Geometrically Regularized Wasserstein Dictionary Learning

Marshall Mueller [1] **Shuchin Aeron** [2] **James M. Murphy** [*] [1] **Abiy Tasissa** [*] [1]

Abstract

Wasserstein dictionary learning is an unsupervised approach to learning a collection of probability distributions that generate observed distributions as Wasserstein barycentric combinations. Existing methods solve an optimization problem that only seeks a dictionary and weights that minimize the reconstruction accuracy. However, there is no a priori reason to believe there are unique solutions in general to this problem. Moreover, the learned dictionary is, by design, optimized to represent the observed data set, and may not be useful for classification tasks or generative modeling. Just as regularization plays a key role in linear dictionary learning, we propose a geometric regularizer for Wasserstein space that promotes representations of a data distribution using nearby dictionary elements. We show that this regularizer leads to barycentric weights that concentrate on dictionary atoms local to each data distribution. When data are generated as Wasserstein barycenters of fixed distributions, this regularizer facilitates the recovery of the generating distributions in cases that are ill-posed for unregularized Wasserstein dictionary learning. Through experimentation on synthetic and real data, we show that our geometrically regularized approach yields more interpretable dictionaries in Wasserstein space which perform better in downstream applications.

1. Introduction

A central goal of statistical signal processing is the discovery of latent structures in complex data. The classical *manifold hypothesis* posits that data living in a high-dimensional ambient space can be well approximated by low-dimensional manifolds or mixtures thereof, which circumvents the curse of dimensionality that plagues high-dimensional statistics. Linear dimensionality reduction methods such as principal component analysis (PCA) (Hotelling, 1933) and non-linear manifold learning approaches that exploit local connectivity structure in the data (Scholkopf et al., 1997; Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003; Coifman & Lafon, 2006) rely on this assumption of intrinsically low-dimensional structure in high-dimensional data to glean insights. These techniques typically output a low-dimensional representation preserving local geometric structures (e.g., pairwise Euclidean or geodesic distances).

An alternative perspective for efficiently representing complex data is the sparse coding and dictionary learning paradigm (Olshausen & Field, 1996; 1997; Barlow et al., 1961; Hromádka et al., 2008). In the simplest setting when data are considered as elements of $\mathbb{R}^d$ (or more generally a normed vector space), the aim of sparse coding is to represent data $\{\mathbf{y}_i\}_{i=1}^n \subset \mathbb{R}^d$, stacked as rows in the matrix $\mathbf{Y} \in \mathbb{R}^{n\times d}$, as a linear combination of vectors $\{\mathbf{d}_j\}_{j=1}^m$, stacked as a *dictionary* matrix $\mathbf{D} \in \mathbb{R}^{m\times d}$ such that $\mathbf{Y} \approx \mathbf{\Lambda}\mathbf{D}$ for some coefficients $\mathbf{\Lambda} \in \mathbb{R}^{n\times m}$, perhaps subject to constraints on $\mathbf{\Lambda}$. When the dictionary $\mathbf{D}$ is fixed, this reduces to an optimization over $\mathbf{\Lambda}$ (Mallat, 1999; Engan et al., 2000). More generally, $\mathbf{D}$ and $\mathbf{\Lambda}$ can be learned simultaneously with some additional constraints on the dictionary or coefficients (Lee & Seung, 1999; Aharon et al., 2006):

$$\arg\min_{\mathbf{D},\mathbf{\Lambda}} \mathcal{L}(\mathbf{Y}, \mathbf{\Lambda}\mathbf{D}) + \rho\mathcal{R}(\mathbf{D}, \mathbf{\Lambda}) \tag{1}$$

for some loss function $\mathcal{L}$ (e.g., $\mathcal{L}(\mathbf{Y}, \mathbf{\Lambda}\mathbf{D}) = \|\mathbf{Y} - \mathbf{\Lambda}\mathbf{D}\|_F^2$) and regularization function $\mathcal{R}$ (e.g., $\mathcal{R}(\mathbf{D}, \mathbf{\Lambda}) = \|\mathbf{\Lambda}\|_1$) balanced by a parameter $\rho > 0$. The regularizers ensure well-posedness of the problem and improve interpretability and robustness. The problem (1) is the *dictionary learning problem* in $\mathbb{R}^d$.

The imposed Euclidean structure is convenient computationally but limiting in practice, as many real data are better modeled as living in spaces with non-Euclidean geometry where instead $\mathbf{Y} \approx \mathcal{F}(\mathbf{D}, \mathbf{\Lambda})$ for some nonlinear reconstruction function $\mathcal{F}$ (Tuzel et al., 2006; 2007; Li et al., 2008; Guo et al., 2010; Harandi et al., 2013; 2015; Cherian & Sra,

[*] Equal contribution [1] Department of Mathematics, Tufts University, Medford, MA, USA [2] Department of Electrical and Computer Engineering, Tufts University, Medford, MA, USA. Correspondence to: James M. Murphy <jm.murphy@tufts.edu>, Abiy Tasissa <abiy.tasissa@tufts.edu>.

Proceedings of the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

2016; Yin et al., 2016; Maggioni et al., 2016; Liu et al., 2018; Schmitz et al., 2018; Tankala et al., 2020). Important questions in this setting are what notion of reconstruction should take the place of linear combination (i.e., $\mathcal{F}$), how reconstruction quality is assessed without the use of a global norm (i.e., $\mathcal{L}$), and what constraints are natural on the coefficients in the nonlinear space (i.e., $\mathcal{R}$).

This paper focuses on dictionary learning for data that are modeled as *probability distributions in Wasserstein space*. This basic framework was pioneered by (Schmitz et al., 2018), where the authors leverage the theory and algorithms of optimal transport to propose the *Wasserstein dictionary learning (WDL)* algorithm, whereby a data point (interpreted as a probability distribution or histogram in $\mathbb{R}^d$) is approximated as a Wasserstein barycenter (Agueh & Carlier, 2011) of the learned dictionary atoms. The resulting framework is focused on learning a dictionary that reconstructs well, but neglects other desirable aspects of a dictionary such as sparsity of the learned coefficients to promote latent feature representations. Moreover, WDL is ill-posed in two senses: (i) for a fixed dictionary, unique coefficients are not assured; (ii) there may be multiple dictionaries that enable perfect reconstruction of the observed data; see Figure 1.

Summary of Contributions: We generalize the classical WDL algorithm (Schmitz et al., 2018) by incorporating a novel Wasserstein *geometric regularizer*. Our regularizer encourages an observed data point to be reconstructed as a barycentric representation from *nearby* (in the sense of Wasserstein distances) dictionary atoms. As we vary the balance parameter for this regularizer, the proposed method interpolates between classical WDL (no regularization) and Wasserstein K-means (strong regularization). Unlike the original formulation, the proposed regularizer learns dictionary atoms with geometric similarity to the training data. Theoretically, we characterize the concentration of learned coefficients on nearby atoms to a data distribution, and show that the original generating atoms are uniquely recovered when data is modeled as a barycenter of two distributions. Empirically, we provide evidence that the regularized problem can learn the generating atoms even when there are more than two generators, and that our scheme yields more interpretable and useful coefficients for downstream classification tasks; the code to reproduce all experiments in this paper can be found here: `https://github.com/MarshMue/GeoWDL-TAG-ML-ICML-2023`.

Notations and Preliminaries: Lowercase and uppercase boldface letters denote (column) vectors and matrices, respectively. We generally use Greek letters to denote measures, with the exception that $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m$ denotes the dictionary when its elements are measures. We denote the Euclidean norm of a vector $\mathbf{x}$ as $||\mathbf{x}||_2$. We let $\langle \cdot, \cdot \rangle$ be the Euclidean inner product when applied to vectors and the trace inner product when applied to matrices. Let $\Delta^m = \{\mathbf{x} \in \mathbb{R}^m \mid \sum_{i=1}^m x_i = 1, \forall i = 1, \dots, m, x_i \geq 0\}$. Softmax, as a change of variables, is defined as $\sigma(\mathbf{x}) := \exp(\mathbf{x})/\langle \exp(\mathbf{x}), \mathbb{1}_N \rangle$, where we take the exponential to be an elementwise operation on the vector and use $\mathbb{1}_N$ to denote the ones vector of size N. When we write $\sigma(\mathbf{X})$ for a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ it means applying σ to each row.

2. Background and Related Work

Classical Dictionary Learning: In (1), using $\mathcal{L}(\mathbf{Y}, \mathbf{\Lambda D}) = \|\mathbf{Y} - \mathbf{\Lambda D}\|_F^2$ and $\rho = 0$ yields an optimization problem with optimal dictionary and coefficients given by the m singular components with largest singular values (Eckart & Young, 1936). To promote sparse coefficients that still realize $\mathbf{Y} \approx \mathbf{\Lambda D}$, the prototypical regularized dictionary learning problem is $\min_{\mathbf{D},\mathbf{\Lambda}} ||\mathbf{Y} - \mathbf{\Lambda D}||_F^2 + \rho||\mathbf{\Lambda}||_1$ where $||\mathbf{\Lambda}||_1 = \sum_{i=1}^n \sum_{j=1}^m |\Lambda_{ij}|$ is a sparsity-promoting regularizer (Donoho, 2006; Elad, 2010). In the non-negative matrix factorization (NMF) paradigm, non-negativity constraints are imposed on the atoms and coefficients (Lee & Seung, 1999; 2000; Berry et al., 2007).

Optimal Transport: We provide basic background on optimal transport; for more general treatments and theory, see (Ambrosio et al., 2005; Villani, 2021; Santambrogio, 2015; Peyré et al., 2019). Let $\mathcal{P}(\mathbb{R}^d)$ be the space of probability measures in $\mathbb{R}^d$. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Let $\Pi(\mu, \nu)$ be the space of measures on $\mathbb{R}^d \otimes \mathbb{R}^d$ with marginals μ, ν. The squared *Wasserstein-2 distance* is defined as:

$$W_2^2(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|_2^2 \, d\pi(\mathbf{x}, \mathbf{y}) \qquad (2)$$

Given measures $\{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ that have finite second moments, along with a vector $\boldsymbol{\lambda} \in \Delta^m$, the *Wasserstein-2 barycenter* (Agueh & Carlier, 2011) is defined as:

$$\mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda}) := \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathbb{R}^d)} \sum_{j=1}^m \lambda_j W_2^2(\mathcal{D}_j, \mu) \qquad (3)$$

The measure $\mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda})$ can be interpreted as a weighted average of the $\{\mathcal{D}_j\}_{j=1}^m$, with the impact of $\mathcal{D}_j$ proportional to λ_j. Wasserstein barycenters have been proven useful in a range of applications, and are in a precise sense the "correct" way to average measures, in that $\mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda})$ preserves the geometric properties of $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m$ in a way that linear mixtures do not (Agueh & Carlier, 2011; Rabin et al., 2011; Cuturi & Doucet, 2014; Bonneel et al., 2016). Wasserstein barycenters are intimately connected to geodesics in Wasserstein space, in the following sense. For π^* optimizing (2), define the *McCann interpolation* of μ, ν as $(P_t)_{\#}\pi^*$ where $P_t(\mathbf{x}, \mathbf{y}) = (1 - t)\mathbf{x} + t\mathbf{y}$ for $t \in [0, 1]$ and where $(P_t)_{\#}$ denotes the pushforward by P_t. The McCann interpolation is the constant-speed geodesic between μ, ν in the

Wasserstein-2 space (McCann, 1997; Ambrosio et al., 2005) and coincides with the Wasserstein barycenter with weight $\boldsymbol{\lambda} = (1-t, t)$ on μ, ν.

The use of Wasserstein distances in linear dictionary learning problem (1) has been considered with $\mathcal{L} = W_2$ and $\mathcal{R}(\mathbf{D}, \boldsymbol{\Lambda}) = -\rho_1 \langle \mathbf{D}, \log \mathbf{D} \rangle - \rho_2 \langle \boldsymbol{\Lambda}, \log \boldsymbol{\Lambda} \rangle$ to promote positivity of $\mathbf{D}$ and $\boldsymbol{\Lambda}$ (Rolet et al., 2016). The problem has the added constraint that the data reconstructions are probability distributions, i.e., $\boldsymbol{\Lambda}\mathbf{D} \subset (\Delta^d)^n$ in order to make $\mathcal{L}$ sensible. One can also use the related Gromov-Wasserstein distance to perform dictionary learning on graphs (Vincent-Cuaz et al., 2021). Both of these methods have a linear generative model that differs from the non-linear generative model that we consider.

3. Geometric Regularization for Wasserstein Dictionary Learning

WDL (Schmitz et al., 2018) aims to find a dictionary of probability distributions $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ such that observed data $\{\mu_i\}_{i=1}^n \subset \mathcal{P}(\mathbb{R}^d)$ can be represented as Wasserstein barycenters of the collection $\mathcal{D}$. The precise optimization problem is

$$\underset{\substack{\mathcal{D} \subset \mathcal{P}(\mathbb{R}^d) \\ \boldsymbol{\Lambda} \in (\Delta^m)^n}}{\arg\min} \sum_{i=1}^n \mathcal{L}(\mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda}_i), \mu_i), \tag{4}$$

where the loss function $\mathcal{L}$ is typically taken to be W_2^2 and $\boldsymbol{\lambda}_i \in \Delta^m$ is a vector of size m, corresponding to a row of $\boldsymbol{\Lambda}$. In other words, solving this problem finds the dictionary of probability distributions yielding the best approximations to each data point μ_i using barycentric combinations of $\mathcal{D}$. WDL was proposed in part as an alternative to geodesic principal component analysis in Wasserstein space (Boissard et al., 2015; Seguy & Cuturi, 2015; Bigot et al., 2017), and has a demonstrated ability to produce meaningful atoms for representing probability distributions. However, the problem formulation is ill-posed: even with a fixed dictionary $\mathcal{D}$ the learned coefficients may not be unique, and more broadly there may be multiple dictionaries that can reconstruct the data perfectly.

Another way to see this is by comparing WDL to linear dictionary learning: (4) is similar to (1) but without a regularization term. Promotion of sparsity of coefficients on the simplex can be done with entropy, projection onto the simplex, and suitable use of the ℓ^2 norm (Donoho et al., 1992; Shashanka et al., 2007; Larsson & Ugander, 2011; Kyrillidis et al., 2013; Li et al., 2020). Our focus, on the other hand, is on geometric regularization as follows. In the linear setting, the *geometric regularizer* $\sum_{j=1}^m \lambda_j ||\mathbf{y} - \mathbf{d}_j||_2^2$ (for an individual data point $\mathbf{y}$ with representation coefficient $\boldsymbol{\lambda}$, constrained to lie on the probability simplex, with respect to dictionary $\{\mathbf{d}_j\}_{j=1}^m$) has been proven to promote sparsity by favoring *local representations*, namely reconstructing using nearby (with respect to Euclidean distances) dictionary atoms (Yu et al., 2009; Tankala et al., 2020; Zhong & Pun, 2020).

We propose to regularize (4) with a novel *Wasserstein geometric regularizer*:

$$\mathcal{R}_G(\mathcal{D}, \boldsymbol{\Lambda}) := \sum_{i=1}^n \sum_{j=1}^m (\boldsymbol{\lambda}_i)_j W_2^2(\mathcal{D}_j, \mu_i). \tag{5}$$

This yields a new, regularized objective:

$$\mathcal{G}(\mathcal{D}, \boldsymbol{\Lambda}, \{\mu_i\}_{i=1}^n) := \sum_{i=1}^n W_2^2(\mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda}_i), \mu_i) + \rho \mathcal{R}_G(\mathcal{D}, \boldsymbol{\Lambda}) \tag{6}$$

where $\rho > 0$ is a tunable balance parameter. We will learn dictionaries

$$(\mathcal{D}^*, \boldsymbol{\Lambda}^*) = \underset{\substack{\mathcal{D} \subset \mathcal{P}(\mathbb{R}^d) \\ \boldsymbol{\Lambda} \in (\Delta^m)^n}}{\arg\min}\ \mathcal{G}(\mathcal{D}, \boldsymbol{\Lambda}, \{\mu_i\}_{i=1}^n).$$

At first glance the geometric regularizer (5) resembles the objective in the definition of the Wasserstein barycenter in (3) and may be appear redundant. However, the barycenter minimization is only indirect in WDL for finding dictionary atoms that provide representational capacity for the data; the atoms could be arbitrarily far from the data with no penalty to the objective (4).

Interpretations of $\mathcal{R}_G(\mathcal{D}, \boldsymbol{\Lambda})$: The regularization term $\mathcal{R}_G(\mathcal{D}, \boldsymbol{\Lambda})$ is analogous to Laplacian smoothing in Euclidean space (Cai et al., 2010; Dornaika & Weng, 2019) and can be interpreted as non-linear archetypal learning (Cutler & Breiman, 1994) in Wasserstein space. The geometric regularizer can also be seen to promote sparsity by penalizing the use of atoms that are far from the data to be represented and thus acts as a weighted ℓ_1 norm on the coefficients (Tasissa et al., 2021).

Connection with Wasserstein K-means: In Wasserstein K-means (Domazakis et al., 2019; Verdinelli & Wasserman, 2019; Zhuang et al., 2022), given observed measures $\{\mu_i\}_{i=1}^n \subset \mathcal{P}(\mathbb{R}^d)$ we want to find "centers" $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ solving the optimization problem $\min_{\mathbf{C}, \mathcal{D}} \sum_{i=1}^n \sum_{j=1}^m C_{ij} W_2^2(\mathcal{D}_j, \mu_i)$, where $\mathbf{C} \in \{0,1\}^{n \times m}$ is such that $\sum_{j=1}^m C_{ij} = 1$ for all $i = 1, \ldots, n$. Suppose $\mathcal{D}^* \subset \mathcal{P}(\mathbb{R}^d)$ and $\boldsymbol{\Lambda}^* \in \mathbb{R}^{n \times m}$ are the optimizers of (6). Note that for *any* feasible $\boldsymbol{\Lambda}$ and with dictionary fixed at $\mathcal{D}^*$, we have

$$\begin{aligned}\mathcal{R}_G(\mathcal{D}^*, \mathbf{\Lambda}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}(\boldsymbol{\lambda}_i)_j W_2^2(\mathcal{D}_j^*, \mu_i) \\ &\geq \sum_{i=1}^{n}\sum_{j=1}^{m}(\boldsymbol{\lambda}_i)_j \min_{1\leq p\leq m} W_2^2(\mathcal{D}_p^*, \mu_i) \\ &= \sum_{i=1}^{n} \min_{1\leq p\leq m} W_2^2(\mathcal{D}_p^*, \mu_i)\end{aligned}$$

Thus, for fixed $\mathcal{D}^*$, coefficients that minimize $\mathcal{R}_G(\mathcal{D}^*, \mathbf{\Lambda})$ have the property that the i^{th} row is all zeros except for a 1 at index $i^* = \arg\min_{1\leq p\leq m} W_2^2(\mathcal{D}_p^*, \mu_i)$. In this sense, for a fixed dictionary $\mathcal{D}^*$ and with each observation $\mu_1, \ldots, \mu_n$ having a unique nearest neighbor in $\mathcal{D}^*$, the optimal solution $\mathbf{\Lambda}^*$ is a matrix whose rows are binary and 1-sparse, which is exactly of the same form as the binary assignment in Wasserstein K-means. When the aforementioned assumption does not hold, uniqueness is not guaranteed but the 1-sparse solution is in the family of optimal solutions.

In this sense, incorporating the geometric regularizer (5) into the main objective in (4) with a scaling parameter ρ enables interpolation between learning a dictionary for pure reconstruction ($\rho = 0$) and one with sparsity promoted via K-means ($\rho \gg 0$). Indeed, (5) is like a *soft Wasserstein K-means objective*, in that it promotes assigning coefficient energy to a single, closest atom.

Learning Coefficients in a Fixed Dictionary: To evaluate the effect of (5), we consider the case where multiple barycentric weights may reconstruct a measure. For a fixed dictionary $\{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ and a target measure μ, we consider the following problem:

$$\begin{aligned}&\arg\min_{\boldsymbol{\lambda}\in\Delta^m} \sum_{j=1}^{m} \lambda_j W_2^2(\mathcal{D}_j, \mu) \\ &\text{subject to } \mu = \mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda}).\end{aligned} \tag{7}$$

This is a coding problem under the constraint that μ is *exactly reconstructed* in the sense of Wasserstein barycenters. The *barycentric coding model* analyzed in (Werenski et al., 2022) gives a characterization of when $\mu = \mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda})$, which can be leveraged to rephrase (7) as follows; a precise statement with explicit regularity assumptions and proof appear in the Appendix.

Proposition 3.1. *Let μ be fixed and let $\{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ be a fixed dictionary. Under suitable regularity assumptions on μ and $\{\mathcal{D}_j\}_{j=1}^m$, the solution to (7) is given by*

$$\arg\min_{\boldsymbol{\lambda}\in\Delta^m} \boldsymbol{\lambda}^T \mathbf{c} \textit{ subject to } \mathbf{A}\boldsymbol{\lambda} = \mathbf{0}, \tag{8}$$

where $\mathbf{c}$ and $\mathbf{A} \in \mathbb{R}^{m\times m}$ are uniquely determined by $\mu, \{\mathcal{D}_j\}_{j=1}^m$.

Importantly, $\mathbf{c}$ and $\mathbf{A}$ are determined by $\{\mathcal{D}\}_{j=1}^m$ for a fixed μ, so that (8) is a linear program in $\boldsymbol{\lambda}$. In general for fixed $\mathcal{D}$ and μ, solving $\mathbf{A}\boldsymbol{\lambda} = \mathbf{0}$ subject to $\boldsymbol{\lambda} \in \Delta^m$ may have multiple solutions (Werenski et al., 2022). Among all the possible barycentric representations of μ, (8) chooses the one "closest" to the dictionary atoms themselves, and thereby promotes unique codes under the hard reconstruction constraint. We note that $\boldsymbol{\lambda}$ is generally robust to perturbations of $\mathcal{D}$; see the Appendix for details.

Characterization of Optimal Solutions to (7): For barycenters whose weights are primarily supported on their nearest neighbors, we can show that solving (7) will obtain weights that are also concentrated on their nearest neighbors. Let $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ be a fixed dictionary. For a positive integer $k \leq n$, consider the generative model $\mu = \mathrm{Bary}(\mathcal{D}, \mathbf{b})$ where $\mathbf{b} \in \Delta^m$ is a vector supported on $N_k \subset \{1, 2., , ., m\}$, the indices of the atoms which are the k-nearest neighbours of μ with respect to W_2. That is, μ is a barycenter of its k-nearest neighbours in $\{\mathcal{D}_j\}_{j=1}^m$.

The following Proposition shows that solutions to (7) concentrate near N_k; its proof appears in the Appendix.

Proposition 3.2. *Let $\mu = \mathrm{Bary}(\mathcal{D}, \mathbf{b})$ with $\mathbf{b}$ supported on the indices of the k-nearest neighbors of μ among $\mathcal{D}$. Suppose without loss of generality that the dictionary elements are ordered in increasing order from μ: $W_2^2(\mathcal{D}_1, \mu) \leq W_2^2(\mathcal{D}_2, \mu) \leq \cdots \leq W_2^2(\mathcal{D}_m, \mu)$. Let $\boldsymbol{\lambda}^*$ be the solution of (7).*

Then for all $i \in \{1, \ldots, m\}$ such that $W_2^2(\mathcal{D}_i, \mu) > W_2^2(\mathcal{D}_1, \mu)$,

$$\lambda_i^* \leq \frac{W_2^2(\mathcal{D}_k, \mu) - W_2^2(\mathcal{D}_1, \mu)}{W_2^2(\mathcal{D}_i, \mu) - W_2^2(\mathcal{D}_1, \mu)}.$$

Proposition 3.2 is vacuous when referring to the coefficients corresponding to the k-nearest neighbors of μ (the measures used to generate μ), but it provides a bound on how much coefficient mass can concentrate *off* these nearest neighbors. As an illustration, consider two representative cases. First, if $W_2^2(\mathcal{D}_{k+1}, \mu) \gg W_2^2(\mathcal{D}_k, \mu)$, then $\lambda_i \ll 1$ for all $i \geq k + 1$ (i.e., very little coefficient mass comes from reference measures outside the k nearest neighbors of μ). Second, if $W_2^2(\mathcal{D}_1, \mu) = W_2^2(\mathcal{D}_2, \mu) = \cdots = W_2^2(\mathcal{D}_k, \mu)$, then for any $i \geq k + 1$ such that $W_2^2(\mathcal{D}_i, \mu) > W_2^2(\mathcal{D}_1, \mu)$, it follows that $\lambda_i = 0$. In particular in this case, if $W_2^2(\mathcal{D}_{k+1}, \mu) > W_2^2(\mathcal{D}_k, \mu)$, then the support set of the generating coefficients is correctly identified.

$\mathcal{R}_G(\mathcal{D}, \mathbf{\Lambda})$ Promotes Unique Solutions to WDL: The unregularized WDL problem (4) does not in general have a unique solution. This can be seen intuitively in the case where the data are generated as barycenters of two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. In this case, any barycenter coincides with a point along the McCann interpolation:

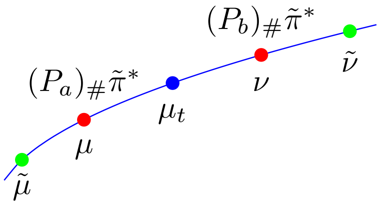


Figure 1. The measures $\tilde{\mu}$ and $\tilde{\nu}$ have the capacity to represent any barycenter μ_t of $\mu = (P_a)_\#\tilde{\pi}$, $\nu = (P_b)_\#\tilde{\pi}$, but they do it in a manner that our geometric regularizer penalizes.

$\text{Bary}(\{\mu,\nu\},(1-t,t)) = (P_t)_\#\pi^*$ where π^* is the optimal transport plan between μ,ν. Then any measures $\tilde{\mu},\tilde{\nu}$ whose McCann interpolation passes through μ,ν will also generate any barycenters of μ,ν. This is visualized in Figure 1. We will show that in this special case, our geometric regularizer (3.1) addresses this ill-posedness of WDL.

Note, this is an issue of *non-uniqueness over dictionaries* $\mathcal{D}$; the simpler issue of *non-uniqueness for a fixed* $\mathcal{D}$ is analyzed in Proposition 3.1. Indeed, the non-uniqueness for a fixed $\mathcal{D}$ is characterized (Werenski et al., 2022) by the solution space to $\mathbf{A}\boldsymbol{\lambda} = \mathbf{0}$ intersecting Δ^m in multiple places. On the other hand, our analysis of non-uniqueness over dictionaries requires an analysis of McCann interpolations.

Definition 3.3. Let $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ have optimal transportation plan π^* and $\tilde{\mu},\tilde{\nu} \in \mathcal{P}(\mathbb{R}^d)$ have optimal transportation plan $\tilde{\pi}^*$. The measures $\tilde{\mu},\tilde{\nu}$ are said to *contain the McCann interpolation* $\{(P_t)_\#\pi^*\}_{t\in[0,1]}$ between μ and ν if there exists an interval $[a,b] \subset [0,1]$ such that $\forall t \in [0,1]$, $\exists s \in [a,b]$ such that $(P_s)_\#\tilde{\pi}^* = (P_t)_\#\pi^*$.

We define the set of all pairs of measures $(\tilde{\mu},\tilde{\nu})$ that contain the McCann interpolation between μ and ν as $M(\mu,\nu)$. Pairs of measures in $M(\mu,\nu)$ can be thought of as generators of "extensions" of the McCann interpolation between μ,ν. In this sense, barycenters of $(\tilde{\mu},\tilde{\nu})$ can perfectly reconstruct any barycenter of (μ,ν) if and only if $(\tilde{\mu},\tilde{\nu}) \in M(\mu,\nu)$. We show that any "extension" of the McCann interpolation from one side results in an increase in the geometric regularizer for any measure in the original interpolation. The proof, which depends on the geodesic properties of McCann interpolation (Ambrosio et al., 2005), appears in the Appendix.

Theorem 3.4. *Consider measures $\mu,\nu,\tilde{\nu}$ with optimal transportation plan π^* between μ and ν and and $\tilde{\pi}^*$ between μ and $\tilde{\nu}$. Suppose $(\mu,\tilde{\nu}) \in M(\mu,\nu)$. For $t \in [0,1]$, let $\mu_t = (P_t)_\#\pi^*$ be in the McCann interpolation between μ and ν, and let $s \in [0,1]$ be such that $(P_s)_\#\tilde{\pi}^* = (P_t)_\#\pi^*$. Then*

$$\begin{aligned}&(1-t)W_2^2(\mu,\mu_t) + tW_2^2(\nu,\mu_t)\\ &\le (1-s)W_2^2(\mu,\mu_t) + sW_2^2(\tilde{\nu},\mu_t)\end{aligned}$$

With this, we can establish that subject to the constraint that $\tilde{\mu},\tilde{\nu}$ provide perfect reconstruction (quantified by $(\tilde{\mu},\tilde{\nu}) \in M(\mu,\nu)$), minimizing the geometric regularizer yields a unique solution that coincides with the true generating atoms in the case of all observed data lying on a McCann interpolation.

Corollary 3.5. *Let $\mu \neq \nu$ be two measures with optimal transport plan π^*. For any $(\tilde{\mu},\tilde{\nu}) \in M(\mu,\nu)$, let $\tilde{\pi}^*$ be the associated optimal transport plan. Then for any barycenter $\mu_t = (P_t)_\#\pi^*$ generated by μ,ν,*

$$(\mu,\nu) = \underset{\substack{(\tilde{\mu},\tilde{\nu})\in M(\mu,\nu),\\ s \text{ s.t. } (P_s)_\#\tilde{\pi}^*=\mu_t}}{\arg\min} (1-s)W_2^2(\tilde{\mu},\mu_t) + sW_2^2(\tilde{\nu},\mu_t).$$

Proof. Apply Theorem 3.4 twice, the second time after reversing parameterization. □

In other words, among all pairs of measures that contain the geodesic between the original data generators μ and ν, the geometric regularizer for every data point is minimized by (μ,ν).

4. Proposed Algorithm

Optimization in Wasserstein space has been infeasible outside of problems with low-dimensional distributions due to the computational complexities of solving the transport problems (Bonneel et al., 2016). We make two primary design choices to make a tractable algorithm:

Shared Fixed Support: We assume all measures lie on the same fixed finite support $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$. So, each distribution μ can be represented as a probability distribution $\mathbf{a} \in \Delta^N$ via $\mu = \sum_{i=1}^N a_i\delta_{\mathbf{x}_i}$. We will abuse notation and write μ in place of $\mathbf{a}$ when referring to discrete measures. Having a fixed support enables us to compute the pairwise costs $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ upfront, which are used repeatedly in the transport and barycenter computations.

Entropic Regularization: We use the entropically regularized Wasserstein distance for all distance and Wasserstein barycenter computations (Cuturi, 2013; Agueh & Carlier, 2011; Benamou et al., 2015) (in particular, we use the entropic estimate of the distance obtained via the dual formulation of the problem as detailed in Chapter 4.5 of (Peyré et al., 2019)). This way we can get simple and relatively cheap estimates of both the Wasserstein distance and barycenter by a few iterations of the Sinkhorn matrix scaling algorithm. We refer the reader to the aforementioned references for detailed discussion of the entropic regularization and its effect on computation. As most operations are now matrix-vector multiplications, our method is well-suited for automatic differentiation and can be efficiently implemented to handle variable updates (Paszke et al., 2019). For all transport computations one could use the unbiased Sinkhorn divergences instead (Feydy et al., 2019; Chizat et al., 2020); we choose not to in order to compare directly to WDL. The impact of entropic regularization on our theoretical results is discussed in the Appendix.

Our main algorithm, which we call *Geometric Wasserstein Dictionary Learning (GeoWDL)*, is detailed in Algorithm 1. Following the original WDL formulation of (Schmitz et al., 2018), we optimize over arbitrary vectors in Euclidean space, each of which represents a unique probability distribution (both for atoms and barycentric weights) via softmax as a change of variables; in our algorithm these are the variables α and β for the dictionary and weights, respectively.

Algorithm 1 Geometric Wasserstein Dictionary Learning (GeoWDL)

```
Input: Training data $\{\mu_i\}_{i=1}^n \subset \Delta^N$, $L$, optimizer
Initialize variables $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^{m\times N}$, $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{n\times m}$ {Use any initialization method}
for $k \leftarrow 1, \dots, L$ do
    $\mathcal{D}^{(k)} \leftarrow \sigma(\boldsymbol{\alpha})$, $\boldsymbol{\Lambda}^{(k)} \leftarrow \sigma(\boldsymbol{\beta})$ {Get updated dictionary/weights}
    loss $\leftarrow \mathcal{G}(\mathcal{D}^{(k)}, \boldsymbol{\Lambda}^{(k)}, \{\mu_i\}_{i=1}^n)$ {Compute the objective function}
    loss.backward() {Compute the gradients with automatic differentiation}
    Update $\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}$ with optimizer.step() {Update variables}
end for
Output: $\mathcal{D}^{(L)}, \boldsymbol{\Lambda}^{(L)}$
```

Atom Initialization: We consider 3 methods of initialization for the atoms: (i) uniform at random samples from Δ^N; (ii) uniform at random data samples: pick m of the data used to learn the dictionary as the initialization; (iii) k-means++ initialization: follow the initialization procedure of k-means++ algorithm using Wasserstein distances (Arthur & Vassilvitskii, 2006) and use those choices as the initial atoms[1]. We expect the data-based initialization schemes to converge faster and to better solutions, particularly when regularizing with $\mathcal{R}_G$ since the probability distributions that resemble the data will be favored, as we assume generating distributions should resemble the data to some degree.

Weight Initialization: We consider 3 methods of initialization for the weights: (i) uniform at random samples from Δ^m; (ii) Wasserstein histogram regression (Bonneel et al., 2016) to match each data point to the initialized atoms; (iii) estimating weights using the quadratic program described by (Werenski et al., 2022); details of this approach are in the Appendix. Empirically, atom initialization was more important than the choice of weight initialization; we use method (i) for all experiments in Section 5.

In each of the initialization methods described above we obtain initial values for α and β by passing the initialized dictionary and weights through $\log$ element-wise, which inverts σ for vectors on the simplex.

5. Experiments

This section summarizes experiments on image and NLP data; further discussion is in the Appendix.

5.1. Identifying Generating Distributions With Synthetic MNIST Data

We demonstrate the utility of our geometric regularizer in identifying the generating probability distributions in a generative model. In this experiment, we randomly select 3 samples from each MNIST (LeCun, 1998) data class $\{0,1,2,\dots,9\}$. We normalize each image to have all pixels sum to 1 and interpret them as probability distributions in $\mathbb{R}^2$ with cost given by squared Euclidean distance between the pixel coordinates. For each of the classes, we use the 3 samples to generate 50 synthetic samples by forming barycenters constructed with weights sampled uniformly from Δ^3. This yields 500 total training points, each of which is a synthetic MNIST digit generated from a total of $3 \times 10 = 30$ real MNIST digits. We then train a dictionary to learn 30 atoms. An optimal solution is to learn the original generating set of 30 distributions.

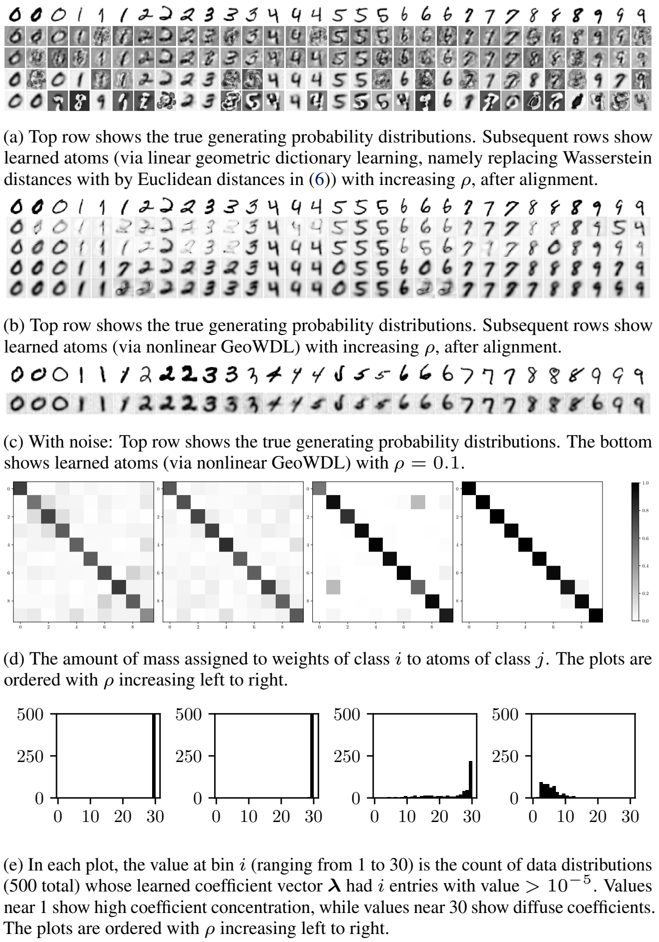
We run Algorithm 1 to compare the effects of geometric regularization by varying the regularization parameter $\rho \in \{0, 10^{-3}, 10^{-1}, 10^1\}$, noting that $\rho = 0$ corresponds to WDL (Schmitz et al., 2018). After learning the dictionary, we match the learned atoms to the true atoms by finding the assignment that minimizes the sum of the transport costs between learned and true atoms; for this we solve the non-entropically regularized transport problem to ensure non-negative assignment costs (Virtanen et al., 2020; Flamary et al., 2021). We visualize the learned vs. true atoms as well as a confusion matrix that demonstrates how well learned coefficients and atoms align with the true class in Figure 2.

For comparison we use code from (Tankala et al., 2020)—based on algorithm unrolling—to solve (1) regularized with the Euclidean *geometric regularizer* as a comparison to our similarly regularized nonlinear dictionary learning method. In essence, this comparison replaces Wasserstein distances with Euclidean distances in both the reconstruction loss and regularizer in (6). The results for the linear method are

[1] For this, one needs distances to be nonnegative. Sinkhorn "distances", as approximated by the value of the entropically regularized problem, may be negative. We work around this by adding the smallest number to the distances to make them positive.

visualized in Figure 2 (a), and should be contrasted with those in (b) for GeoWDL. We see that the linear method learns much more distorted and in some cases noisy digits, compared to our proposed GeoWDL method. We also note that the learned atoms from GeoWDL are more similar perceptually to the true atoms, aiding in interpretability.

To show robustness of GeoWDL to noise, we add noise to the synthetic data and show that we can still learn meaningful atoms. Details of the noise model can be found in the Appendix along with an example noisy digit shown in Figure 4. Learned atoms under noise are shown in Figure 2 (c).



(a) Top row shows the true generating probability distributions. Subsequent rows show learned atoms (via linear geometric dictionary learning, namely replacing Wasserstein distances with by Euclidean distances in (6)) with increasing ρ, after alignment.

(b) Top row shows the true generating probability distributions. Subsequent rows show learned atoms (via nonlinear GeoWDL) with increasing ρ, after alignment.

(c) With noise: Top row shows the true generating probability distributions. The bottom shows learned atoms (via nonlinear GeoWDL) with $\rho = 0.1$.

(d) The amount of mass assigned to weights of class i to atoms of class j. The plots are ordered with ρ increasing left to right.

(e) In each plot, the value at bin i (ranging from 1 to 30) is the count of data distributions (500 total) whose learned coefficient vector $\boldsymbol{\lambda}$ had i entries with value $> 10^{-5}$. Values near 1 show high coefficient concentration, while values near 30 show diffuse coefficients. The plots are ordered with ρ increasing left to right.

Figure 2. The learned digits shown in (a) and (b) are most interpretable in the moderate regularization regime (e.g., $\rho > 0$ but not too large). We show robustness to noise in (c). As ρ increases, (d) shows the concentration of coefficient energy on atoms that belong to the same class as the test data point, with atoms assigned by finding their closest training data. Moreover, we see in (e) that increasing ρ increases the essential sparsity of the coefficients learned, with many coefficient vectors having most values $< 10^{-5}$ when $\rho = 10^1$.

As observed empirically in Figure 2 (d), increasing ρ only helps the learned coefficients to be placed more correctly on their class after matching the atoms. This illustrates a trade-off associated with increasing ρ: more geometric regularization promotes concentration of weights on fewer learned atoms, but the learned atoms may not resemble the true generating atoms as well. Indeed, in Figure 2 (e) we see that increasing ρ concentrates the coefficients, as we would expect given the connection between GeoWDL with large ρ and Wasserstein K-means as well as Proposition 3.2 which establishes coefficient concentration; see Section 3. Examples of the atom learning dynamics, learned reconstructions, and training loss can be found in Figure 5 in the Appendix.

5.2. Learned Coefficients as Features for Document Classification

Here we demonstrate the effectiveness of the geometric regularizer when used for a classification task comparing word documents. We represent documents as probability distributions with a bag-of-words (i.e., a vector of counts for each word in the document normalized to lie in the probability simplex) approach where we use learned embeddings of the words from (Huang et al., 2016) as the support (words embedded in $\mathbb{R}^{300}$). The ground cost is the squared Euclidean distance in the embedding space. The documents considered come from the BBCSPORT dataset which consists of 737 documents; each document is one of five classes overall.

We *learn* a fixed size dictionary for each reference class and then compare our *learned* dictionaries to equally sized sets of *random* reference documents as follows. We vary the size the dictionary/set of random reference documents from 1 to 12 per class. Each dictionary is learned with a number of samples proportional to the size of the dictionary; for a dictionary of m atoms we use $4m$ training samples. To compare our learned dictionaries to random documents, we sample 100 documents for testing and classify them using the methods described below, where the reference documents in the classifier are either the output of our GeoWDL method or chosen randomly from all available documents of a particular class. Ideally, the learned dictionary atoms should be better representatives of the class than the random documents from the class. The sets of documents used for (i) learning the dictionaries representing each class (ii) randomly chosen for baseline comparison in the classifiers and (iii) testing are mutually disjoint. We repeat this test 30 times and report results averaged across these trials. The results are visualized in Figure 3 in the main text and Figure 6 in the Appendix. We note that these plots compare against two baselines in terms of what reference documents to use: (i) random samples from the data; (ii) WDL, which corresponds to $\rho = 0$ in the GeoWDL framework.

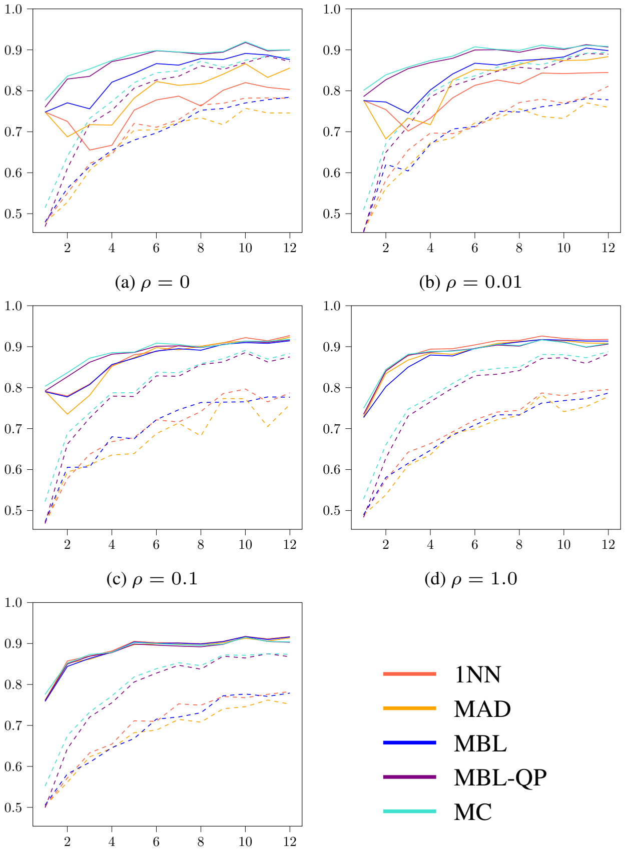
We notice that at all levels of geometric regularization tested, learned dictionaries enable the barycenter focused methods to outperform all methods that use random samples of the data. Perhaps more interestingly, we note that increasing the level of geometric regularization coincides with an increase in the performance of the “simpler” methods of 1NN and MAD. This suggests that the geometric regularizer is promoting dictionary atoms that are more informative generally *as individual atoms* as opposed to the information contained in their collective representational capacity via Wasserstein barycenters. As mentioned in the discussion of the geometric regularizer in Section 3, the unregularized WDL objective is minimized by the dictionary probability distributions only with respect of their representative ability. On the other hand, as evidenced here, GeoWDL encourages the learned atoms to be representative themselves of the data they model, even generalizing to unseen data.

Classification Methods: Below, we state the 5 classification methods considered in our NLP experiments, each of which requires a set of reference documents per class in order to label the testing data.

1. **1-Nearest Neighbor (1NN):** classifies based on the class of the W_2-nearest reference document;
2. **Minimum Average Distance (MAD):** selects the class with reference documents on average W_2-closest;
3. **Minimum Barycentric Loss (MBL):** we learn the barycentric weights to represent the test document by solving the quadratic program in (Werenski et al., 2022) for each reference class. We then compute the barycentric representation for each class and classify with the one that minimizes W_2 to the test document;
4. **Minimum Barycenter Loss (MBL-QP):** selects the class that minimizes the aforementioned quadratic program's objective (a proxy for the W_2 distance between the barycentric representation and test document);
5. **Maximum Coordinate (MC):** also solves the aforementioned quadratic program to estimate barycentric weights when using the reference documents *of all* classes to represent the test document. The class is then assigned based on the class whose total portion of the estimated weights is largest.

6. Conclusion

We have extended the WDL framework by introducing a geometric regularizer that interpolates between WDL and Wasserstein K-means according to a tuneable parameter ρ. We have shown the geometric regularizer itself is useful in solving uniqueness and identifiability problems relating to the dictionary and weights, by leveraging characterizations of the nonlinear problem of exact barycentric reconstruction as well as geometric properties of Wasserstein geodesics. Additionally, we have shown the usefulness of our extension in providing improvements to classification methods on real data. In particular our regularized framework improves over classical WDL in terms of atom interpretability and performance in classification.



(a) $\rho = 0$

(b) $\rho = 0.01$

(c) $\rho = 0.1$

(d) $\rho = 1.0$

(e) $\rho = 10.0$

Figure 3. Accuracy vs. number of reference documents. Solid lines denote reference documents being learned dictionary atoms while dashed lines denote the baseline of reference documents being random documents from each class. Here, we group by regularization parameter and show the impact of different methods and learning versus random atoms. Increasing ρ brings the non-barycentric based methods to performance parity with the barycentric based classification approaches. Note that $\rho = 0$ is unregularized WDL (Schmitz et al., 2018).

Future Work: Computational runtime remains a burden for both GeoWDL and classical WDL. While automatic differentiation provides for simple implementations, there may exist more specific algorithms to the dictionary learning framework; in particular the use of the geometric regularizer may enable faster algorithms as is done in the linear case (Mallat & Zhang, 1993). Relatedly, notions of linear optimal transport have important computational potential (Wang et al., 2013; Moosmüller & Cloninger, 2020; Hamm et al., 2022) in speeding up runtime of pairwise W_2-calculations for certain classes of measures.

The analysis of the encoding step in Section 3 does not immediately extend to case when the dictionary $\mathcal{D}$ is changing (which causes $\mathbf{A}$ to change). Understanding how the matrix $\mathbf{A}$ changes with $\mathcal{D}$ is a topic of ongoing research, and may allow for a closed-form solution to optimizers of (6). Relatedly, Theorem 3.4 applies only to the case when the original data lie exactly on a Wasserstein barycenter between $m = 2$ distributions; extending to $m \geq 3$ is a topic of ongoing research.

In the linear setting, the use of the geometric regularizer can be shown to promote sparsity (in the sense of few non-zero entries) in coefficients (Tankala et al., 2020). We believe that it should be possible to show the same in the Wasserstein setting, which would be a stronger claim than the coefficient concentration that we show in Proposition 3.2. Figure 2 (e) gives evidence to this claim as well and we believe the fact that entries are not exactly 0 to be caused by the use of softmax for change of variables.

Acknowledgements

We thank Demba Ba, JinCheng Wang, and Matt Werenski for insightful discussions. MM was partially supported by NSF DMS 1924513 and CCF-1934553. SA was partially supported by NSF CCF 1553075 and NSF DRL 1931978. JMM was partially supported by NSF DMS 1912737, NSF DMS 1924513, and The Camille & Henry Dreyfus Foundation. AT was partially supported by NSF DMS 2208392. The authors acknowledge the Tufts University High Performance Compute Cluster (https://it.tufts.edu/high-performance-computing) which was utilized for the research reported in this paper.

References

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Aharon, M., Elad, M., and Bruckstein, A. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Barlow, H. B. et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 1961.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

Best, M. J. and Chakravarti, N. Stability of linearly constrained convex quadratic programs. *Journal of Optimization Theory and Applications*, 64:43–53, 1990.

Bigot, J., Gouet, R., Klein, T., and López, A. Geodesic pca in the wasserstein space by convex pca. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53 (1):1–26, 2017.

Boissard, E., Le Gouic, T., and Loubes, J.-M. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21 (2):740–759, 2015.

Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

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

Cai, D., He, X., Han, J., and Huang, T. S. Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1548–1560, 2010.

Cherian, A. and Sra, S. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE transactions on neural networks and learning systems*, 28 (12):2859–2871, 2016.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Cominetti, R. and Martín, J. S. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67:169–187, 1994.

Cutler, A. and Breiman, L. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

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

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.

Delalande, A. and Merigot, Q. Quantitative stability of optimal transport maps under variations of the target measure. *arXiv preprint arXiv:2103.05934*, 2021.

Domazakis, G., Drivaliaris, D., Koukoulas, S., Papayiannis, G., Tsekrekos, A., and Yannacopoulos, A. Clustering measure-valued data with wasserstein barycenters. *arXiv preprint arXiv:1912.11801*, 2019.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67, 1992.

Dornaika, F. and Weng, L. Sparse graphs with smoothness constraints: Application to dimensionality reduction and semi-supervised classification. *Pattern Recognition*, 95: 285–295, 2019.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Elad, M. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

Engan, K., Aase, S. O., and Husoy, J. H. Multi-frame compression: Theory and design. *Signal Processing*, 80 (10):2121–2140, 2000.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021.

Guo, K., Ishwar, P., and Konrad, J. Action recognition using sparse representation on covariance manifolds of optical flow. In *2010 7th IEEE international conference on advanced video and signal based surveillance*, pp. 188–195. IEEE, 2010.

Hamm, K., Henscheid, N., and Kang, S. Wassmap: Wasserstein isometric mapping for image manifold learning. *arXiv preprint arXiv:2204.06645*, 2022.

Harandi, M., Sanderson, C., Shen, C., and Lovell, B. C. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *Proceedings of the IEEE international conference on computer vision*, pp. 3120–3127, 2013.

Harandi, M. T., Hartley, R., Lovell, B., and Sanderson, C. Sparse coding on symmetric positive definite manifolds using bregman divergences. *IEEE transactions on neural networks and learning systems*, 27(6):1294–1306, 2015.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Hromádka, T., DeWeese, M. R., and Zador, A. M. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS biology*, 6(1):e16, 2008.

Huang, G., Guo, C., Kusner, M. J., Sun, Y., Sha, F., and Weinberger, K. Q. Supervised word mover's distance. *Advances in neural information processing systems*, 29, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kyrillidis, A., Becker, S., Cevher, V., and Koch, C. Sparse projections onto the simplex. In *International Conference on Machine Learning*, pp. 235–243. PMLR, 2013.

Larsson, M. and Ugander, J. A concave regularization technique for sparse mixture models. *Advances in Neural Information Processing Systems*, 24, 2011.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lee, D. and Seung, H. S. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788–791, 1999.

Li, P., Rangapuram, S. S., and Slawski, M. Methods for sparse and low-rank recovery under simplex constraints. *Statistica Sinica*, 30(2):557–577, 2020.

Li, X., Hu, W., Zhang, Z., Zhang, X., Zhu, M., and Cheng, J. Visual tracking via incremental log-euclidean riemannian subspace learning. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2008.

Liu, T., Shi, Z., and Liu, Y. Kernel sparse representation on grassmann manifolds for visual clustering. *Optical Engineering*, 57(5):053104, 2018.

Maggioni, M., Minsker, S., and Strawn, N. Multiscale dictionary learning: non-asymptotic bounds and robustness. *The Journal of Machine Learning Research*, 17(1):43–93, 2016.

Mallat, S. *A wavelet tour of signal processing*. Elsevier, 1999.

Mallat, S. G. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

McCann, R. J. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Moosmüller, C. and Cloninger, A. Linear optimal transport embedding: Provable wasserstein classification for certain rigid transformations and perturbations. *arXiv preprint arXiv:2008.09165*, 2020.

Nutz, M. and Wiesel, J. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1-2):401–424, 2022.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Phu, H. X. and Yen, N. D. On the stability of solutions to quadratic programming problems. *Mathematical programming*, 89:385–394, 2001.

Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

Robinson, S. M. A characterization of stability in linear programming. *Operations Research*, 25(3):435–447, 1977.

Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed wasserstein loss. In *Artificial Intelligence and Statistics*, pp. 630–638. PMLR, 2016.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Scholkopf, B., Smola, A., and Müller, K.-R. Kernel principal component analysis. In *International conference on artificial neural networks*, pp. 583–588. Springer, 1997.

Seguy, V. and Cuturi, M. Principal geodesic analysis for probability measures under the optimal transport metric. *Advances in Neural Information Processing Systems*, 28, 2015.

Shashanka, M., Raj, B., and Smaragdis, P. Sparse overcomplete latent variable decomposition of counts data. *Advances in neural information processing systems*, 20, 2007.

Smith, C. S. and Knott, M. Note on the optimal transportation of distributions. *Journal of Optimization Theory and Applications*, 52(2):323–329, 1987.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Tankala, P., Tasissa, A., Murphy, J. M., and Ba, D. K-deep simplex: Deep manifold learning via local dictionaries. *arXiv preprint arXiv:2012.02134*, 2020.

Tasissa, A., Tankala, P., and Ba, D. Weighed ℓ_1 on the simplex: Compressive sensing meets locality. In *2021 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 476–480. IEEE, 2021.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Tuzel, O., Porikli, F., and Meer, P. Region covariance: A fast descriptor for detection and classification. In *European conference on computer vision*, pp. 589–600. Springer, 2006.

Tuzel, O., Porikli, F., and Meer, P. Human detection via classification on riemannian manifolds. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

Verdinelli, I. and Wasserman, L. Hybrid wasserstein distance and fast distribution clustering. *Electronic Journal of Statistics*, 13(2):5088–5119, 2019.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Vincent-Cuaz, C., Vayer, T., Flamary, R., Corneli, M., and Courty, N. Online graph dictionary learning. In *International Conference on Machine Learning*, pp. 10564–10574. PMLR, 2021.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., and Rohde, G. K. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

Werenski, M. E., Jiang, R., Tasissa, A., Aeron, S., and Murphy, J. M. Measure estimation in the barycentric coding model. In *International Conference on Machine Learning*, pp. 23781–23803. PMLR, 2022.

Yin, M., Guo, Y., Gao, J., He, Z., and Xie, S. Kernel sparse subspace clustering on symmetric positive definite manifolds. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5157–5164, 2016.

Yu, K., Zhang, T., and Gong, Y. Nonlinear learning using local coordinate coding. *Advances in neural information processing systems*, 22, 2009.

Zhong, G. and Pun, C.-M. Subspace clustering by simultaneously feature selection and similarity learning. *Knowledge-Based Systems*, 193:105512, 2020.

Zhuang, Y., Chen, X., and Yang, Y. Wasserstein k-means for clustering probability distributions. *arXiv preprint arXiv:2209.06975*, 2022.

A. Precise Statement and Proof of Proposition 3.1

In order to establish this result, it is essential to note that if measures μ and ν satisfy the constraints that they have finite second moments and do not give mass to small sets (e.g., are absolutely continuous) (Villani, 2021), their optimal plan π^* concentrates on the graph of $T^* = \nabla\phi$ for a strictly convex ϕ, so that

$$W_2^2(\mu,\nu) = \int_{\mathbb{R}^d} \|T^*(\mathbf{x}) - \mathbf{x}\|_2^2 d\mu(\mathbf{x})$$

where T^* satisfies the pushforward constraint $T^*_\#\mu = \nu$ and is called the *optimal transport map* (Smith & Knott, 1987; Brenier, 1991).

In order to precisely state Proposition 3.1, we require a few regularity assumptions **A1**-**A3** on the dictionary $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m$ and measure μ. These are required to invoke Theorem 1 in (Werenski et al., 2022), which characterizes exactly when

$$\mu = \mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda})$$

for some $\boldsymbol{\lambda} \in \Delta^m$.

A1: The measures $\{\mathcal{D}_j\}_{j=1}^m$ and μ are absolutely continuous and supported on either all of $\mathbb{R}^d$ or a bounded open convex subset. Call this shared support set Ω.

A2: The measures $\{\mathcal{D}_j\}_{j=1}^m$ and μ have respective densities $\{g_j\}_{j=1}^m$ and g which are bounded above and $g_1, ..., g_m$ are strictly positive on Ω.

A3: If $\Omega = \mathbb{R}^d$ then $\{g_j\}_{j=1}^m$ and g are locally Hölder continuous. Otherwise $\{g_j\}_{j=1}^m$ and g are bounded away from zero on Ω.

Proposition A.1 (Formal Statement of Proposition 3.1)**.** *Let μ be fixed and let $\{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ be a fixed dictionary. Consider*

$$\operatorname*{arg\,min}_{\boldsymbol{\lambda}\in\Delta^m} \sum_{j=1}^m \lambda_j W_2^2(\mathcal{D}_j, \mu) \text{ subject to } \mu = \mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda}). \tag{9}$$

*If $\mathcal{D}$ and μ satisfy the assumptions **A1**-**A3**, the solution to* (9) *is given by*

$$\operatorname*{arg\,min}_{\boldsymbol{\lambda}\in\Delta^m} \boldsymbol{\lambda}^T\mathbf{c} \text{ subject to } \mathbf{A}\boldsymbol{\lambda} = \mathbf{0}, \tag{10}$$

where $\mathbf{c}$ *and* $\mathbf{A} \in \mathbb{R}^{m\times m}$ *are uniquely determined by* μ, $\{\mathcal{D}_j\}_{j=1}^m$.

Proof. Let $\{T_j\}_{j=1}^m$ be the optimal transport maps between μ and $\mathcal{D}_j$. Define $\mathbf{A} \in \mathbb{R}^{m\times m}$ by

$$A_{j\ell} = \int_{\mathbb{R}^d} \langle T_j(\mathbf{x}) - \mathbf{x}, T_\ell(\mathbf{x}) - \mathbf{x}\rangle d\mu(\mathbf{x}).$$

Then by Theorem 1 in (Werenski et al., 2022), which holds because **A1**-**A3** hold,

$$\mu = \mathrm{Bary}(\mathcal{D}, \boldsymbol{\lambda}) \iff \boldsymbol{\lambda}^T\mathbf{A}\boldsymbol{\lambda} = 0.$$

Since $\mathbf{A}$ is symmetric and positive semidefinite (it is in fact a Gram matrix), $\boldsymbol{\lambda}^T\mathbf{A}\boldsymbol{\lambda} = 0$ is equivalent to $\mathbf{A}\boldsymbol{\lambda} = \mathbf{0}$. Letting $\mathbf{c}$ be defined as $c_j = W_2^2(\mathcal{D}_j, \mu)$ gives the result. □

We remark that since μ is absolutely continuous and $A_{j\ell} = \int_{\mathbb{R}^d}\langle T_j(\mathbf{x}) - \mathbf{x}, T_\ell(\mathbf{x}) - \mathbf{x}\rangle d\mu(\mathbf{x})$, $\mathbf{A}$ is robust to small perturbations in the optimal transport maps. Under certain conditions, $\|T_{\mu\to\nu} - T_{\mu\to\tilde{\nu}}\|^2_{L^2(\mathbb{R}^d,\mu)} \leq CW_2^q(\nu,\tilde{\nu})$ for constants C, q that depends only on μ and the smoothness and decay properties of the measures $\nu, \tilde{\nu}$ (e.g., dimension of support, largest bounded moment); see (Delalande & Merigot, 2021) for a comprehensive survey of results of this type. In particular, under the assumptions that results of this type can be invoked, the matrix $\mathbf{A}$ is stable under small deformations of $\mathcal{D}$ for μ fixed. Let $\mathbf{A}$ and $\mathbf{A}'$ denote the original $\mathbf{A}$ and perturbed $\mathbf{A}'$ respectively. Assuming that $\mu = \mathrm{Bary}(\mathcal{D}', \boldsymbol{\lambda})$, where $\mathcal{D}'$ denotes perturbed $\mathcal{D}$, we can consider the following program:

$$\operatorname*{arg\,min}_{\mathbf{z}\in\Delta^m} \mathbf{z}^T\mathbf{c} \text{ subject to } \mathbf{A}'\mathbf{z} = \mathbf{0}. \tag{11}$$

To relate the solution of the above program to the solution of (10), we can use linear programming stability results in (Robinson, 1977).

B. Proof of Proposition 3.2

Proof. For clarity, let $d_i = W_2^2(\mathcal{D}_i, \mu)$ so that by assumption $d_1 \leq d_2 \leq \ldots \leq d_m$. If $\boldsymbol{\lambda}^*$ is an optimal solution then noting $b_j = 0$ for $j \geq k+1$, we have that:

$$\sum_{j=1}^{m} \lambda_j^* d_j \leq \sum_{j=1}^{m} b_j d_j = \sum_{j=1}^{k} b_j d_j.$$

Isolating λ_i^* from the left hand side above yields

$$\begin{aligned}
\lambda_i^* d_i &\leq \sum_{j=1}^{k} b_j d_j - \sum_{j \neq i} \lambda_j^* d_j \\
&\leq \sum_{j=1}^{k} b_j d_k - \sum_{j \neq i} \lambda_j^* d_1 \\
&\leq d_k - \sum_{j \neq i} \lambda_j^* d_1 \\
&= d_k - (1 - \lambda_i^*) d_1
\end{aligned}$$

The result follows by solving for λ_i^*. □

C. Proof of Theorem 3.4

Proof. Rearranging, we aim to show

$$0 \leq (t-s) W_2^2(\mu, \mu_t) + s W_2^2(\mu_t, \tilde{\nu}) - t W_2^2(\mu_t, \nu).$$

Noting that McCann interpolants are in fact constant-speed geodesics in Wasserstein space (Ambrosio et al., 2005), we have that

$$t = \frac{W_2(\mu, \mu_t)}{W_2(\mu, \nu)}, \quad s = \frac{W_2(\mu, \mu_t)}{W_2(\mu, \tilde{\nu})}$$

and

$$\begin{aligned}
W_2(\mu, \tilde{\nu}) &= W_2(\mu, \nu) + W_2(\nu, \tilde{\nu}), \\
W_2(\mu_t, \tilde{\nu}) &= W_2(\mu_t, \nu) + W_2(\nu, \tilde{\nu}).
\end{aligned}$$

In particular, $s < t$ and so it suffices to show

$$\begin{aligned}
& t W_2^2(\mu_t, \nu) \leq s W_2^2(\mu_t, \tilde{\nu}) \\
\iff & \frac{W_2(\mu, \mu_t)}{W_2(\mu, \nu)} W_2^2(\mu_t, \nu) \leq \frac{W_2(\mu, \mu_t)}{W_2(\mu, \tilde{\nu})} W_2^2(\mu_t, \tilde{\nu}) \\
\iff & \frac{W_2^2(\mu_t, \nu)}{W_2(\mu, \nu)} \leq \frac{W_2^2(\mu_t, \tilde{\nu})}{W_2(\mu, \tilde{\nu})} \\
\iff & \frac{W_2^2(\mu_t, \nu)}{W_2(\mu, \nu)} \leq \frac{(W_2(\mu_t, \nu) + W_2(\nu, \tilde{\nu}))^2}{W_2(\mu, \nu) + W_2(\nu, \tilde{\nu})} \\
\iff & W_2^2(\mu_t, \nu)(W_2(\mu, \nu) + W_2(\nu, \tilde{\nu})) \leq W_2(\mu, \nu)(W_2(\mu_t, \nu) + W_2(\nu, \tilde{\nu}))^2 \\
\iff & W_2^2(\mu_t, \nu)(W_2(\mu, \nu) + W_2(\nu, \tilde{\nu})) \leq W_2(\mu, \nu)(W_2^2(\mu_t, \nu) + 2W_2(\mu_t, \nu) W_2(\nu, \tilde{\nu}) + W_2^2(\nu, \tilde{\nu})) \\
\iff & W_2^2(\mu_t, \nu) W_2(\nu, \tilde{\nu}) \leq W_2(\mu, \nu)(2W_2(\mu_t, \nu) W_2(\nu, \tilde{\nu}) + W_2^2(\nu, \tilde{\nu})).
\end{aligned}$$

If $\tilde{\nu} = \nu$, the result follows trivially. So, assume $W_2(\nu, \tilde{\nu}) > 0$. Then the above reduces to

$$W_2^2(\mu_t, \nu) \leq W_2(\mu, \nu)(2W_2(\mu_t, \nu) + W_2(\nu, \tilde{\nu})).$$

The result follows by noting that $W_2(\mu_t, \nu) \leq W_2(\mu, \nu)$ and that $W_2(\nu, \tilde{\nu}) \geq 0$. □

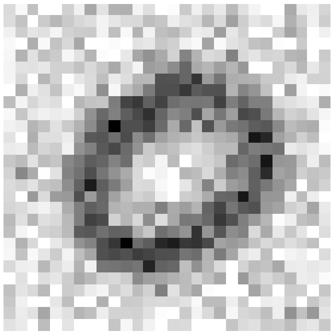


Figure 4. A noisy MNIST digit

C.1. Effect of Entropy on Theoretical Results

In general, and under suitable regularity conditions, entropic quantities converge to their unregularized counterparts as $\epsilon \to 0^+$ where ε is the entropic regularization parameter. (Cominetti & Martín, 1994) and (Peyré et al., 2019) discuss convergence of entropic distances and (Pooladian & Niles-Weed, 2021) and (Nutz & Wiesel, 2022) discuss the convergence of entropic transport maps. Additionally, the work in (Werenski et al., 2022) establishes that the matrix $\mathbf{A}$ of (8) is entry-wise close to its non-entropic counterpart.

The implication of this to our theoretical results is as follows:

Proposition 3.1: If we consider the matrix $\mathbf{A}$ to be computed via entropic regularization, Proposition 3.1 does not continue to hold since the entropically computed $\mathbf{A}$ will not likely have the same zero eigenvectors of $\mathbf{A}$. However, we note that $\boldsymbol{\lambda}^T\mathbf{A}\boldsymbol{\lambda} = 0$ when $\boldsymbol{\lambda}$ is a true set of barycentric weights. With that, one can relax the optimization problem by considering instead $\rho\boldsymbol{\lambda}^T\mathbf{c} + \boldsymbol{\lambda}^T\mathbf{A}\boldsymbol{\lambda}$. The entrywise closeness of $\mathbf{c}$ and $\mathbf{A}$ suggests that for small ε there might be stability in the computed solution of the relaxed program i.e., solutions obtained for the original and entropic variations may be close. Precisely quantifying this statement will depend on applying results on the stability of a quadratic program (e.g., (Phu & Yen, 2001) and (Best & Chakravarti, 1990)) to our setting.

Proposition 3.2: Similarly if one instead considers the relaxed problem discussed above, then we would expect the weights obtained from the entropic problem to be close and thus have a similar concentration.

Theorem 3.4 and Corollary 3.5: These results strongly depend on the true Wasserstein distance due to the discussion of geodesics. Empirically, we observe a similar result should be true as evidenced by Figure 2 where the generating measures are able to be approximately identified.

D. Experimental Details

For each of the experiments we report specific parameters used to generate the results. We also report the timings based on our code.

D.1. MNIST

Noise Model: We add scaled pixel-wise noise to each data distribution. The noise is $\sim \mathcal{N}(0,1)$ and scaled by 0.0005. After adding the noise we clip the entries of the image that became negative to a small positive number, before renormalizing the image to lie on the probability simplex.

Specific parameter choices:

- Atom initialization: Wasserstein K-Means++.

- Weight initialization: Uniform samples from the simplex.
- Optimizer: Adam (Kingma & Ba, 2014) with default parameters except for learning rate as 0.25.
- We use $L = 250$ iterations for reasonable convergence; learned atoms generally show no visual change for about 50 iterations and loss had flat-lined for a similar amount of iterations.
- Entropic regularization parameter set to 0.003 for both Wasserstein distance and barycenter computations.
- We use $L_s = 50$ Sinkhorn iterations for both Wasserstein distance and barycenter computations.
- Entropic transport computations were accelerated with Convolutional Wasserstein (Solomon et al., 2015).

This experiment took 8 hours total using all cpu cores of an Apple M1 chip (no gpu). This code has not been optimized. We show more examples in Figure 5.

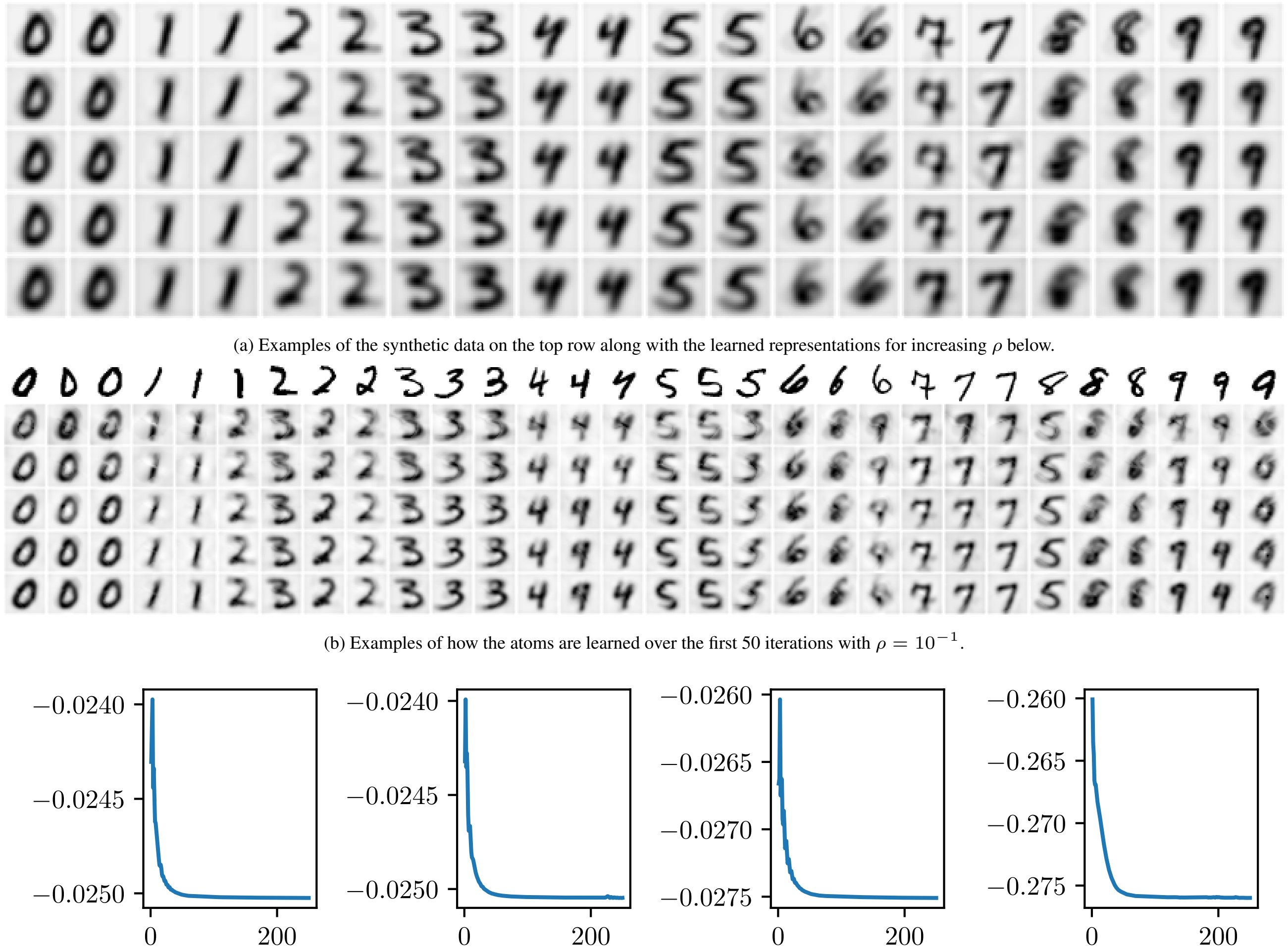
D.2. NLP

In Figure 6 we plot a different view of the data from Figure 3 to clearly show the effect of the regularization parameter per experiment. We plot the 1 standard deviation bars for the NLP experiments in Figure 7.

Specific parameter choices:

- Atom initialization: Wasserstein K-Means++.
- Weight initialization: Each weight is initialized as a vector with uniform random samples and then normalized to lie on the simplex. This differs from uniform samples from the simplex, but in practice there were no performance differences.
- Optimizer: Adam with default parameters except for learning rate as 0.25.
- We use $L = 300$ iterations for reasonable convergence; loss had generally flat-lined for about 50 iterations.
- Entropic regularization set to 0.1 for both Wasserstein distance and barycenter computations.
- We use $L_s = 25$ Sinkhorn iterations for both Wasserstein distance and barycenter computations.
- Since the dictionary atoms must have fixed support, we fix the support as the union of words present in the training documents for each dictionary.

Running one of the thirty trials of the experiment for all number of the references took at most 2 days on an HPC node using 2 cpu cores and one Nvidia gpu of type T4, RTX 6000, V100, or P100 (depending on node availability). This code has not been optimized.

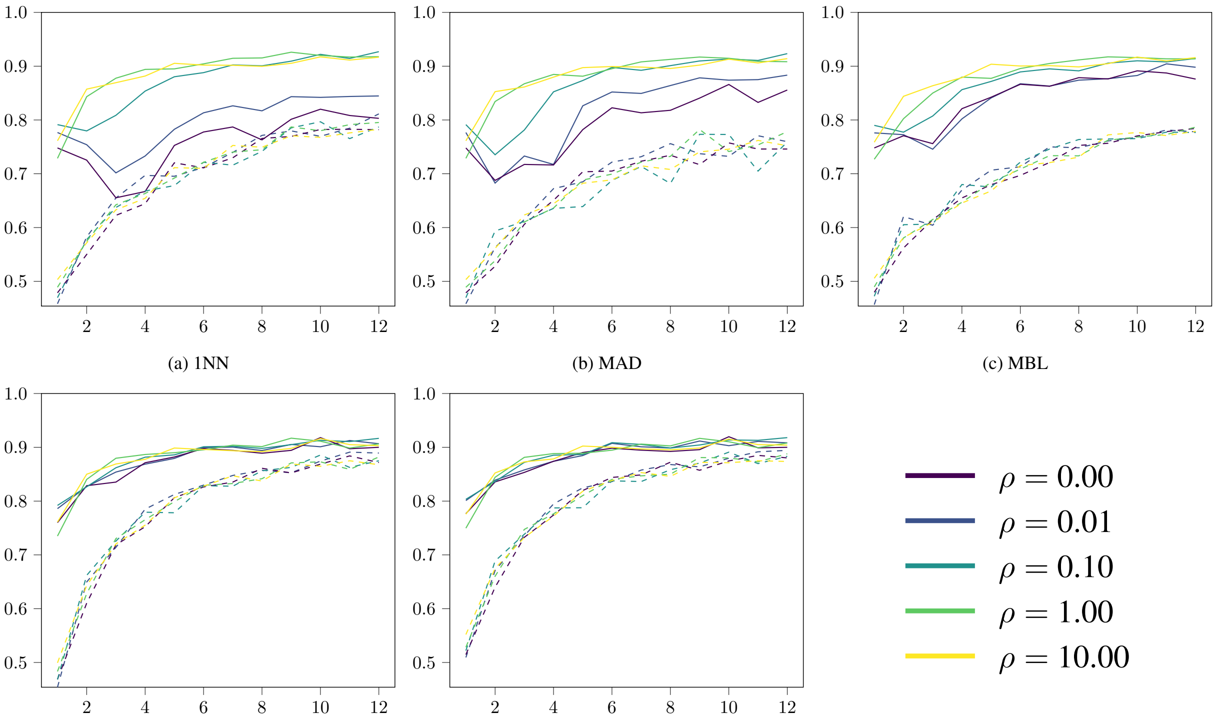


(a) Examples of the synthetic data on the top row along with the learned representations for increasing ρ below.

(b) Examples of how the atoms are learned over the first 50 iterations with $\rho = 10^{-1}$.

(c) The corresponding training loss plots, displaying (6), associated with the above runs in (a) and (b). Note that the negative objective values are a result of our entropic estimate being a lower bound on the true value.

Figure 5. (a) Learned Representations: In the top row we include two example synthetic digits per class. In each subsequent row we show the learned reconstruction of the data point for $\rho \in \{0, 10^{-3}, 10^{-1}, 10^{1}\}$ (top-to-bottom). (b) Learning Dynamics: On the top row we show the generating atoms and each subsequent row shows how the subsequently matched atoms evolve over the learning process at 10, 20, 30, 40, and 50 iterations. Atoms here are not the same as in the main text, but representative of the learning dynamics. (c) The corresponding training loss graphs for $\rho \in \{0, 10^{-3}, 10^{-1}, 10^{1}\}$ (left-to-right). *Note:* the sample of digits from MNIST used as generating atoms here is not the same as shown in the main text.



(a) 1NN

(b) MAD

(c) MBL

(d) MBL-QP

(e) MC

Figure 6. Accuracy vs. number of reference documents. Solid lines denote reference documents being learned dictionary atoms while dashed lines denote the baseline of reference documents being random documents from each class. Here, we group by method and show the impact of different levels of regularization and learning versus random atoms. Except for small ρ in 1NN, using learned reference documents significantly outperforms the use randomly sampled documents across all number of references. Note that $\rho = 0$ is unregularized WDL (Schmitz et al., 2018).

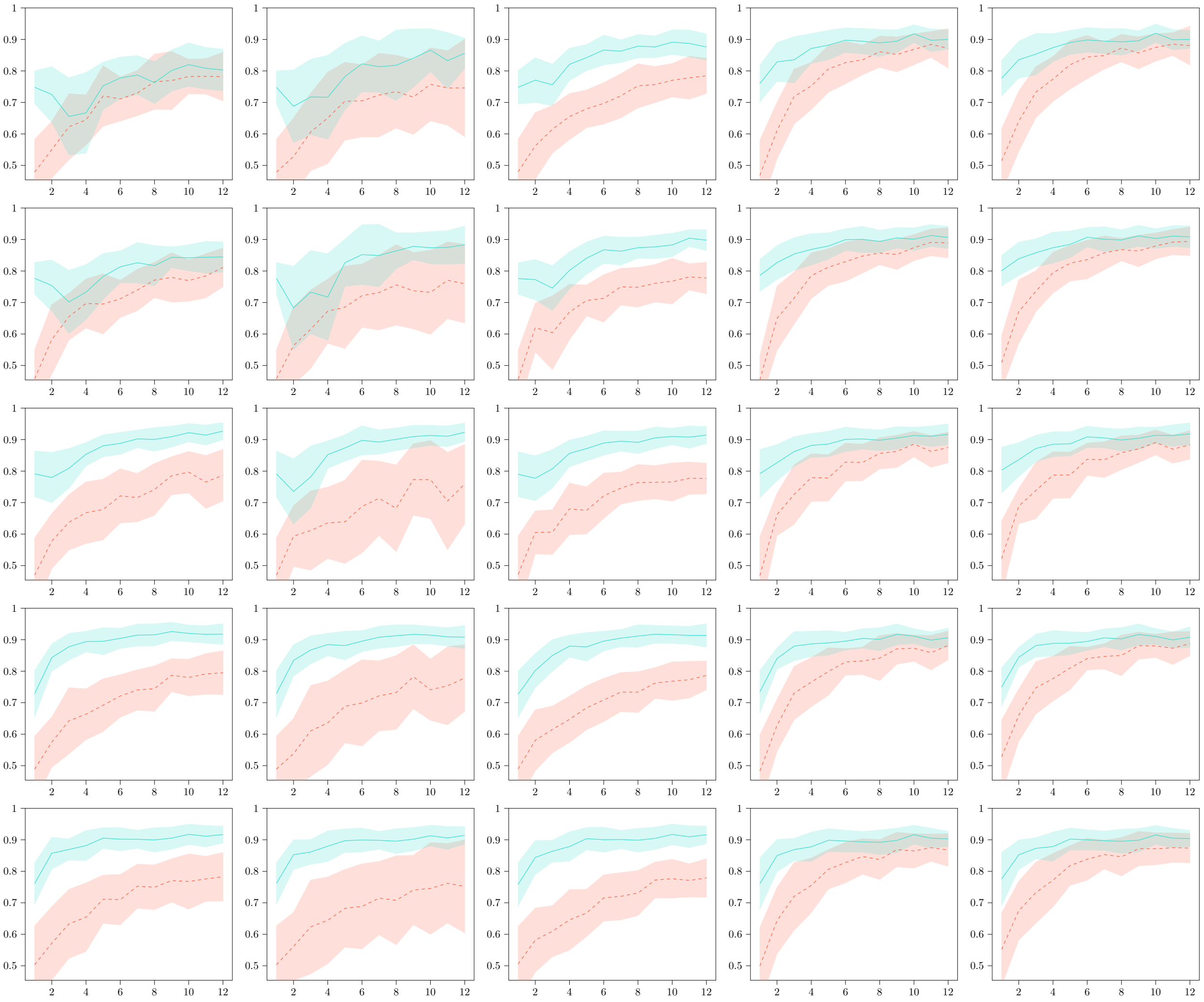


Figure 7. Fraction of test documents classified correctly vs number of representative documents. Solid turquoise line corresponds and dashed tomato line correspond to learned and random reference documents, respectively. Top to bottom: each row corresponds to ρ increasing in $\{0.0, 0.01, 0.1, 1.0, 10.0\}$. Left to right: each column corresponds to the methods $\{$1NN, MAD, MBL, MBL-QP, MC$\}$. We observe that learned documents outperform random documents in every experiment for all levels of ρ. The smaller variance of the learned documents is explained by the fact that the learned documents were trained with more documents than were used.