benefit, as one must only compute N optimal transport maps to define the N^2 pairwise distances between N distributions. We demonstrate the benefits of LOT on a number of distribution classification problems.

Keywords: optimal transport; linear embedding; Wasserstein distance; classification.

1. Introduction

The problem of supervised learning is most commonly formulated as follows. Given data of the form $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^n$, learn a function $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x_i) \approx y_i$. However, in many applications, the data points are not simply points in $\mathbb{R}^n$ but are instead probability measures μ_i on $\mathbb{R}^n$, or even finite samples $X_i = \{x_j^{(i)}\}_{j=1}^{N_i}$ for $x_j^{(i)} \sim \mu_i$. Applications where this problem arises are surveys broken into demographic or location groups [12], topic modeling from a bag of words model [34] and flow cytometry and other measurements of cell or gene populations per person [7, 11, 35].

The most natural way to solve the supervised learning problem on data $\{(\mu_i, y_i)\}_{i=1}^N$ is to embed μ_i into a Hilbert space and then apply traditional machine learning techniques on this embedding. Simple versions of this embedding would be through moments $\mu_i \mapsto \mathbb{E}_{X\sim\mu_i}[X]$ [26] or through a mean embedding $\mu_i \mapsto \mathbb{E}_{X\sim\mu_i}K(\cdot, X)$ for some kernel K [24]. However, these embeddings either throw away pertinent information about μ_i (e.g. higher order moments) or induce a complex nonlinear geometric relationship between distributions (e.g. $\|\mathbb{E}_{X\sim\mu(x)}K(\cdot, X) - \mathbb{E}_{X\sim\mu(x-\tau)}K(\cdot, X)\| \approx \|\mathbb{E}_{X\sim\mu(x)}K(\cdot, X) - \mathbb{E}_{X\sim\mu(x-2\tau)}K(\cdot, X)\|$ for τ significantly larger than the bandwidth of the kernel). These issues motivate

© The Author(s) 2022. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

the need for a transformation that is both injective and induces a simple geometric structure in the embedding space, so that one can learn an easy classifier.

The natural distance between distributions is Wasserstein-2 distance [32], where the distance between distributions μ and ν is

$$W_2(\mu,\nu)^2 = \min_{T\in\Pi_\mu^\nu} \int \|T(x)-x\|^2 d\mu(x), \tag{1.1}$$

where Π_μ^ν is the collection of all measure preserving maps from μ to ν. The arg min of (1.1) is referred to as the 'optimal transport map' and we denote it by T_μ^ν (see Section 2 for a full description). Wasserstein distance is a more natural distance between distributions as it is a metric on distributions (unlike distances between a finite number of moments as above) and the distance does not saturate as the distributions move further apart (unlike mean embeddings as described above). Optimal transport has been of significant importance in machine learning, including as a cost for generative models [4], natural distances between images [29], pattern discovery for data cubes of neuronal data [23] and general semi-supervised learning [31]. There are two main drawbacks to optimal transport in machine learning. The first is that the computation of each transport map is slow, though this has motivated a number of approximations for computational speed up [13, 19, 30]. The second drawback is that it is difficult to incorporate supervised learning into optimal transport, as the distance is defined for a pre-defined cost function and eq. (1.1), as stated, does not generate a feature embedding of μ and ν that can be fed into traditional machine learning techniques.

This motivated the introduction of Linear Optimal Transportation (LOT) [33], also called Monge embedding in [22]. 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

$$\mu \mapsto T_\sigma^\mu. \tag{1.2}$$

The power of this transform lies in the fact that the nonlinear space of distributions is mapped into the linear space of L^2 functions. In addition, eq. (1.2) is an embedding with convex image.

In one-dimensional space, the optimal transport map is simply the generalized cdf of the distribution (if $\sigma = \text{Unif}([0,1])$, this is exactly the traditional cdf). In [27], the authors define the LOT as the Cumulative Distribution Transform (CDT), and the main theory and applications presented in [27] concern linear separability of data consisting of one-dimensional densities.

However, LOT is more complicated on $\mathbb{R}^n$ for $n > 1$. For $n = 1$, the cdf is the only monotone non-decreasing measure preserving map from μ_i to σ, and thus is the optimal transport between the distributions. Similarly, it can be computed explicitly. This is not the case for $n > 1$: there are a large number of measure preserving maps, with the optimal transport map being the map that requires minimal work, see (1.1). Similarly, there are a much larger family of potential simple continuous perturbations that can be done to μ_i when $n > 1$ (e.g. shearings, rotations) than exist for $n = 1$.

In [17], the CDT is combined with the Radon transform to apply results from [27] in general dimensions $n > 1$. While this construction can be considered a variant of LOT, a linear separability result for LOT in $n > 1$ is still missing. A proof of linear separability in LOT space for $n > 1$ is one of the main contributions of this paper (see Section 1.1).

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

The LOT embedding eq. (1.2) comes with yet another advantage. One can define a distance between two distributions μ_i and μ_j as the L^2-norm of their images under LOT

$$W_2^{\mathrm{LOT}}(\mu_i,\mu_j)^2 := \|T_\sigma^{\mu_i} - T_\sigma^{\mu_j}\|_\sigma^2 = \int \|T_\sigma^{\mu_i}(x) - T_\sigma^{\mu_j}(x)\|^2 \, d\sigma(x).$$

In this paper, we prove that W_2 equals W_2^{LOT} if the family of distributions μ_i is generated by shifts and scalings of a fixed distribution μ. We further show that W_2 is well approximated by W_2^{LOT} for perturbations of shift and scalings (see Section 1.1).

We wish to highlight the computational importance of establishing approximate equivalence between LOT distance and Wasserstein-2 distance. Given N distributions, computing the exact Wasserstein-2 distance between all distributions naively requires computing $\binom{N}{2}$ expensive OT optimization problems. However, if the distributions come from a family of distributions generated by perturbations of shifts and scalings, one can instead compute N expensive OT optimization problems mapping each distribution to σ and compute $\binom{N}{2}$ cheap Euclidean distances between the transport maps, and this provably well approximates the ground truth distance matrix.

1.1 *Main contributions*

The main contributions of this paper are as follows:

- We establish the following with regards to building simple classifiers:

  THEOREM 1.1 (Informal Statement of Theorem 4.1). If $\mathscr{P} = \{\mu_i : y_i = 1\}$ are ε-perturbations of shifts and scalings of μ, and $\mathscr{Q} = \{\nu_i : y_i = -1\}$ are ε-perturbations of shifts and scalings of ν, and $\mathscr{P}$ and $\mathscr{Q}$ have a small minimal distance depending on ε (and satisfy a few technical assumptions), then $\mathscr{P}$ and $\mathscr{Q}$ are linearly separable in the LOT embedding space.

- We establish the following with regards to LOT distance:

  PROPOSITION 1.2 (Informal Statement of Proposition 4.1). If μ and ν are ε-perturbations by shifts and scalings of one another, then

  $$W_2(\mu,\nu) \le W_2^{\mathrm{LOT}}(\mu,\nu) \le W_2(\mu,\nu) + C_\sigma \varepsilon + \overline{C_\sigma}\varepsilon^{1/2}.$$

  In particular, this implies that the LOT embedding is an isometry on the subset of measures related via shifts and scalings, i.e. when $\varepsilon = 0$.

- We demonstrate that in applications to MNIST images, the LOT embedding space is near perfectly linearly separable between classes of images.

2. Preliminaries: optimal mass transport

Let $\mathscr{P}(\mathbb{R}^n)$ be the set of probability measures on $\mathbb{R}^n$. By $\mathscr{P}_2(\mathbb{R}^n)$, we denote those measures in $\mathscr{P}(\mathbb{R}^n)$ with bounded second moment, i.e. $\sigma \in \mathscr{P}(\mathbb{R}^n)$ that satisfy

$$\int \|x\|_2^2 \, d\sigma(x) < \infty.$$

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

For $\sigma \in \mathscr{P}_2(\mathbb{R}^n)$, we also consider the space $L^2(\mathbb{R}^n, \sigma)$ with squared norm

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

In case of the L^2-norm with respect to the Lebesgue measure λ, we simply write $\|f\|$.

For a map $T : \mathbb{R}^n \to \mathbb{R}^n$ and a measure σ we define the *pushforward measure* $T_\sharp\sigma$ by

$$T_\sharp\sigma(A) = \sigma(T^{-1}(A)),$$

where $A \subset \mathbb{R}^n$ is measurable and $T^{-1}(A)$ denotes the preimage of A under T.

If $\sigma \in \mathscr{P}_2(\mathbb{R}^n)$ is absolutely continuous with respect to the Lebesgue measure λ, which we denote by $\sigma \ll \lambda$, then there exists a density $f_\sigma : \mathbb{R}^n \to \mathbb{R}$ such that

$$\sigma(A) = \int_A f_\sigma(x) \, d\lambda(x), \quad A \subseteq \mathbb{R}^n \text{ measurable.} \tag{2.1}$$

In terms of densities, the pushforward relation $\nu(A) = \sigma(T^{-1}(A))$ is given by

$$\int_{T^{-1}(A)} f_\sigma(x) \, d\lambda(x) = \int_A f_\nu(y) \, d\lambda(y), \quad A \subseteq \mathbb{R}^n \text{ measurable.} \tag{2.2}$$

In case the map T is invertible and differentiable, we can rewrite (2.2) as

$$f_\nu(y) = f_\sigma\left(T^{-1}(y)\right) |\det D_y T^{-1}|. \tag{2.3}$$

Given two measures σ, ν, there can exist many maps T that push σ to ν. For that reason, we focus on an additional assumption for the map, namely that it is the optimal transport map between σ and ν [32]. The map T is required to minimize a cost function of the form

$$\int_{\mathbb{R}^n} c(T(x), x) \, d\sigma(x), \tag{2.4}$$

under the constraint $T_\sharp\sigma = \nu$, which is equivalent to $T \in \Pi_\sigma^\nu$. In this paper, we consider the cost $c(x, y) = ||x - y||_2^2$. Other cost functions are possible as well, most notably, p-norms can be studied instead of 2-norms [32]. If the optimization has a solution, then

$$W_2(\sigma, \nu)^2 = \min_{T : T_\sharp\sigma = \nu} \int_{\mathbb{R}^n} \|T(x) - x\|_2^2 \, d\sigma(x)$$

is the 2-*Wasserstein distance* between the measures σ and ν. In this paper, we will refer to W_2 as *the* Wasserstein distance, as we only consider this case. The map T that minimizes (2.4) is called *optimal transport map*.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

We introduce the notation T_σ^ν to denote the optimal transport map from σ to ν. With this notation, we have the identity

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

The minimization involving T might not have a solution, which has led to a relaxation of the formulation of (2.4) introduced by Kantorovich: instead of a transport map T, one seeks a transport plan or coupling $\gamma \in \Gamma_\sigma^\nu$, where Γ_σ^ν is the set of all measures on $\mathbb{R}^n \times \mathbb{R}^n$, whose marginals along the two coordinate directions are σ and ν. The problem now becomes

$$\inf_{\gamma\in\Gamma_\sigma^\nu} \int_{\mathbb{R}^n\times\mathbb{R}^n} c(x,y)d\gamma(x,y),$$

which gives rise to the 2-Wasserstein distance, even when the optimal map does not exist. In case an optimal transport map T_σ^ν does exist, then the optimal coupling has the form $\gamma = (\mathrm{id}, T_\sigma^\nu)_\sharp\sigma$.

We now cite a result concerning existence and uniqueness of the optimal transport map which is used throughout this paper.

THEOREM 2.1 ((6), formulation taken from (28)). Let $\sigma,\nu \in \mathscr{P}_2(\mathbb{R}^n)$ and consider the cost function $c(x,y) = ||x-y||_2^2$. If σ is absolutely continuous with respect to the Lebesgue measure, then there exists a unique map $T \in L^2(\mathbb{R}^n,\sigma)$ pushing σ to ν, which minimizes (2.4). Furthermore, the map T is uniquely defined as the gradient of a convex function φ, $T(x) = \nabla\varphi(x)$, where φ is the unique (up to an additive constant) convex function such that $(\nabla\varphi)_\sharp\sigma = \nu$.

There exist many generalizations of this result, for example to more general cost functions or to Riemannian manifolds [3, 6, 21, 32].

3. LOT and its properties

In this section, we introduce the *LOT* as defined in [33] (also called *Monge embedding* in [22]) and present its basic properties.

LOT is an embedding of $\mathscr{P}_2(\mathbb{R}^n)$ into $L^2(\mathbb{R}^n,\sigma)$ based on a fixed measure σ. It is defined as

$$\nu \mapsto T_\sigma^\nu. \tag{3.1}$$

The power of this embedding lies in the fact that the target space is a Hilbert space, and as we will show, will allow us to use linear hyperplanes to separate naturally clustered sets of measures. This allows to apply linear methods to inherently nonlinear problems in $\mathscr{P}_2(\mathbb{R}^n)$ (see, for example, the application to classification problems in Section 5 and [22, 27]).

The map (3.1) can be thought of as a linearization of the Riemannian manifold $\mathscr{P}_2(\mathbb{R}^n)$ endowed with the Wasserstein distance. The tangent space of $\mathscr{P}_2(\mathbb{R}^n)$ at σ lies in $L^2(\mathbb{R}^n,\sigma)$, hence (3.1) is an inverse to the exponential map [15, 32, 33].

The map eq. (3.1) has been studied by others authors as well, mainly with respect to its regularity. Reference [15] shows 1/2-Hölder regularity of a time-dependent version of eq. (3.1) under regularity assumptions on the measures σ,ν (we discuss this result in Appendix 7.1). Reference [22] prove a weaker Hölder bound, but without any regularity assumptions on the measures. It is also shown in both [15] and [22] that in general, the regularity of (3.1) is not better than 1/2.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

Bounds for a variant of (3.1) in which the source measure, rather than the target measure is varied, can be found in [5].

We now define LOT and summarize its basic properties.

DEFINITION 3.1 (LOT (33)). Fix a measure $\sigma \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$. We define the *LOT* , F_σ, which assigns a function in $L^2(\mathbb{R}^n, \sigma)$ to a measure in $\mathscr{P}_2(\mathbb{R}^n)$

$$F_\sigma(\nu) = T_\sigma^\nu, \quad \nu \in \mathscr{P}_2(\mathbb{R}^n).$$

We now show that LOT is an embedding with convex image.

LEMMA 3.1. For fixed $\sigma \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$, we have the following

1. F_σ embeds $\mathscr{P}_2(\mathbb{R}^n)$ into $L^2(\mathbb{R}^n, \sigma)$;
2. the image $F_\sigma(\mathscr{P}_2(\mathbb{R}^n))$ is convex in $L^2(\mathbb{R}^n, \sigma)$.

Proof. The proof is an application of Theorem 2.1. The first part is also shown in [22]. For the convenience of the reader, we summarize the proof in Section 7.3. □

We introduce a compatibility condition between LOT and the pushforward operator, which is one of the key ingredients for the results in Section 4.

Fix two measures $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n), \sigma \ll \lambda$. F_σ is called *compatible* with μ-pushforwards of a set of functions $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \mu)$ if for every $h \in \mathscr{H}$, we have

$$F_\sigma(h_\sharp \mu) = h \circ F_\sigma(\mu). \tag{3.2}$$

This condition has also been introduced by [1] on the level of densities.

REMARK 3.1. For $\sigma = \mu$, the compatibility condition reads as $T_\sigma^{h_\sharp \sigma} = h$. This means that a function h is required to be the optimal transport from σ to $h_\sharp \sigma$. This is a rather strong condition, and not satisfied for a general function h. In particular, by Brenier's theorem for $\sigma \ll \lambda$, $T_\sigma^{h_\sharp \sigma} = h$ if and only if h is the gradient of a convex function.

The compatibility condition can also be understood in terms of operators. The pushforward operator $h \mapsto h_\sharp \sigma$, which in Riemannian geometry is an exponential map, is left-inverse to F_σ. The compatibility condition requires that it is also right-inverse.

We mention below that the compatibility condition is satisfied for shifts and scalings, a fact also shown in [1] on the level of densities. Reference [1] also prove that shifts and scalings are the only transformation that satisfy (3.2) for all μ.

For $a \in \mathbb{R}^n$ denote by $S_a(x) = a + x$ the shift by a. Similarly, for $c > 0$ denote by $R_c(x) = cx$ the scaling by c. We denote by $\mathscr{E} := \{cx + a : c > 0, a \in \mathbb{R}^n\}$ the group generated by shifts and scalings.

LEMMA 3.2 (Compatibility on $\mathbb{R}$ and with shifts and scalings). Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$.

1. If $n = 1$, i.e. on $\mathbb{R}$, F_σ is compatible with μ-pushforwards of monotonically increasing functions.
2. For general $n \geq 1$, F_σ is compatible with μ-pushforwards of shifts and scalings, i.e. functions in $\mathscr{E}$.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

Proof. The proof is an application of Theorem 2.1 and can be found in [1] (the first part can also be found in [27]); for the convenience of the reader, we show details in Section 7.3. □

4. Geometry of LOT embedding space

In this section, we characterize the geometry of the LOT embedding space under families of compatible transformations in $\mathscr{E}$ (i.e. shifts and scalings), as well as for approximately compatible transformations in $\mathscr{G}_{\lambda,R,\varepsilon}$ (eq. 4.3), where λ denotes the Lebesgue measure.

For a measure μ and a set of functions $\mathscr{H}$, we denote by $\mathscr{H} \star \mu$ the set of all pushforwards of μ under $\mathscr{H}$, i.e.

$$\mathscr{H} \star \mu = \{h_\sharp \mu : h \in \mathscr{H}\}.$$

In this section, we are mainly interested in conditions under which two families of distributions defined by pushforwards of $\mathscr{G} \subset \mathscr{G}_{\lambda,R,\varepsilon}$, $\mathscr{G} \star \mu$ and $\mathscr{G} \star \nu$ are linearly separable in the LOT embedding space.

Before stating the main results of this section, we briefly describe linear separability and its importance in machine learning. Linear separability of two disjoint sets in a Hilbert space implies the existence of a hyperplane $\langle w, x\rangle = b$ such that

$$\begin{aligned} \langle w, \mu_i\rangle &< b, & \forall \mu_i &\in \mathscr{H} \star \mu \\ \langle w, \nu_i\rangle &> b, & \forall \nu_i &\in \mathscr{H} \star \nu. \end{aligned}$$

The existence of such a hyperplane can be established through the Hahn–Banach separation theorem. The theorem simply assumes that the two sets ($\mathscr{H} \star \mu$, $\mathscr{H} \star \nu$) are convex and that one is closed and the other is compact [25].

Linear separability is a strong and important condition for many machine learning applications and supervised learning generally. This is because learning a linear classifier is very straightforward and does not require many training points to accurately estimate w and b. This implies that once the distributions are mapped to the LOT embedding space, it is possible to learn a classifier that perfectly separates the two families with only a small amount of labeled examples.

We note that the result on $\mathscr{G}_{\lambda,\varepsilon,R}$ (Theorem 4.2) is the main result of this section, but we list several other results for completeness. We also note that, for ease of understanding, we frame all theorems in this section for subsets of shifts/scalings or perturbations of such. However, Corollary 4.3 and Theorem 4.2 actually have versions in Appendix 7.2 (Theorems 7.4 and 7.5, respectively) for the family of all approximately compatible transformations. Furthermore, in the case of $\mathscr{G}_{\lambda,\varepsilon,R}$ (Theorem 4.2), through Corollary 7.2, we can give an explicit characterization of the minimal distance δ required between the two families of distributions, $\mathscr{G} \star \mu$ and $\mathscr{G} \star \nu$, to guarantee linear separability.

Finally, both theorems can be strengthened if we make additional assumptions on the regularity of the reference and target distributions. These assumptions, referred to as the Caffarelli's regularity assumptions, are used in Theorem 7.1, but we highlight the assumptions here as well. The assumptions on both the target and source are as follows.

1. $\mathrm{supp}(\sigma), \mathrm{supp}(\mu)$ are C^2 and uniformly convex,
2. for some $\alpha \in (0, 1)$, the densities f_σ, f_μ are $C^{0,\alpha}$ continuous on their supports, and
3. the densities are bounded from above and below.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

By assuming this, we can attain much sharper rates when we move to $\mathscr{G}_{\lambda,\varepsilon,R}$, but weaker versions of the results are similarly established without the regularity conditions.

4.1 *Approximation of the Wasserstein distance*

From Lemma 3.1, we know that LOT embeds $\mathscr{P}_2(\mathbb{R}^n)$ into $L^2(\mathbb{R}^n, \sigma)$. In general, this embedding is not an isometry.

In this section, we derive the error that occurs when approximating the Wasserstein distance by the L^2 distance obtained in the LOT embedding. We are thus interested in the accuracy of the following approximation:

$$W_2(\mu, \nu) \approx \|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma. \tag{4.1}$$

Note that if $W_2(\mu, \nu)$ is approximated well by $\|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma$, LOT is very powerful, as the Wasserstein distance between k different measures can be computed from only k transports instead of $\binom{k}{2}$. Indeed, in this section, we show that (4.1) is exact, i.e. the LOT embedding is an isometry, for two important cases: On $\mathbb{R}$, and on $\mathbb{R}^n$ if both μ and ν are pushforwards of a fixed measure under shifts and scalings. We further show that it is almost exact for pushforwards of functions close to shifts and scalings.

It is important to note that in most applications, distributions are not exact shifts or scalings of one another. In many applications, perturbations such as rotation, stretching, shearing or overall noise are commonly encountered. Thus, it is important to consider the behavior of LOT under such perturbations and demonstrate that the LOT distance continues to be a quasi-isometry with respect to Wasserstein-2 distance and that the deformation constants depend smoothly on the size of the pertubation.

Let $\mu \in \mathscr{P}_2(\mathbb{R}^n)$, $R > 0$ and $\varepsilon > 0$. Recall that we denote by $\mathscr{E} := \{cx + a : c > 0, a \in \mathbb{R}^n\}$ the group generated by shifts and scalings. We define the sets

$$\mathscr{E}_{\mu,R} = \{h \in \mathscr{E} : \|h\|_\mu \leq R\} \tag{4.2}$$

and

$$\mathscr{G}_{\mu,R,\varepsilon} = \{g \in L^2(\mathbb{R}^n, \mu) : \exists h \in \mathscr{E}_{\mu,R} : \|g - h\|_\mu \leq \varepsilon\}. \tag{4.3}$$

This can be thought of as the ε tube around the set of shifts and scalings or as the set of almost compatible transformations.

Proposition 4.1. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n), \sigma, \mu \ll \lambda$. Let $R > 0, \varepsilon > 0$.

1. For $g_1, g_2 \in \mathscr{G}_{\mu,R,\varepsilon}$ and σ the Lebesgue measure on a convex, compact subset of $\mathbb{R}^n$, we have

$$0 \leq \|F_\sigma(g_{1\sharp}\mu) - F_\sigma(g_{2\sharp}\mu)\|_\sigma - W_2(g_{1\sharp}\mu, g_{2\sharp}\mu) \leq C\varepsilon^{\frac{2}{15}} + 2\varepsilon.$$

2. If σ, μ satisfy the assumptions of Caffarelli's regularity theorem (Theorem 7.1), then for $g_1, g_2 \in \mathscr{G}_{\mu,R,\varepsilon}$, we have

$$0 \leq \|F_\sigma(g_{1\sharp}\mu) - F_\sigma(g_{2\sharp}\mu)\|_\sigma - W_2(g_{1\sharp}\mu, g_{2\sharp}\mu) \leq \overline{C}\,\varepsilon^{1/2} + C\,\varepsilon.$$

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

The constants depend on σ, μ and R.

Proof. The main ingredient for these results is Hölder bounds as derived in [15, 22]. We show a detailed proof in Section 7.4. □

We mention that through the application of results derived from [15] (Corollary 7.2), the constants appearing in the second part of this theorem can be characterized explicitly, see Section 7.4.

The theorem states that for functions close to 'ideal' functions (shifts and scalings), the LOT embedding is an almost isometry. Also note the trade-off between Hölder regularity and regularity assumptions on σ, μ: through [22], we can achieve a $2/15$ bound without strong regularity assumptions on σ, μ; the bound improves through [15], when σ, μ are regular in the sense of Theorem 7.1.

Without perturbation, i.e. when $\varepsilon = 0$, Proposition 4.1 implies the followings.

COROLLARY 4.1. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$. Then for $h_1, h_2 \in \mathscr{E}$, we have

$$W_2(h_{1\sharp}\mu, h_{2\sharp}\mu) = \|F_\sigma(h_{1\sharp}\mu) - F_\sigma(h_{2\sharp}\mu)\|_\sigma = \|h_1 - h_2\|_\mu.$$

This means that F_σ restricted to $\mathscr{E} \star \mu := \left\{h_\sharp \mu : h \in \mathscr{E}\right\}$ is an isometry.

We also have the following result, which has also been shown in [27]:

COROLLARY 4.2. On $\mathbb{R}$, F_σ is an isometry.

Proof. We prove in Lemma 7.5 that compatibility of F_σ with μ-pushforwards implies eq. (7.9). Thus, the result follows from Lemma 3.2. □

4.2 *Linear separability results*

We establish the main result of this paper, which covers approximately compatible transforms in $\mathscr{G}_{\lambda,\varepsilon,R}$, the ε-tube around the bounded shifts and scalings $\mathscr{E}_{\lambda,R}$. Theorem 4.2 establishes the case for the tube around $\mathscr{E}_{\lambda,R}$, and Theorem 7.5 (in the appendix) establishes the condition for almost compatible transformations; indeed, Theorem 4.2 follows from the more general result presented in Theorem 7.5. In both cases to show linear separability in the LOT embedding space, one must now assume that the two families of distributions are not just disjoint but actually have a non-trivial minimal distance.

THEOREM 4.2. Let $\sigma, \mu, \nu \in \mathscr{P}_2(\mathbb{R}^n), \sigma, \mu, \nu \ll \lambda$. Let $R > 0, \varepsilon > 0$. Consider $\mathscr{G} \subset \mathscr{G}_{\lambda,R,\varepsilon}$ and let $\mathscr{G}$ be convex. Let $\mathscr{G} \star \mu$ and $\mathscr{G} \star \nu$ be compact. If either

1. σ is the Lebesgue measure on a convex, compact subset of $\mathbb{R}^n$ or
2. σ, μ, ν satisfy the assumptions of Caffarelli's regularity theorem (Theorem 7.1),

then there exists a $\delta > 0$ such that whenever $W_2(g_1 \star \mu, g_2 \star \nu) > \delta$ for all $g_1, g_2 \in \mathscr{G}$, we have that $F_\sigma(\mathscr{G} \star \mu)$ and $F_\sigma(\mathscr{G} \star \nu)$ are linearly separable. Moreover, δ is explicitly computable in both cases (see Remark 4.1) and $\delta = O(\varepsilon^{\frac{2}{15}})$ in Case 1 and $\delta = O(\varepsilon^{\frac{1}{2}})$ in Case 2.

Proof. We show a detailed proof in Section 7.4. □

REMARK 4.1. We note here that for both cases of Theorem 4.2, the sufficient minimal distance δ can be made explicit.

1. In this case, the Hölder bound by [22] can be used, see (7.13). With $\psi(\mu) = C\|f_\mu\|_\infty^{1/15}\varepsilon^{2/15} + \|f_\mu\|_\infty^{1/2}\varepsilon$, where f_μ is the density of μ with respect to the Lebesgue measure, the choice $\delta =$

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

$6 \max\{\psi(\mu), \psi(\nu)\}$ is sufficient. The constant C is the same constant appearing in the derivations by [22].

2. In this case, a Hölder bound following from [15] can be used, see Corollary 7.1. With

$$\overline{\psi}(\mu) := \left(\sqrt{\frac{4R}{K_\mu^\sigma}} + 2\right) \|f_\mu\|_\infty^{1/2} \varepsilon + \left(4R \|f_\mu\|_\infty^{1/2} \frac{W_2(\sigma,\mu) + R + \|\operatorname{Id}\|_\mu}{K_\mu^\sigma}\right)^{1/2} \varepsilon^{1/2},$$

the choice $\delta = 6 \max\{\overline{\psi}(\mu), \overline{\psi}(\nu)\}$ is sufficient. The constant K_σ^μ is defined in Definition 7.1.

Note that a minimal distance $\delta > 0$ is needed since we consider perturbations of 'ideal' functions (shifts and scalings). A version of ψ (respectively $\overline{\psi}$) also appears in characterizing the amount that LOT distance deviates from Wasserstein-2 distance (Proposition 4.1). A parallel of ψ (respectively $\overline{\psi}$) could be established for any approximately compatible transformations by proving a result similar to Lemma 7.1 for some compatible transformation other than shifts and scalings.

The ε appears in both ψ and $\overline{\psi}$ since functions in $\mathscr{G}$ are ε-close to compatible functions, while the $\varepsilon^{2/15}$, respectively, $\varepsilon^{1/2}$ come from the general Hölder bounds for LOT as proved in [22], respectively [15].

Remark 4.2. Theorem 4.2 is written in terms of $\mathscr{G}_{\lambda,R,\varepsilon}$ for a fixed base measure λ and $\mu, \nu \ll \lambda$. This was done to define a family of allowable perturbations that is extrinsic and independent of the target distributions μ, ν. One could alternatively define the intrinsic family of push-forwards $\mathscr{G}_{\mu+\nu,R,\varepsilon}$ and all proofs follow through similarly. The benefit of this perspective is it allows for the theory to apply to discrete distributions μ, ν because we could drop the need for absolute continuity. This would change the constants of the general Hölder bound to $\psi(\mu) = C\varepsilon^{2/15} + \varepsilon$.

Remark 4.3. We mention that the Hölder bounds for the LOT embedding used in Theorem 4.2 can be further improved, see for example the recent paper [14].

As a corollary to Theorem 4.2 with $\varepsilon = 0$ and $\delta = 0$, we establish simple conditions under which LOT creates linearly separable sets for distributions in $\mathscr{P}_2(\mathbb{R}^n)$. This effectively creates a parallel of Theorem 4.4 and Theorem 5.6 of [27] for the higher dimensional cases of LOT, and under the particular compatibility conditions required for higher dimensions. Theorem 4.3 states this for $\mathscr{E}$ (shifts and scalings), and Theorem 7.4 in the Appendix provides an equivalent form for subsets of arbitrary compatible transforms.

Corollary 4.3. Let $\sigma, \mu, \nu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$, and let $\mathscr{H} \subseteq \mathscr{E}$ and let $\mathscr{H}$ be convex. If $\mathscr{H} \star \mu$ is closed and $\mathscr{H} \star \nu$ is compact, and these two sets are disjoint, then $F_\sigma(\mathscr{H} \star \mu)$ and $F_\sigma(\mathscr{H} \star \nu)$ are linearly separable.

We also note the separability result on $\mathbb{R}$, which follows directly from the results established above. It is also proved in [27].

Corollary 4.4. Let $\sigma, \mu, \nu \in \mathscr{P}_2(\mathbb{R})$, $\sigma \ll \lambda$, and let $\mathscr{H}$ be a convex set of monotonically increasing functions $\mathbb{R} \to \mathbb{R}$. If $\mathscr{H} \star \mu$ is closed and $\mathscr{H} \star \nu$ is compact, and these two sets are disjoint, then $F_\sigma(\mathscr{H} \star \mu)$ and $F_\sigma(\mathscr{H} \star \nu)$ are linearly separable.

Remark 4.4. Note that Corollary 4.4 is also proved in [27]. In [27], $\mathbb{H}$ (equivalent to our $\mathscr{H}$) is defined as a convex subgroup of the monotonic functions (Definitions 5.5 and 5.6 (i)–(iii) of [27]). We are able to relax the assumption from subgroup to subset, however. Definition 5.5 of [27] also

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

assumes differentiability of functions in $\mathbb{H}$, which is needed because constructions are considered from the viewpoint of densities, which means that (2.3) should hold. Since our approach uses the more general framework of measures rather than densities, we can drop this assumption.

REMARK 4.5. We note that each Theorem in this section can be trivially extended to an arbitrary set $\mathscr{H}$ (or $\mathscr{G}$) that is not required to be convex, so long as their convex hulls conv($\mathscr{H}$) (or conv($\mathscr{G}$)) satisfy the needed assumptions of closedness, compactness and disjointness.

REMARK 4.6. We note that the assumption of compactness for $\mathscr{H} \star \mu$ is not limiting in practice, as this simply requires that the family of $\mu-$pushforwards not to be unbounded (e.g. one cannot expect an arbitrarily large shift/dilation of a distribution). This boundedness is a common assumption when working on a real-world, finite number of training points data set.

5. Example: linear separability of MNIST data set

We linearly separate two classes of digits from the MNIST data set [18] with LOT to verify the linear separability result (Theorem 4.2) numerically. We mention that we have an additional experiment on Wasserstein distance approximation in the appendix (Section 7.6).

We consider the classes of 1s and 2s from the MNIST data set. Since the MNIST digits are centered in the middle of the image, and the images have a similar size, we applied an additional (random) scale and shift to every image. Scalings were applied between 0.4 and 1.2 using MATLAB's 'imresize' function. These values have been chosen based on the heuristics that smaller scales make some digits unrecognizable and with larger scales some digits are larger than the image. The images are non-negative and were normalized to 1 after scaling, so they are probability measures supported on $\mathbb{R}^2$.

Within each class, the digits can be considered as shifts, scalings and perturbations of each other. Therefore, the aim of this section is to show the LOT embedding works well to separate 1s from 2s.

The data consisting of images of 1s and 2s are embedded in L^2 via the LOT embedding, where we choose as reference density σ an isotropic Gaussian, and every image μ is interpreted as a density on a grid $R \subset \mathbb{R}^2$. To approximate the continuous transport map, we project σ to R and compute the optimal coupling between σ and μ. This is then projected to a transport map by taking the center of $P_\sigma^\mu(x, y)$ for each x. This means that every image is assigned to the function $T_\sigma^\mu : \text{supp}(\sigma) \to R$. Since $\text{supp}(\sigma) \subset R$ is discrete, $T_\sigma^\mu(\text{supp}(\sigma))$ is a vector in $\mathbb{R}^{2n}$, where n is the number of grid points in $\text{supp}(\sigma)$. For each μ of the data set, we use this vector as input for the linear classification scheme (we use MATLAB's 'fitcdiscr' function). We note that while the discretization of the transport plan no longer satisfies the regularity assumptions of the reference distribution needed for some of the guarantees, this experimental setup demonstrates the robustness of the LOT embedding even beyond the regularity assumptions.

The experiment is conducted in the following way: we fix the number of testing data to 100 images from each class (i.e. in total, the testing data set consists of 200 images). Note that we only fix the number of testing data; the actual testing images are chosen randomly from the MNIST data set for each experiment. For the training data set, we randomly choose N images from each class, where $N = 40, 60, 80$ and 100. For each choice N, we run 20 experiments. In each experiment, after embedding the points using LOT, we perform linear SVM on the training data set and the classification error of the test data is computed. Then, the mean and standard deviation for every N is computed. The mean classification error is shown in Figure 1 (blue graph labeled 'LOT') as a function of N.

We compare the classification performance of LOT with regular L^2 distance between the images. Since we only use a small number of training data ($N = 40, 60, 80$ and 100 for each digit), and the size of an image is $28 \times 28 = 784$, the dimension of the feature space is much larger than the data

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

point dimension. Such a set-up leads to zero within-class variance in LDA. To prevent this, and in order to allow for a fair comparison, we first apply PCA to reduce the dimension of the images to the same dimension as is used in LOT. The feature space dimension used in LOT is the size of the support of σ, which consists of ≈ 70 grid points in these experiments. Thus, the dimension is 140. LDA is then applied to the PCA embeddings of the images. The resulting mean classification error is shown in Figure 1 (red graph labeled 'PCA') as a function of N. We chose PCA as a comparison to demonstrate the robustness of LOT to geometric transformations of the data, something that subspace methods such as PCA are not equipped for without preprocessing (if possible).

We also compare the classification result to two different Gaussian Mixture Model classification schemes. In the first approach, we train a GMM for each class (i.e. one model for MNIST digit 1, and one model for MNIST digit 2). We then compute the probability of a testing data point belonging to model 1 or model 2 and assign the class that results in higher probability. Note that we again need to apply dimension reduction, since the number of data points needs to be larger than the number of variables for GMM training. We used PCA for dimension reduction, and the dimension was either chosen as the dimension for the LOT embedding, or if the LOT dimension is too large for GMM training, it was chosen as the largest possible based on the training data. We also mention that we used a small regularizer (on the order of 10^{-6}) to guarantee convergence. To train the GMM model for each class, we used AIC to determine the optimal number of Gaussians to fit the data (between 1 and 8 Gaussians). The resulting mean classification error is shown in Figure 1 (purple graph labeled 'GMM image') as a function of N.

In the second approach, we first extract every 7×7 patch from the images and train a GMM for each class on the patches. Then for each patch of each test point, we perform a least squares regression to approximate the patch by a linear combination of the GMM means. After all patches are approximated, we reconstruct the image from the patches (when windows overlap we average over the patches). This is done for both models, and the test point is classified as the digit whose model yielded smaller L^2 reconstruction error. We used a small regularizer (on the order of 10^{-6}) to guarantee convergence and used the AIC to determine the optimal number of Gaussians (chosen from the set $\{5, 10, 25, 50\}$). The resulting mean classification error is shown in Figure 1 (cyan graph labeled 'GMM patch') as a function of N.

We also compare these classification results to training a convolutional neural network (CNN) [16] on small amounts of data. This is not necessarily a perfect comparison, as LOT and PCA are building unsupervised embeddings followed by a supervised classifier in that space, whereas a CNN is building an end-to-end supervised feature extraction and classification. In theory, this should benefit the CNN if the only method of validation is the overall classification error. However, as we will demonstrate, in the small data regime, the CNN's performance still does not compete with the LOT embedding and linear classification.

To show this, we construct two CNNs to be shown in Figure 1. The first (labeled 'Small CNN') is a network constructed with three convolutional layers, each with 2 3x3 filters, followed by two fully connected layers, all with ReLU activation units. In total, this CNN has 182 trainable parameters, which is of a similar size to the 140 parameters used in the LOT embedding. The second (labeled 'Large CNN') is a similar architecture, but with 8 3x3 filters, and 3650 trainable parameters. The CNNs are given the same training data sets as the LOT embeddings, and the testing error is also averaged over 20 experiments. We chose to compare to a CNN to demonstrate the benefit of LOT embeddings and linear classifiers for small data, even when compared with neural networks that are in the interpolation regime (such as the 'Large CNN'). This also demonstrates that while CNN layers are naturally designed to handle translations, the shifted and scaled MNIST data we have created are not automatically handled by a CNN.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

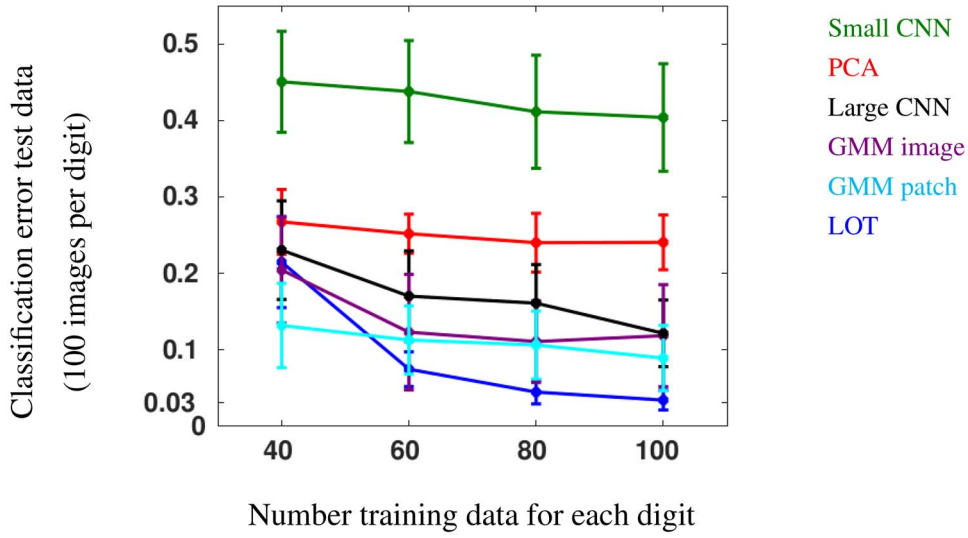

FIG. 1. Classification of MNIST digits 1s and 2s with small CNN (green), PCA with a linear classifier (red), large CNN (black), GMM on the image (purple), GMM on the patches (cyan) and LOT embedding with a linear classifier (blue). We fix the number of testing data (100 images per digit) and vary the number of training data, $N = 40, 60, 80$ and 100 (the actual training and testing sets are chosen randomly for each experiment). We train a linear classifier on the LOT embedding (blue) and the PCA embedding (red), a GMM on PCA reduced images (purple) and on patches (cyan), and two CNNs (black and green). The figure shows the mean and standard deviation of the classification error of the testing data over 20 experiments for each N.

It is clear from the figure that the mean error decreases as the number of training data increases for the LOT embedding, while the mean error stagnates for the PCA embedding. The GMM classification performs better than PCA, but the classification error does not drop as significantly as for LOT. Note that we start with a very small amount of training data (40 images from each class), and test on 100 images from each class. The resulting LOT mean error is only ≈ 0.2. When we train on the same amount as we test (100 images per class), the LOT mean error is already down to ≈ 0.03. Similarly, the LOT mean error significantly outperforms both the small and large CNNs. This is perhaps unsurprising as neural networks are known to require large corpuses of training data [20] but still serves to demonstrate the strength of embedding into a linearly separable space.

The LOT classification result is also visualized via LDA embedding plots in Figure 2 for two experiments. These plots again underline the fact that separation improves as the training data are increased. While training on 100 images per class (right plot of Figure 2) leads to almost perfect separation, training on 40 images per class (left plot of Figure 2) still performs very well considering the small size of the training set.

In addition to the fact that the LOT embedding is capable of producing good separation results on small training data, there is yet another benefit connected to the dimensionality of the problem. To run LDA (or any linear classifier), a matrix of data points versus features needs to be constructed. If we were to compare the original images, the feature space would have dimension equal to the number of grid points. In the LOT embedding, only the grid points in $\text{supp}(\sigma)$ need to be considered, rather than the whole grid, which drastically reduces the dimension of the feature space. In the experiments we ran with MNIST, the grid is of size 28×28, which leads to dimension $28^2 = 784$, while the support of σ is ≈ 70 grid points, hence the dimension is 140.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

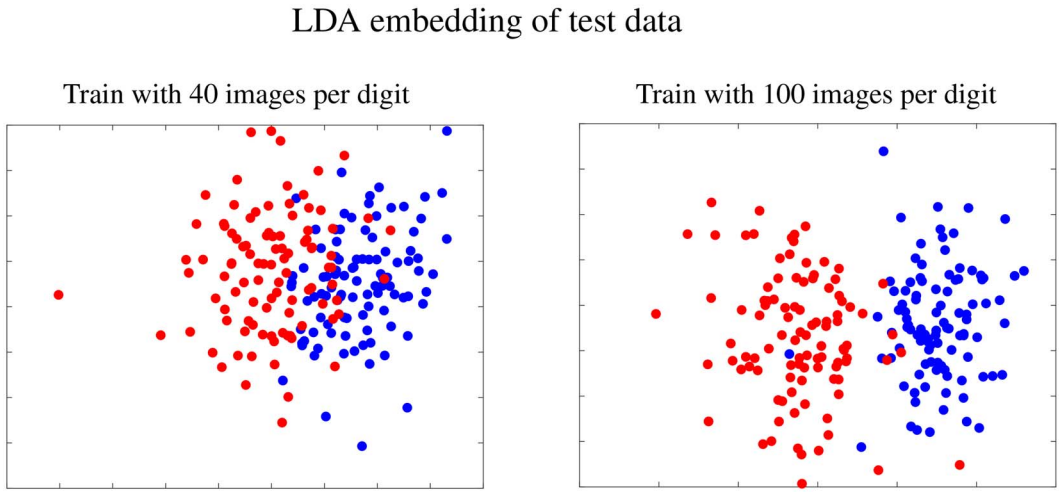

FIG. 2. LDA embedding plots for the MNIST classification of digits 1 and 2 using LOT. As Figure 1, these plots underline that the classification improves with the amount of training data. *Left:* We choose one of the experiments carried out for $N = 40$ training data for each digit. The testing data (100 images per digit) is embedded in $\mathbb{R}^2$ through the LDA coordinates. The mean error (Figure 1) is ≈ 0.2, which corresponds to ≈ 40 digits being misclassified. *Right:* We choose one of the experiments carried out for $N = 100$ training data for each digit. The testing data (100 images per digit) is embedded in $\mathbb{R}^2$ through the LDA coordinates. The mean error (Figure 1) is ≈ 0.03, which corresponds to ≈ 6 digits being misclassified.

This dimension reduction allows us to run LDA on small training data as we did in these experiments. If the feature dimension is very high, one also needs a lot of training data to prevent zero within-class variance, or one has to first apply PCA as we did for Figure 1.

6. Conclusion

In summary, LOT provides a useful framework for embedding certain families of distributions into a linearly separable Hilbert space. These families can consist of shifts, scalings and perturbations of a base distribution, and the results are strengthened when the base distribution satisfies the Caffarelli's regularity assumptions (support is convex and has a smooth boundary, densities are Hölder-continuous and the densities have a minimum height).

There are a number of directions of future work that are currently being considered. First, we are examining how the set of compatible push-forwards can be significantly increased if we make assumptions about the reference distribution σ and how it relates to the base distribution μ. Similarly, one can consider multiple reference sets. Second, we are examining how these results can be extended to other forms of optimal transport, including entropic regularization and graph transport. Third, we are considering LOT on point clouds sampled from the measure μ and how the guarantees scale with the size of the samples.

Data availability statement

There are no new data associated with this article.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

Acknowledgement

We thank the referees for their careful reading and their suggestions to improve the paper.

Funding

NSF DMS (grants 1819222 and 2012266 to AC, 2111322 to CM); Russell Sage Foundation (grant 2196 to AC); AMS-Simons Travel Grant (to CM).

References

1. Aldroubi, A., Li, S. & Rohde, G. K. (2021) Partitioning signal classes using transport transforms for data analysis and machine learning. *Sampl. Theory Signal Process. Data Anal.*, **19**, 6.
2. Aliprantis, C. D. & Border, K. C. (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide*. London: Springer, Berlin.
3. Ambrosio, L. & Gigli, N. (2013) *A User's Guide to Optimal Transportpp*. Berlin Heidelberg, Berlin, Heidelberg: Springer, pp. 1–155.
4. Arjovsky, M., Chintala, S. & Bottou, L. (2017) Wasserstein generative adversarial networks. *Proceedings of Machine Learning Research*, vol. **70**. (D. Precup & Y. W. Teh eds). Sydney, NSW, Australia: PMLR, pp. 214–223.
5. Berman, R. (2021) Convergence rates for discretized Monge-Ampére equations and quantitative stability of optimal transport. *Found Comput Math.*, **21**, 1099–1140.
6. Brenier, Y. (1991) Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, **44**, 375–417.
7. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. (2014) Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci.*, **111**, E2770–E2777.
8. Caffarelli, L. A. (1992a) Boundary regularity of maps with convex potentials. *Comm. Pure Appl. Math.*, **45**, 1141–1151.
9. Caffarelli, L. A. (1992b) The regularity of mappings with a convex potential. *J. Amer. Math. Soc.*, **5**, 99–104.
10. Caffarelli, L. A. (1996) Boundary regularity of maps with convex potentials–II. *Ann. Math. Second Series*, **144**, 453–496.
11. Cheng, X., Cloninger, A. & Coifman, R. R. (2020) Two-sample statistics based on anisotropic kernels. *Inf. Inference.*, **9**, 677–719.
12. Cloninger, A., Roy, B., Riley, C. & Krumholz, H. M. (2019) People mover's distance: Class level geometry using fast pairwise data adaptive transportation costs. *Appl. Comput. Harmon. Anal.*, **47**, 248–257.
13. Cuturi, M. (2013) Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., **28**, pp. 2292–2300.
14. Delalande, A. & Mérigot, Q. (2021) Quantitative stability of optimal transport maps under variations of the target measure. arXiv:2103.05934.
15. Gigli, N. (2011) On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proc. Edinburgh Math. Soc. (2)*, **54**, 401–409.
16. Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016) *Deep Learning*. Cambridge, MA, USA: The MIT Press.
17. Kolouri, S., Park, S. R. & Rohde, G. K. (2016) The radon cumulative distribution transform and its application to image classification. *IEEE Trans. Image Process.*, **25**, 920–934.
18. LeCun, Y. & Cortes, C. (1998) The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.
19. Leeb, W. & Coifman, R. (2016) Hoelder-Lipschitz norms and their duals on spaces with semigroups, with applications to earth mover's distance. *J. Fourier Anal. Appl.*, **22**, 910–953.
20. Marcus, G. (2018) Deep learning: a critical appraisal. arXiv preprint arXiv:1801.00631.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

21. McCann, R. J. (2001) Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.*, **11**, 589–608.
22. Mérigot, Q., Delalande, A. & Chazal, F. (2020) Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa & R. Calandra eds.), vol. **108** of *Proceedings of Machine Learning Research*. PMLR, pp. 3186–3196.
23. Mishne, G., Talmon, R., Meir, R., Schiller, J., Lavzin, M., Dubin, U. & Coifman, R. R. (2016) Hierarchical coupled-geometry analysis for neuronal structure and activity pattern discovery. *IEEE J. Sel. Top. Signal Process.*, **10**, 1238–1253.
24. Muandet, K., Fukumizu, K., Sriperumbudur, B. & Schölkopf, B. (2017) Now Foundations and Trends. *Kernel Mean Embedding of Distributions: A Review and Beyond*, **10**, 1–141.
25. Narici, L. & Beckenstein, E. (2010) *Topological Vector Spaces*. Boca Raton, FL, USA: CRC Press.
26. Newey, W. K. & West, K. D. (1987) Hypothesis testing with efficient method of moments estimation. *Internat. Econom. Rev.*, **28**, 777–787.
27. Park, S. R., Kolouri, S., Kundu, S. & Rohde, G. K. (2018) The cumulative distribution transform and linear pattern classification. *Appl. Comput. Harmon. Anal.*, **45**, 616–641.
28. Peyré, G. & Cuturi, M. (2019) *Computational Optimal Transport. Foundations and Trends in Machine Learning*, Delft, Netherlands: Now Publishers Inc., vol. **11**, pp. 355–607.
29. Rubner, Y., Tomasi, C. & Guibas, L. J. (2000) The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, **40**, 99–121.
30. Shirdhonkar, S. & Jacobs, D. W. (2008) Approximate earth mover's distance in linear time. *2008 IEEE Conference on Computer Vision and Pattern Recognition*.Anchorage, AK, USA: IEEE, pp. 1–8.
31. Solomon, J., Rustamov, R., Guibas, L. & Butscher, A. (2014) Wasserstein propagation for semi-supervised learning. *International Conference on Machine Learning*. **32**. (Xing, Eric P. and Jebara, Tony eds). Bejing, China: PMLR, pp. 306–314.
32. Villani, C. (2009) *Optimal Transport*. Berlin Heidelberg: Springer.
33. Wang, W., Slepčev, D., Basu, S., Ozolek, J. A. & Rohde, G. K. (2013) A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comput. Vis.*, **101**, 254–269.
34. Zhang, Y., Jin, R. & Zhou, Z.-H. (2010) Understanding bag-of-words model: a statistical framework. *Int. J. Mach. Learn. Cybern*, **1**, 43–52.
35. Zhao, J., Jaffe, A., Li, H., Lindenbaum, O., Sefik, E., Jackson, R., Cheng, X., Flavell, R. A. & Kluger, Y. (2021) Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proc. Natl. Acad. Sci.*, **118**, e2100293118.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

Appendix

The aim of this appendix is to establish the proofs of Section 4 (Geometry of LOT space). We actually derive a more general theory (true for all compatible transformations) in Section A.2 from which the results of Section 4 follow. The rest of the appendix contains auxiliary results (Sections A.1 and A.5) and additional experiments (Section A.6); see Table A1 for an overview.

A.1 *Regularity of the LOT embedding*

The main results of this paper are based on Hölder regularity-type properties of the LOT embedding, which we discuss in more detail in this section.

One of the main ingredients is a version of a theorem on the regularity of the optimal transport map proved by L. A. Caffarelli [8–10]. The formulation of the theorem is taken from [15].

THEOREM A.1 (Caffarelli's regularity theorem). Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$ with $\sigma, \mu \ll \lambda$. Assume that $\operatorname{supp}(\sigma), \operatorname{supp}(\mu)$ are C^2 and uniformly convex. Further assume that for some $\alpha \in (0,1)$, the densities f_σ, f_μ are $C^{0,\alpha}$ continuous on their supports and assume that they are bounded from above and below, i.e. there exist constants $c, C, \overline{c}, \overline{C} > 0$ such that

$$0 < c \leq \|f_\sigma\|_\infty \leq C,$$
$$0 < \overline{c} \leq \|f_\mu\|_\infty \leq \overline{C}.$$

Then, T_μ^σ is the gradient of a $C^{2,\alpha}$ function on $\operatorname{supp}(\mu)$.

DEFINITION A.1. We introduce the concept of k-strong convexity.

1. Let $f : X \to \mathbb{R}$ with $X \subseteq \mathbb{R}^n$ convex. f is called k-strongly convex if $g_k(x) = f(x) - \frac{1}{2}k\|x\|^2$ is convex.
2. For two measures $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$ with $\operatorname{supp}(\sigma)$ convex, we denote K_σ^μ to be the supremum over all k such that φ with $\nabla\varphi = T_\sigma^\mu$, is k-strongly convex on $\operatorname{supp}(\sigma)$.

In [15, Corollary 3.2], it is proved that if σ, μ satisfy the assumptions of Caffarelli's regularity theorem (Theorem A.1), then $K_\sigma^\mu > 0$. We further cite the following result from [15].

THEOREM A.2 ([15, Proposition 3.3]). Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$ and assume they satisfy the same assumptions as in Caffarelli's regularity theorem (Theorem A.1). Then for every S that pushes σ to

TABLE A1 *Overview of the appendix chapters*

Section	Title
Section A.1	Regularity of the LOT embedding
Section A.2	Set-up for linear separability results
Section A.3	Proofs of Section 3
Section A.4	Proofs of Section 4
Section A.5	A useful result in normed spaces
Section A.6	Experiment: Wasserstein distance approximation

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

μ, we have

$$\|S - T_\sigma^\mu\|_\sigma^2 \leq \frac{1}{K_\mu^\sigma}\left(\|S - \mathrm{Id}\,\|_\sigma^2 - W_2(\sigma,\mu)^2\right).$$

Note that in the formulation of this theorem in [15], $2/K_\mu^\sigma$ appears instead of $1/K_\mu^\sigma$ in the bound. From the proof presented in [15], it can be seen, however, that 2 can be replaced by 1.

We now prove a bound on the LOT embedding. The proof is based on Theorem A.2 and [15, Corollary 3.4].

THEOREM A.3. Let $\sigma, \nu_1, \nu_2 \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma, \nu_1, \nu_2 \ll \lambda$. Suppose that σ and ν_2 satisfy the assumptions of Caffarelli's regularity theorem (Theorem A.1). Then,

$$\|F_\sigma(\nu_1) - F_\sigma(\nu_2)\|_\sigma \leq \left(\frac{2}{K_{\nu_2}^{\sigma\,1/2}} + 1\right) W_2(\nu_1,\nu_2) + 2\left(\frac{W_2(\sigma,\nu_2)}{K_{\nu_2}^\sigma}\right)^{1/2} W_2(\nu_1,\nu_2)^{1/2}.$$

Proof. Let $S = T_{\nu_1}^{\nu_2}$. We aim at finding a bound on $\|T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma$.

The triangle inequality and change-of-variables formula imply

$$\begin{aligned}\|S \circ T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma &\geq \|T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma - \|S \circ T_\sigma^{\nu_1} - T_\sigma^{\nu_1}\|_\sigma = \|T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma - \|S - \mathrm{Id}\,\|_{\nu_1} \\ &= \|T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma - W_2(\nu_1,\nu_2).\end{aligned}$$

Thus, we get

$$\|T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma \leq \|S \circ T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma + W_2(\nu_1,\nu_2), \tag{A.1}$$

Theorem A.2 implies

$$\|S \circ T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma^2 \leq \frac{1}{K_{\nu_2}^\sigma}\left(\|S \circ T_\sigma^{\nu_1} - \mathrm{Id}\,\|_\sigma^2 - W_2(\sigma,\nu_2)^2\right). \tag{A.2}$$

Again by the triangle inequality and the change-of-variables formula, we have

$$\begin{aligned}\|S \circ T_\sigma^{\nu_1} - \mathrm{Id}\,\|_\sigma &\leq \|S \circ T_\sigma^{\nu_1} - T_\sigma^{\nu_1}\|_\sigma + \|T_\sigma^{\nu_1} - \mathrm{Id}\,\|_\sigma = W_2(\nu_1,\nu_2) + W_2(\sigma,\nu_1) \\ &\leq 2W_2(\nu_1,\nu_2) + W_2(\sigma,\nu_2).\end{aligned}$$

Combining this with (A.2), we obtain

$$\|S \circ T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma^2 \leq \frac{4}{K_{\nu_2}^\sigma}\left(W_2(\nu_1,\nu_2)^2 + W_2(\nu_1,\nu_2)W_2(\sigma,\nu_2)\right).$$

Taking the square root and using the fact that $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$, we obtain

$$\|S \circ T_\sigma^{\nu_1} - T_\sigma^{\nu_2}\|_\sigma \leq \frac{2}{K_{\nu_2}^{\sigma\,1/2}}\left(W_2(\nu_1,\nu_2) + (W_2(\nu_1,\nu_2)W_2(\sigma,\nu_2))^{1/2}\right).$$

Now (A.1) implies the result. □

Note that the 'constants' in Theorem A.3 depend on ν_2 (namely $K_{\nu_2}^\sigma$ and $W_2(\sigma,\nu_2)$). This can be avoided by considering $\nu_2 \in \mathscr{E} \star \mu$ for a fixed $\mu \in \mathscr{P}_2(\mathbb{R}^n)$, where $\mathscr{E}$ denotes the set of shifts and scalings. As a preparation for this result, we need the following lemma:

LEMMA A.1. Let $f : X \to \mathbb{R}$ be differentiable with $X \subseteq \mathbb{R}^n$ convex. Then, we have the following:

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

1. f is k-strongly convex on X if and only if $f \circ S_a$ is k-strongly convex on $S_a^{-1}(X)$.
2. f is k-strongly convex on X if and only if $R_c^{-1} \circ f \circ R_c$ is (kc)-strongly convex on $R_c^{-1}(X)$.

Proof. We first note that X is convex if and only if $h^{-1}(X)$ is convex for $h = S_a$ or $h = R_c$. Furthermore, f is k-strongly convex if and only if

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq k\|x - y\|^2, \quad x, y \in X. \tag{A.3}$$

For $\bar{x}, \bar{y} \in S_a^{-1}(X)$, eq. (A.3) implies that $f \circ S_a$ is k-strongly convex if and only if

$$(\nabla f \circ S_a(\bar{x}) - \nabla f \circ S_a(\bar{y}))^T (\bar{x} - \bar{y}) \geq k\|\bar{x} - \bar{y}\|^2,$$

which is the same as

$$(\nabla f(S_a(\bar{x})) - \nabla f(S_a(\bar{y})))^T (S_a(\bar{x}) - S_a(\bar{y})) \geq k\|S_a(\bar{x}) - S_a(\bar{y})\|^2.$$

As this is only a transformation $x = S_a(\bar{x})$ and $y = S_a(\bar{y})$ compared with eq. (A.3), k-strong convexity of f and $f \circ S_a$ are equivalent.

For $\bar{x}, \bar{y} \in R_c^{-1}(X)$, eq. (A.3) implies that $R_c^{-1} \circ f \circ R_c$ is (kc)-strongly convex if and only if

$$(\nabla (R_c^{-1} \circ f \circ R_c)(\bar{x}) - \nabla (R_c^{-1} \circ f \circ R_c)(\bar{y}))^T (\bar{x} - \bar{y}) \geq kc\, \|\bar{x} - \bar{y}\|^2,$$

which is the same as

$$c^{-1} (\nabla f(R_c(\bar{x})) - \nabla f(R_c(\bar{y})))^T (R_c(\bar{x}) - R_c(\bar{y})) \geq kc\, c^{-2} \|R_c(\bar{x}) - R_c(\bar{y})\|^2,$$

resulting in

$$(\nabla f(R_c(\bar{x})) - \nabla f(R_c(\bar{y})))^T (R_c(\bar{x}) - R_c(\bar{y})) \geq k\|R_c(\bar{x}) - R_c(\bar{y})\|^2.$$

As this is only a transformation $x = R_c(\bar{x})$ and $y = R_c(\bar{y})$ compared with eq. (A.3), k-strong convexity of f and (kc)-strong convexity of $R_c^{-1} \circ f \circ R_c$ are equivalent. □

COROLLARY A.1. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma, \mu \ll \lambda$. Further assume that σ and μ satisfy the assumptions of Caffarelli's regularity theorem (Theorem A.1). Let $R > 0$ and consider $h \in \mathscr{E}_{\mu,R}$ (bounded shifts/scalings, see eq. (4.2)) as well as $g \in L^2(\mathbb{R}^n, \mu)$. Then, we have

$$\begin{aligned}\|F_\sigma(g_\sharp \mu) - F_\sigma(h_\sharp \mu)\|_\sigma \leq & \left(\sqrt{\frac{4R}{K_\mu^\sigma}} + 1\right) W_2(g_\sharp \mu, h_\sharp \mu) \\ & + \left(4R\, \frac{W_2(\sigma, \mu) + R + \|\operatorname{Id}\|_\mu}{K_\mu^\sigma}\right)^{1/2} W_2(g_\sharp \mu, h_\sharp \mu)^{1/2}.\end{aligned}$$

Note that we now have a bound with constants that do not depend on h or g. They only depend on the fixed measures σ, μ and on the radius R.

Proof. Let $\nu_1 = g_\sharp \mu$ and $\nu_2 = h_\sharp \mu$. First note that since μ and σ satisfy the assumptions of Caffarelli's regularity theorem, also ν_2 and σ satisfy them. Therefore, we can apply Theorem A.3.

We now bound $W_2(\sigma, \nu_2)$ and $K_{\nu_2}^\sigma$ from Theorem A.3 by constants that only depend on σ, μ and R. Such bounds then imply the result.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

The triangle inequality, Lemma 3.2, and the assumption $\|h\|_\mu < R$ imply

$$\begin{aligned} W_2(\sigma,\nu_2) &\leq W_2(\sigma,\mu) + W_2(\mu,h_\sharp\mu) = W_2(\sigma,\mu) + \|T_\mu^{h_\sharp\mu} - \mathrm{Id}\,\|_\mu \\ &= W_2(\sigma,\mu) + \|h - \mathrm{Id}\,\|_\mu \\ &< W_2(\sigma,\mu) + R + \|\,\mathrm{Id}\,\|_\mu. \end{aligned}$$

We now show that $K_{\nu_2}^\sigma$ can be lower bounded by something that only depends on σ, μ and R but does not depend on h. First consider $h = S_a$. Note that $T_{\nu_2}^\sigma = T_\mu^\sigma \circ S_a^{-1}$ and $T_\mu^\sigma = \nabla\psi$ implies $T_\mu^\sigma \circ S_a^{-1} = \nabla\psi \circ S_a^{-1}$. Also, $\psi \circ S_a^{-1}$ is convex on $S_a(\mathrm{supp}(\mu))$. This implies that $\varphi = \psi \circ S_a^{-1}$, where $\nabla\varphi = T_{\nu_2}^\sigma$.

Lemma A.1 implies that ψ is k-strongly convex if and only if $\varphi = \psi \circ S_a^{-1}$ is k-strongly convex. Therefore, the modulus of uniform convexity of $\psi \circ S_a^{-1}$ equals the modulus of uniform convexity of ψ. Thus, $K_{\nu_2}^\sigma = K_\mu^\sigma$, which is independent of S_a.

Now consider $h = R_c$. Again, we have $T_{\nu_2}^\sigma = T_\mu^\sigma \circ R_c^{-1}$ and $T_\mu^\sigma = \nabla\psi$ implies $T_\mu^\sigma \circ R_c^{-1} = \nabla R_c \circ \psi \circ R_c^{-1}$. Also, $R_c \circ \psi \circ R_c^{-1}$ is convex on $R_c(\mathrm{supp}(\mu))$. This implies that $\varphi = R_c \circ \psi \circ R_c^{-1}$, where $\nabla\varphi = T_{\nu_2}^\sigma$.

Lemma A.1 implies that ψ is k-strongly convex if and only if $\varphi = R_c \circ \psi \circ R_c^{-1}$ is kc^{-1}-strongly convex. Therefore, $K_{\nu_2}^\sigma = K_\mu^\sigma c^{-1}$. Since by assumption $|c| = \|R_c\|_\mu < R$, we have

$$\frac{1}{K_{\nu_2}^\sigma} = \frac{1}{K_\mu^\sigma}|c| < \frac{R}{K_\mu^\sigma},$$

which gives a bound independent of R_c. □

We now combine Corollary A.1 with the Lipschitz continuity of the pushforward map $g \mapsto g_\sharp\sigma$ to obtain a Hölder regularity-type result for LOT. We first cite the result on the Lipschitz continuity of the pushforward map, which can be found in e.g. [3, Equation (2.1)]

$$W_2(g_\sharp\mu, h_\sharp\mu) \leq \|g - h\|_\mu. \tag{A.4}$$

Corollary A.2. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma, \mu \ll \lambda$. Further assume that σ and μ satisfy the assumptions of Caffarelli's regularity theorem (Theorem A.1). Let $R > 0$, $h \in \mathscr{E}_{\mu,R}$ (see eq. (4.2)) and $g \in L^2(\mathbb{R}^n, \mu)$. Then, we have

$$\|F_\sigma(g_\sharp\mu) - F_\sigma(h_\sharp\mu)\|_\sigma \leq \left(\sqrt{\frac{4R}{K_\mu^\sigma}} + 1\right)\|g - h\|_\mu + \sqrt{4R\,\frac{W_2(\sigma,\mu) + R + \|\,\mathrm{Id}\,\|_\mu}{K_\mu^\sigma}}\,\|g - h\|_\mu^{1/2}.$$

Remark A.1. In [15, Corollary 3.4], it is proved that for fixed σ and a Lipschitz continuous curve μ_t of absolutely continuous measures, $t \in [0,1]$, 1/2-Hölder regularity of $t \mapsto F_\sigma(\mu_t)$ can be achieved. Indeed, it is proved that

$$\|F_\sigma(\mu_t) - F_\sigma(\mu_0)\|_\sigma \leq C\sqrt{t}.$$

Corollary A.2 can be considered a generalization of this result. We prove that the map $h \mapsto F_\sigma(h_\sharp\mu)$ can achieve Hölder-type regularity between an element of $\mathscr{E}$ (comparable to μ_0) and an element of $L^2(\mathbb{R}^n, \mu)$ (comparable to μ_t). Note that like μ_t, the 'curve' $h \mapsto h_\sharp\mu$ is Lipschitz continuous (eq. (A.4)). The restriction to bounded shifts and scalings (via $R > 0$) relates to the fact that $[0,1]$ is bounded. We also

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

mention that the 'linear term' $\|g-h\|_\mu$, if small enough, can be bounded by $\|g-h\|_\mu^{1/2}$, relating our result more closely to [15]. Indeed, in [15], a linear term in t is also present but can always be bounded by $\sqrt{t}$ since $t \in [0,1]$.

A.2 *Set-up for linear separability results*

In this section, we build up the theory needed for the results on linear separability presented in Section 4. The proofs for these results can then be derived easily from results of this section, see Section A.4.

Throughout this section, let $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \sigma)$. Then, $\mathscr{H}$ acts on $\mathscr{P}_2(\mathbb{R}^n)$ by push-forward

$$h \star \mu = h_\sharp \mu, \qquad h \in \mathscr{H}, \mu \in \mathscr{P}_2(\mathbb{R}^n).$$

This is a group action if $\mathscr{H}$ is a subgroup of $L^2(\mathbb{R}^n, \sigma)$.

Fix $\mu \in \mathscr{P}_2(\mathbb{R}^n)$. Using the notation from Corollary 4.1, we denote by

$$\mathscr{H} \star \mu = \{h \star \mu : h \in \mathscr{H}\}$$

the orbit of μ with respect to the action of $\mathscr{H}$.

Note that $\mathscr{H}$ also acts on $L^2(\mathbb{R}^n, \sigma)$ by composition, i.e. $h \star f = h \circ f$ for $f \in L^2(\mathbb{R}^n, \sigma)$ and $h \in \mathscr{H}$. We also denote this action by $\star$.

We now derive some properties of this action in connection with the LOT embedding F_σ.

Definition A.2. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$ and let $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \sigma)$. We say that F_σ is *compatible with μ-orbits with respect to the action of $\mathscr{H}$* if

$$F_\sigma(h \star \mu) = h \star F_\sigma(\mu), \qquad h \in \mathscr{H}. \tag{A.5}$$

Remark A.2. Note that Equation (A.5) is exactly eq. (3.2). We just introduced a new notation via $\star$.

As is shown in Lemma 3.2, Condition (A.5) is satisfied by shifts and scalings in arbitrary dimension and by all monotonically increasing functions in dimension $n = 1$.

A version of the following lemma is also proved in [1].

Lemma A.2. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n), \sigma \ll \lambda$ and let $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \sigma)$ be convex. If F_σ is compatible with μ-orbits with respect to the action of $\mathscr{H}$ (Definition A.2), then $F_\sigma(\mathscr{H} \star \mu)$ is convex.

Proof. We prove that for $f \in L^2(\mathbb{R}^n, \sigma)$, convexity of $\mathscr{H}$ implies convexity of $\mathscr{H} \star f$. This together with condition (A.5) then implies convexity of $F_\sigma(\mathscr{H} \star \mu)$.

Let $c \in [0,1]$ and let $h_1, h_2 \in \mathscr{H}$. Then,

$$(1-c)(h_1 \circ f) + c\,(h_2 \circ f) = ((1-c)h_1 + c\,h_2) \circ f \in \mathscr{H} \star f. \qquad \square$$

Theorem A.4. Let $\sigma, \mu, \nu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$ and let $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \sigma)$ be convex. Further assume that F_σ is compatible with both μ- and ν-orbits with respect to the action of $\mathscr{H}$ (Definition A.2). If $\mathscr{H} \star \mu$ is closed and $\mathscr{H} \star \nu$ is compact, and these two sets are disjoint, then $F_\sigma(\mathscr{H} \star \mu)$ and $F_\sigma(\mathscr{H} \star \nu)$ are linearly separable.

Proof. Since F_σ is continuous, $F_\sigma(\mathscr{H} \star \mu)$ is compact and $F_\sigma(\mathscr{H} \star \nu)$ is closed. Since F_σ is injective (Lemma 3.1), they are also disjoint. Lemma A.2 implies that both images are convex. Therefore, the Hahn–Banach Theorem implies separability. $\square$

Definition A.2 is a strong condition which is satisfied for shifts and scalings. In the following, we show a linear separability result which relaxes this condition. Indeed, we show that Theorem A.4 is also

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

true if we extend $\mathscr{H}$ by functions which are ε-close to shifts and scalings in $L^2(\mathbb{R}^n, \sigma)$. In analogy to Definition A.2, we define compatibility of F_σ with respect to μ-orbits up to an error ε.

DEFINITION A.3. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$, let $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \sigma)$ and let $\varepsilon > 0$. We say that F_σ is *ε-compatible with μ-orbits with respect to the action of $\mathscr{H}$* if

$$\|F_\sigma(h \star \mu) - h \star F_\sigma(\mu)\|_\sigma < \varepsilon \qquad h \in \mathscr{H}.$$

There is also an analog to Lemma A.2:

LEMMA A.3. Let $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n), \sigma \ll \lambda$, let $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \sigma)$ be convex and let $\varepsilon > 0$. If F_σ is ε-compatible with μ-orbits with respect to the action of $\mathscr{H}$ (Definition A.3), then $F_\sigma(\mathscr{H} \star \mu)$ is 2ε-convex (Definition A.4).

Proof. Let $h_1, h_2 \in \mathscr{H}$ and $c \in [0, 1]$. Define $h = (1-c)h_1 + ch_2 \in \mathscr{H}$. We aim at proving that

$$\|(1-c)F_\sigma(h_1 \star \mu) + cF_\sigma(h_2 \star \mu) - F_\sigma(h \star \mu)\|_\sigma < 2\varepsilon.$$

To this end, we apply Definition A.3:

$$\begin{aligned}
&\|(1-c)F_\sigma(h_1 \star \mu) + cF_\sigma(h_2 \star \mu) - F_\sigma(h \star \mu)\|_\sigma \\
&\quad \leq (1-c)\|F_\sigma(h_1 \star \mu) - h_1 \star F_\sigma(\mu)\|_\sigma + c\|F_\sigma(h_2 \star \mu) - h_2 \star F_\sigma(\mu)\|_\sigma \\
&\qquad + \|h \star F_\sigma(\mu) - F_\sigma(h \star \mu)\|_\sigma \\
&\quad < (1-c)\varepsilon + c\varepsilon + \varepsilon = 2\varepsilon.
\end{aligned}$$

□

This lemma allows us to establish the most general form of the linear separability theorem, which simply requires the additional assumption that the two families generated by action $\mathscr{H}$, $\mathscr{H} \star \mu$ and $\mathscr{H} \star \nu$, have a minimal distance greater than 6ε.

THEOREM A.5. Let $\sigma, \mu, \nu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma \ll \lambda$, let $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \sigma)$ be convex and let $\varepsilon > 0$. Further assume that F_σ is ε-compatible with both μ- and ν-orbits with respect to the action of $\mathscr{H}$ (Definition A.3). If $\mathscr{H} \star \mu$ and $\mathscr{H} \star \nu$ are compact, and $W_2(h_1 \star \mu, h_2 \star \nu) > 6\varepsilon$ for all $h_1, h_2 \in \mathscr{H}$, then $F_\sigma(\mathscr{H} \star \mu)$ and $F_\sigma(\mathscr{H} \star \nu)$ are linearly separable.

Proof. Since F_σ is continuous, both $A = F_\sigma(\mathscr{H} \star \mu)$ and $B = F_\sigma(\mathscr{H} \star \nu)$ are compact. Now consider the closed convex hull of these sets, i.e. consider $\overline{\text{conv}(A)}$ and $\overline{\text{conv}(B)}$. The closed convex hull of compact sets is compact again in a completely metrizable locally convex space [2, Theorem 5.35]. Thus, in order to apply the Hahn–Banach theorem to $\overline{\text{conv}(A)}$ and $\overline{\text{conv}(B)}$, we only need to show that these sets are disjoint.

Lemma A.4 implies

$$6\varepsilon < W_2(h_1 \star \mu, h_2 \star \nu) \leq \|F_\sigma(h_1 \star \mu) - F_\sigma(h_2 \star \nu)\|_\sigma,$$

for $h_1, h_2 \in \mathscr{H}$. Therefore, $d(A, B) > 6\varepsilon$, where d denotes the distance between sets.

Since F_σ is ε-compatible with respect to both μ- and ν-orbits, Lemma A.3 implies that both A and B are 2ε-convex (Definition A.4). This means that $d(\text{conv}(A), A) < 2\varepsilon$ and $d(\text{conv}(B), B) < 2\varepsilon$.

Lemma A.6 now implies that $d(\text{conv}(A), \text{conv}(B)) > \varepsilon$. Therefore, the closure of these sets has positive distance, $d(\overline{\text{conv}(A)}, \overline{\text{conv}(B)}) > 0$, which implies that $\overline{\text{conv}(A)} \cap \overline{\text{conv}(B)} = \emptyset$.

REMARK A.3. The essential part of Theorem A.5 is about proving that convexity (or almost convexity) is preserved under LOT. We impose this property by assuming the compatibility condition to hold. Since

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

this condition is quite restrictive (up to now: shifts and scalings), one may consider other ways to infer convexity in the embedding space. A possibility would be to try relate convexity in the embedding space to geodesic convexity in $\mathscr{P}_2(\mathbb{R}^n)$ and Wasserstein barycenters. This is strongly connected to the question of how well LOT barycenters approximate Wasserstein barycenters (a question that is still open). This is part of future research.

A.3 *Proofs of Section 3*

Proof of Lemma 3.1. To prove part 1 of the lemma, we show continuity and injectivity of F_σ.

The stability of transport maps as described in [32, Corollary 5.23] implies that F_σ is continuous.

If $F_\sigma(\nu_1) = F_\sigma(\nu_2)$, then $T_\sigma^{\nu_1} = T_\sigma^{\nu_2}$. In particular, this implies

$$\nu_1 = T_\sigma^{\nu_1}{}_\sharp \sigma = T_\sigma^{\nu_2}{}_\sharp \sigma = \nu_2.$$

This implies injectivity of F_σ.

To prove part 2 of the lemma, let $c \in [0,1]$ and let $\nu_1, \nu_2 \in \mathscr{P}_2(\mathbb{R}^n)$. We define

$$T(x) := (1-c)\, F_\sigma(\nu_1)(x) + c\, F_\sigma(\nu_2)(x), \quad x \in \mathbb{R}^n.$$

We need to show that there exists $\nu_3 \in \mathscr{P}_2(\mathbb{R}^n)$ such that $T = F_\sigma(\nu_3)$. To this end, we define $\nu_3 := T_\sharp \sigma$. By definition, T pushes σ to ν_3. We now show that T can be written as the gradient of a convex function.

By Theorem 2.1, there exist convex functions φ_1, φ_2 such that $T_\sigma^{\nu_1}$ and $T_\sigma^{\nu_2}$ can be written uniquely as $T_\sigma^{\nu_j}(x) = \nabla \varphi_j(x), j = 1, 2, x \in \mathbb{R}^n$. This implies that $T(x) = \nabla \varphi_3(x)$, with the convex function

$$\varphi_3(x) = (1-c)\, \varphi_1(x) + c\, \varphi_2(x), \quad x \in \mathbb{R}^n.$$

Theorem 2.1 thus implies that $T = T_\sigma^{\nu_3}$, which proves $T = F_\sigma(\nu_3)$. □

Proof of Lemma 3.2. On $\mathbb{R}$ recall that

$$T_\sigma^\nu = G_\nu^{-1} \circ G_\sigma, \tag{A.6}$$

where G_σ denotes the cdf of σ defined by $G_\sigma(x) = \sigma((-\infty, x])$. Now if h is monotonically increasing, we have $G_{h_\sharp \mu} = G_\mu \circ h^{-1}$, which implies compatibility.

Let $n \geq 1$ and $h \in \mathscr{E}$. We first consider the case $h = S_a$ for some $a \in \mathbb{R}^n$. By Theorem 2.1, both T_σ^ν and $T_\sigma^{S_a{}_\sharp \nu}$ exist. We now prove $T_\sigma^{S_a{}_\sharp \nu} = S_a \circ T_\sigma^\nu$, which shows the result for $h = S_a$.

Again, by Theorem 2.1, there exists a unique convex function φ such that $T_\sigma^\nu = \nabla \varphi$. Then,

$$\left(S_a \circ T_\sigma^\nu\right)(x) = \nabla \varphi(x) + a = \nabla\left(\varphi(x) + < a, x >\right) = \nabla \psi(x),$$

where ψ is also convex.

Due to the general property

$$(\widetilde{T} \circ T)_\sharp \sigma = \widetilde{T}_\sharp (T_\sharp \sigma) \tag{A.7}$$

for maps $T, \widetilde{T}$, we have that $S_a \circ T_\sigma^\nu$ pushes σ to $S_{a\sharp}\nu$. Therefore, Theorem 2.1 implies that $S_a \circ T_\sigma^\nu = T_\sigma^{S_a{}_\sharp \nu}$.

We now consider the case $h = R_c$ for some $c \in \mathbb{R}$. By Theorem 2.1, both T_σ^ν and $T_\sigma^{R_c{}_\sharp \nu}$ exist. We now prove that $T_\sigma^{R_c{}_\sharp \nu} = R_c \circ T_\sigma^\nu$, which implies the result for $h = R_c$.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

Again, by Theorem 2.1, there exists a unique convex function φ such that $T_\sigma^\nu = \nabla\varphi$. Then,

$$\left(R_c \circ T_\sigma^\nu\right)(x) = c\nabla\varphi(x) = \nabla c\varphi(x) = \nabla\psi(x),$$

where ψ is also convex. Furthermore, by eq. (A.7), $R_c \circ T_\sigma^\nu$ pushes σ to $R_{c\sharp}\nu$. Therefore, Theorem 2.1 implies $T_\sigma^{R_{c\sharp}\nu} = R_c \circ T_\sigma^\nu$. □

A.4 *Proofs of Section 4*

We first establish an approximation result:

LEMMA A.4. Let $\sigma, \mu, \nu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma, \mu \ll \lambda$, then we have

$$W_2(\mu,\nu) \le \|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma \le W_2(\mu,\nu) + \|T_\mu^\nu - T_\sigma^\nu \circ T_\mu^\sigma\|_\mu.$$

We also have an upper bound by the triangle inequality

$$\|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma \le W_2(\mu,\sigma) + W_2(\sigma,\nu).$$

Proof. By the change-of-variables formula, we have

$$\|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma = \|T_\sigma^\mu - T_\sigma^\nu\|_\sigma = \|\operatorname{Id} - T_\sigma^\nu \circ T_\mu^\sigma\|_\mu. \tag{A.8}$$

Since $T_\sigma^\nu \circ T_\mu^\sigma$ pushes μ to ν, $W_2(\mu,\nu) \le \|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma$ follows.

For the first upper bound on $\|F_\sigma(\mu) - F_\sigma(\nu)\|_\sigma$ note the following

$$\|\operatorname{Id} - T_\sigma^\nu \circ T_\mu^\sigma\|_\mu \le \|\operatorname{Id} - T_\mu^\nu\|_\mu + \|T_\mu^\nu - T_\sigma^\nu \circ T_\mu^\sigma\|_\mu \le W_2(\mu,\nu) + \|T_\mu^\nu - T_\sigma^\nu \circ T_\mu^\sigma\|_\mu.$$

The second upper bound by $W_2(\mu,\sigma) + W_2(\sigma,\nu)$ follows from the triangle inequality. □

Lemma A.4 shows that the error occurring in the LOT approximation of the Wasserstein distance is determined by the L^2-error between the map $T_\sigma^\mu \circ T_\mu^\sigma$ and the correct transport map T_μ^ν. This means that the LOT embedding replaces the transport T_μ^ν by $T_\sigma^\nu \circ T_\mu^\sigma$ and computes the Wasserstein distance from this map.

Lemma A.4 shows that in case the relation

$$T_\mu^\nu = T_\sigma^\nu \circ T_\mu^\sigma \tag{A.9}$$

is satisfied, the LOT embedding is an isometry. Also, if eq. (A.9) is satisfied up to an error $\varepsilon > 0$, then ε is also the maximal error between the LOT embedding and the correct Wasserstein distance.

LEMMA A.5. Fix $\sigma, \mu \in \mathscr{P}_2(\mathbb{R}^n)$, $\sigma, \mu \ll \lambda$. If F_σ is compatible with μ-pushforwards of a set of functions $\mathscr{H} \subseteq L^2(\mathbb{R}^n, \mu)$ (see eq. 3.2) then for $h_1, h_2 \in \mathscr{H}$, we have

$$T_{h_{1\sharp}\mu}^{h_{2\sharp}\mu} = T_\sigma^{h_{2\sharp}\mu} \circ T_{h_{1\sharp}\mu}^\sigma.$$

Proof. Denote by $\nu_1 = h_{1\sharp}\mu, \nu_2 = h_{2\sharp}\mu$. Compatibility (eq. (3.2)) implies

$$T_\sigma^{\nu_j} = h_j \circ T_\sigma^\mu, \quad j = 1,2. \tag{A.10}$$

This implies

$$T_\sigma^{\nu_2} \circ T_{\nu_1}^\sigma = h_2 \circ T_\sigma^\mu \circ (h_1 \circ T_\sigma^\mu)^{-1} = h_2 \circ h_1^{-1}.$$

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

Again by eq. (3.2), we obtain

$$T_{\nu_1}^{\nu_2} = h_2 \circ T_{\nu_1}^{\mu} = h_2 \circ (T_{\mu}^{\nu_1})^{-1} = h_2 \circ (h_1 \circ T_{\mu}^{\mu})^{-1} = h_2 \circ h_1^{-1}.$$

□

Proof of Proposition 4.1. Since $g_1, g_2 \in \mathscr{G}_{\mu,R,\varepsilon}$, there exist $h_1, h_2 \in \mathscr{E}_{\mu,R}$ such that $\|g_1 - h_1\|_\mu < \varepsilon$ and $\|g_2 - h_2\|_\mu < \varepsilon$. The triangle inequality implies

$$\begin{aligned}\|F_\sigma(g_{1\sharp}\mu) - F_\sigma(g_{2\sharp}\mu)\|_\sigma &\le \|F_\sigma(g_{1\sharp}\mu) - F_\sigma(h_{1\sharp}\mu)\|_\sigma \\ &\quad + \|F_\sigma(h_{1\sharp}\mu) - F_\sigma(h_{2\sharp}\mu)\|_\sigma + \|F_\sigma(h_{2\sharp}\mu) - F_\sigma(g_{2\sharp}\mu)\|_\sigma. \end{aligned} \tag{A.11}$$

Lemma A.5, eq. (A.4) and the triangle inequality imply

$$\begin{aligned}\|F_\sigma(h_{1\sharp}\mu) - F_\sigma(h_{2\sharp}\mu)\|_\sigma &= W_2(h_{1\sharp}\mu, h_{2\sharp}\mu) \\ &\le W_2(h_{1\sharp}\mu, g_{1\sharp}\mu) + W_2(g_{1\sharp}\mu, g_{2\sharp}\mu) + W_2(g_{2\sharp}\mu, h_{2\sharp}\mu) \\ &\le \|h_1 - g_1\|_\mu + W_2(g_{1\sharp}\mu, g_{2\sharp}\mu) + \|g_2 - h_2\|_\mu \\ &\le 2\varepsilon + W_2(g_{1\sharp}\mu, g_{2\sharp}\mu). \end{aligned} \tag{A.12}$$

Now, we distinguish the two cases of the theorem

1. For this part, we use the following Hölder-$\frac{2}{15}$ regularity result by [22]:

$$\|F_\sigma(\nu_1) - F_\sigma(\nu_2)\|_\sigma \le C\, W_2(\nu_1, \nu_2)^{2/15}, \tag{A.13}$$

for $\nu_1, \nu_2 \in \mathscr{P}_2(\mathbb{R}^n)$. For $i = 1, 2$, we get

$$\|F_\sigma(g_{i\sharp}\mu) - F_\sigma(h_{i\sharp}\mu)\|_\sigma \le C\, W_2(g_{i\sharp}\mu, h_{i\sharp}\mu)^{2/15} \le C\, \|g_i - h_i\|_\mu^{2/15} < C\, \varepsilon^{2/15}.$$

This, together with (A.11) and (A.12), gives the overall bound

$$0 \le \|F_\sigma(g_{1\sharp}\mu) - F_\sigma(g_{2\sharp}\mu)\|_\sigma - W_2(g_{1\sharp}\mu, g_{2\sharp}\mu) \le 2C\, \varepsilon^{2/15} + 2\varepsilon.$$

2. With regularity assumptions on σ, μ, Corollary A.2 implies that there exist constants $C_{\sigma,\mu,R}, \overline{C}_{\sigma,\mu,R}$ such that

$$\begin{aligned}\|F_\sigma(g_{i\sharp}\mu) - F_\sigma(h_{i\sharp}\mu)\|_\sigma &\le C_{\sigma,\mu,R}\|g_i - h_i\|_\mu + \overline{C}_{\sigma,\mu,R}\|g_i - h_i\|_\mu^{1/2} \\ &\le C_{\sigma,\mu,R}\, \varepsilon + \overline{C}_{\sigma,\mu,R}\, \varepsilon^{1/2}, \end{aligned} \tag{A.14}$$

for $i = 1, 2$. Note that the same constants can be used for $i = 1$ and $i = 2$ since R bounds both h_1 and h_2. This, together with (A.11) and (A.12), gives the overall bound

$$0 \le \|F_\sigma(g_{1\sharp}\mu) - F_\sigma(g_{2\sharp}\mu)\|_\sigma - W_2(g_{1\sharp}\mu, g_{2\sharp}\mu) \le 2(C_{\sigma,\mu,R} + 1)\, \varepsilon + 2\overline{C}_{\sigma,\mu,R}\, \varepsilon^{1/2},$$

which concludes the proof. □

Proof of Corollaries 4.3 and 4.4. By Remark A.2, the compatibility condition (A.5) is satisfied. Thus, we can apply Theorem A.4. □

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

Proof of Theorem 4.2. We show that F_σ is δ-compatible with both μ- and ν-orbits with respect to the action of $\mathscr{G}$. Then, the result follows from Theorem A.5. We note that the value of δ will be as in Remark 4.1.

Let $g \in \mathscr{G}$ and $h \in \mathscr{E}_{\lambda,R}$ such that $\|g - h\| \leq \varepsilon$. Since $h \in \mathscr{E}_{\lambda,R}$, it is compatible with μ-orbits. First note that

$$\|F_\sigma(g \star \mu) - g \star F_\sigma(\mu)\|_\sigma \leq \|F_\sigma(g \star \mu) - F_\sigma(h \star \mu)\|_\sigma + \|h \star F_\sigma(\mu) - g \star F_\sigma(\mu)\|_\sigma.$$

We further note that

$$\|h \star F_\sigma(\mu) - g \star F_\sigma(\mu)\|_\sigma = \|h \circ T_\sigma^\mu - g \circ T_\sigma^\mu\|_\sigma = \|h - g\|_\mu \leq \|f_\mu\|_\infty^{1/2} \varepsilon.$$

To bound $\|F_\sigma(g \star \mu) - F_\sigma(h \star \mu)\|_\sigma$, we distinguish the two cases as in the theorem:

1. We use the Hölder bound (A.13) and (A.4):

$$\|F_\sigma(g \star \mu) - F_\sigma(h \star \mu)\|_\sigma \leq C W_2(g \star \mu, h \star \mu)^{2/15} \leq C \|g - h\|_\mu^{2/15} \leq C \left(\|f_\mu\|_\infty^{1/2} \varepsilon\right)^{2/15}.$$

   Therefore, overall, F_σ is δ-compatible with $\delta = \|f_\mu\|_\infty^{1/2} \varepsilon + C \left(\|f_\mu\|_\infty^{1/2} \varepsilon\right)^{2/15}$.

2. Corollary A.2 implies

$$\|F_\sigma(g_\sharp \mu) - F_\sigma(h_\sharp \mu)\|_\sigma \leq \left(\sqrt{\frac{4R}{K_\mu^\sigma}} + 1\right) \|g - h\|_\mu + \sqrt{4R \frac{W_2(\sigma, \mu) + R + \|\operatorname{Id}\|_\mu}{K_\mu^\sigma}} \|g - h\|_\mu^{1/2}$$

$$\leq \left(\sqrt{\frac{4R}{K_\mu^\sigma}} + 1\right) \left(\|f_\mu\|_\infty^{1/2} \varepsilon\right) + \sqrt{4R \frac{W_2(\sigma, \mu) + R + \|\operatorname{Id}\|_\mu}{K_\mu^\sigma}} \left(\|f_\mu\|_\infty^{1/2} \varepsilon\right)^{1/2}.$$

   Therefore, overall, F_σ is δ-compatible with

$$\delta = \left(\sqrt{\frac{4R}{K_\mu^\sigma}} + 2\right) \left(\|f_\mu\|_\infty^{1/2} \varepsilon\right) + \sqrt{4R \frac{W_2(\sigma, \mu) + R + \|\operatorname{Id}\|_\mu}{K_\mu^\sigma}} \left(\|f_\mu\|_\infty^{1/2} \varepsilon\right)^{1/2}.$$

Similarly, it can be shown that F_σ is δ-compatible with ν-orbits (now δ depending on ν). Thus by taking the maximum between those δ values and multiplying by 6 (distance conditions in Theorem A.5), all the assumptions of Theorem A.5 are satisfied and linear separability follows. □

A.5 *A useful result in normed spaces*

In this section, we derive a result on almost convex sets for general normed spaces. It states that if two almost convex sets are separated by a positive value, then their convex hull can also be separated.

This result is needed for the almost linear separability proof for perturbed shifts and scalings (Theorem 4.2 and A.5).

Definition A.4. Let $(X, \|\cdot\|)$ be a normed space and let $\varepsilon > 0$. X is called ε-convex if for every $x_1, x_2 \in X$ and $c \in [0, 1]$ there exists $x \in X$ such that

$$\|(1 - c)\, x_1 + c\, x_2 - x\| < \varepsilon.$$

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022

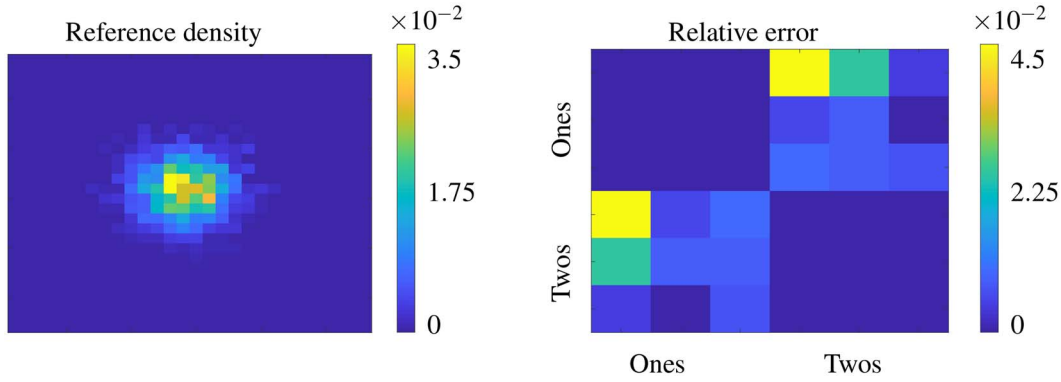

FIG. A1. *Left:* The reference density σ (Gaussian) used in the LOT embedding to approximate the Wasserstein distances between images of 1s and 2s from the MNIST data set (right). *Right:* We consider images of 1s and 2s from the MNIST data set. Both the images of 1s and 2s are shifts of each other. The right panel shows the relative error between the ground truth Wasserstein distances of the data set and the distances computed by the LOT embedding. The error is 0 in each individual class of 1s and 2s since the LOT embedding is an isometry on these subsets (since they have been produced by shifts). In the inbetween classes, i.e. between 1s and 2s, an error of order 10^{-2} is observed.

This definition states that for an ε-convex set X, $d(\text{conv}(X), X) < \varepsilon$, where $\text{conv}(X)$ denotes the convex hull of X and d is the distance between sets.

LEMMA A.6. Let $(X, \|\cdot\|)$ be a normed space and let $\varepsilon > 0$. Consider two ε-convex sets $A, B \subset X$. If $d(A, B) > 3\varepsilon$, then $d(\text{conv}(A), \text{conv}(B)) > \varepsilon$.

Proof. Let $a \in A$ and $c_b \in \text{conv}(B)$. Let $b \in B$ such that $\|c_b - b\| < \varepsilon$ Then,

$$\|a - c_b\| \geq \|a - b\| - \|b - c_b\| > 3\varepsilon - \varepsilon = 2\varepsilon.$$

Therefore, $d(A, \text{conv}(B)) > 2\varepsilon$. Similarly one can prove that $d(B, \text{conv}(A)) > 2\varepsilon$.

Now let $c_a \in \text{conv}(A)$ and $c_b \in \text{conv}(B)$ and choose $b \in B$ such that $\|c_b - b\| < \varepsilon$. Then, we have

$$\|c_a - c_b\| \geq \|c_a - b\| - \|b - c_b\| > 2\varepsilon - \varepsilon = \varepsilon,$$

which implies that $d(\text{conv}(A), \text{conv}(B)) > \varepsilon$. □

A.6 *Experiment: Wasserstein distance approximation*

We show an experiment related to the Wasserstein distance approximation result of Section 4.1. We consider MNIST images (1s and 2s), where both the 1s and the 2s are shifts of each other. Figure A1 shows the approximation error of LOT for the individual and inbetween classes, using a Gaussian as reference. We observe that there is no error in the class of 1s and 2s, because they are shifts of each other (isometry result of Corollary 4.1).

We mention that [22] has additional Wasserstein distance approximation results and experiments.

Downloaded from https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaac023/6713906 by guest on 05 October 2022