

---

# Measure Estimation in the Barycentric Coding Model

---

Matthew Werenski<sup>1</sup> Ruijie Jiang<sup>2</sup> Abiy Tasissa<sup>3</sup> Shuchin Aeron<sup>2</sup> James M. Murphy<sup>3</sup>

## Abstract

This paper considers the problem of measure estimation under the *barycentric coding model* (BCM), in which an unknown measure is assumed to belong to the set of Wasserstein-2 barycenters of a finite set of known measures. Estimating a measure under this model is equivalent to estimating the unknown barycentric coordinates. We provide novel geometrical, statistical, and computational insights for measure estimation under the BCM, consisting of three main results. Our first main result leverages the Riemannian geometry of Wasserstein-2 space to provide a *procedure for recovering the barycentric coordinates as the solution to a quadratic optimization problem* assuming access to the true reference measures. The essential geometric insight is that the parameters of this quadratic problem are determined by inner products between the optimal displacement maps from the given measure to the reference measures defining the BCM. Our second main result then establishes an *algorithm for solving for the coordinates in the BCM when all the measures are observed empirically via i.i.d. samples*. We prove precise *rates of convergence* for this algorithm—determined by the smoothness of the underlying measures and their dimensionality—thereby guaranteeing its statistical consistency. Finally, we demonstrate the utility of the BCM and associated estimation procedures in three application areas: (i) *covariance estimation* for Gaussian measures; (ii) *image processing*; and (iii) *natural language processing*.

## 1. Introduction

A number of recent machine learning applications including computer vision (Schmitz et al., 2018; Bonneel et al., 2016), domain adaptation (Montesuma & Mboula, 2021; Redko et al., 2019), natural language processing (NLP) (Singh et al., 2020; Colombo et al., 2021; Xu et al., 2018), and unsupervised segmentation of multivariate time series data (Cheng et al., 2021) have shown the utility of representing and modeling high-dimensional data as probability distributions. The essential insight in these applications is to utilize the Riemannian geometry of the space of probability distributions induced by the Wasserstein-2 metric and the Wasserstein-2 barycenters as alternatives to the Euclidean distance and linear combinations, respectively.

In this context, the methods that use Wasserstein barycenters for data modeling and inference involve solving two core problems which we term the *synthesis problem* and *analysis problem*. To be precise, let  $\mathcal{P}_{2,ac}(\mathbb{R}^d)$  denote the space of absolutely continuous distributions on  $\mathbb{R}^d$  with finite second moment. For  $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ ,

$$W_2^2(\mu, \nu) \triangleq \min_{T \# \mu = \nu} \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 d\mu(x) \quad (1)$$

defines the *Wasserstein-2 metric* where the minimum is over all measurable maps  $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$  and the *pushforward*  $T \# \mu = \nu$  is such that for all Borel sets  $B$  we have  $\nu[B] = \mu[T^{-1}(B)]$  (Villani, 2003; Santambrogio, 2015). Under the assumption  $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ , unique minimizing  $T$  are known to exist and the resulting  $W_2$  distance is finite (Villani, 2003). We refer to the space  $\mathcal{P}_{2,ac}(\mathbb{R}^d)$  equipped with the metric  $W_2$  as the *Wasserstein-2 space*, denoted  $\mathcal{W}_2$ .

Let  $\Delta^p = \{\lambda = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p : \lambda_i \geq 0, \sum_{i=1}^p \lambda_i = 1\}$  and let  $\{\mu_i\}_{i=1}^p$  be known *reference measures*. The  $W_2$  *barycenter* for coordinates  $\lambda \in \Delta^p$  with respect to  $\{\mu_i\}_{i=1}^p$  is defined as

$$\nu_\lambda \triangleq \arg \min_{\nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)} \frac{1}{2} \sum_{i=1}^p \lambda_i W_2^2(\nu, \mu_i). \quad (2)$$

The measure  $\nu_\lambda$  is an advection of  $\{\mu_i\}_{i=1}^p$  and blends their geometric features; it may be thought of as a “displacement interpolation” (McCann, 1997; Villani, 2003). In contrast, the linear mixture  $\sum_{i=1}^p \lambda_i \mu_i$  simply sets  $\lambda_i$  to be the

---

<sup>1</sup>Department of Computer Science, Tufts University

<sup>2</sup>Department of Electrical and Computer Engineering, Tufts University <sup>3</sup>Department of Mathematics, Tufts University . Correspondence to: Matthew Werenski <matthew.werenski@tufts.edu>.

rate we draw samples from  $\mu_i$ . Figure 1 demonstrates the important difference between linear mixture models and Wasserstein barycenters on Gaussian distributions: the linear mixture is a standard Gaussian mixture model (Murphy, 2012), while the Wasserstein barycenter is itself Gaussian. The important takeaway is that Wasserstein barycenters offer a more geometrically meaningful interpolation between measures compared to mixtures, and as indicated above have been successfully leveraged in a number of applications.

Now, let  $\text{Bary}(\{\mu_i\}_{i=1}^p) \triangleq \{\nu_\lambda : \lambda \in \Delta^p\}$  denote the set of all of possible barycenters of  $\{\mu_i\}_{i=1}^p$ . The *synthesis* problem (2) seeks to combine the reference measures  $\{\mu_i\}_{i=1}^p$  in a notion of weighted average, with relative contributions determined by  $\lambda$ . This problem is well-studied (McCann, 1997; Agueh & Carlier, 2011; Kroshnin et al., 2019) and known to have a unique solution under mild assumptions on the  $\{\mu_i\}_{i=1}^p$ . Moreover, there exist fast algorithms for computing or approximating  $\nu_\lambda$  when samples from the  $\{\mu_i\}_{i=1}^p$  are available (Rabin et al., 2011; Cuturi & Doucet, 2014; Bonneel et al., 2015; Clatici et al., 2018; Yang et al., 2021), which have spurred the application of Wasserstein barycenters to machine learning problems.

In this paper, our focus is on the less well-studied *analysis* problem: given  $\mu_0$  and reference measures  $\{\mu_i\}_{i=1}^p$ , solve

$$\arg \min_{\lambda \in \Delta^p} W_2^2(\mu_0, \nu_\lambda). \quad (3)$$

Instead of *combining* the measures  $\{\mu_i\}_{i=1}^p$  according to the weights  $\lambda \in \Delta^p$  as in (2), the analysis problem (3) *decomposes* a given measure  $\mu_0$  into its optimal representation in the set  $\text{Bary}(\{\mu_i\}_{i=1}^p)$  as parameterized by the learned barycentric coordinates  $\lambda \in \Delta^p$ . If the measure  $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$ , then  $\min_{\lambda \in \Delta^p} W_2^2(\mu_0, \nu_\lambda) = 0$  and solving (3) is equivalent to learning the barycentric coordinates  $\lambda \in \Delta^p$  such that  $\nu_\lambda = \mu_0$ . We refer to this general model of parameterizing measures  $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$  through the associated coordinates  $\lambda \in \Delta^p$  as the *barycentric coding model* (BCM).

**Main Contributions:** This paper addresses the analysis problem (3) as a measure estimation problem under the BCM. For this general problem, we make the following geometrical, statistical, and computational contributions:

- (1) When all the measures are directly observed, Theorem 1 provides a *geometric* characterization of the solution to (3) as corresponding to a solution to a quadratic program. Our analysis leverages the Riemannian geometry of  $\mathcal{W}_2$  space and establishes that the quadratic program is determined by the angles between the optimal displacement maps between the measures. When specialized to the case of Gaussian measures, Corollary 1

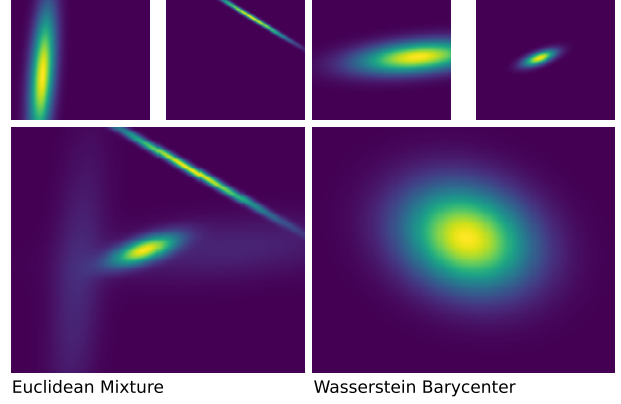


Figure 1: *Top four*: reference Gaussians. *Bottom left*: linear mixture with  $\lambda = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ . *Bottom right*: Wasserstein barycenter with same  $\lambda$ . (Intensities adjusted to show detail.)

provides a closed-form characterization of the solution to the analysis problem.

- (2) When the measures are indirectly observed via i.i.d. samples, we provide an algorithm for estimating the barycentric coordinates  $\lambda$  of a given measure with respect to the reference measures. Theorem 3 gives precise rates of convergence of the angles between the optimal displacement map by a computationally tractable estimator that exploits entropic regularization, thereby establishing the statistical consistency of our approach in Corollary 2.
- (3) We showcase the utility of the BCM and the proposed algorithm on several data analysis problems: covariance estimation for Gaussian measures; image inpainting and denoising; and document classification from limited labeled data. We show that the BCM provides a simple yet effective approach that outperforms competing methods in the regimes considered.<sup>1</sup>

## 1.1. Related Work

The synthesis problem (2) is well-studied. (Agueh & Carlier, 2011) propose  $W_2$  barycenters for a finite set of reference measures  $\{\mu_i\}_{i=1}^p$  and prove their existence and uniqueness in a very general setting. (Le Gouic & Loubes, 2017) extend this to barycenters constructed from infinitely many reference measures. Statistical rates of estimation in the setting of i.i.d. samples from the measures  $\{\mu_i\}_{i=1}^p$  are established in (Kroshnin et al., 2019). Building on fundamental advances in estimating transport maps between measures using entropic regularization (Cuturi, 2013), fast estimation methods for solving the synthesis problem have been developed for entropically regularized barycenters (Cuturi & Doucet, 2014). Approaches to computing a barycenter based on al-

<sup>1</sup>Code to reproduce results is available at <https://github.com/MattWerenski/BCM>

ternating direction of multipliers on the dual problem (Yang et al., 2021) and parallelization schemes (Claici et al., 2018) have also been considered.

The literature on the analysis problem (3) is much sparser. To the best of our knowledge, our geometrical characterization of solutions to the analysis problem (Theorem 1) and sample complexity results for estimating the optimal coordinates  $\lambda$  (Theorem 3, Corollary 2) are the first of their kind. Note that our approach extends (Bonneel et al., 2016; Schmitz et al., 2018)—which consider measure estimation under the BCM but only for measures with finite support—in three fundamental ways. First, we enable the application of the BCM to absolutely continuous measures. Second, we prove precise theoretical characterizations of the solutions to (3). Third, we avoid the expensive procedure of differentiating through an iterative algorithm as required in (Bonneel et al., 2016) and instead provide the direct Algorithm 1 that enjoys statistical consistency (Corollary 2). At the level of technical analysis, our Theorem 1 relies on an alternative characterization of  $W_2$  barycenters that leverages known conditions under which Karcher means are guaranteed to be  $W_2$  barycenters (Panaretos & Zemel, 2020).

For the special case when all the measures are zero-mean Gaussians, the *synthesis* problem is again well-studied, with a fixed-point algorithm for computing the barycenter given in (Álvarez-Esteban et al., 2016) that is shown to have a linear convergence rate in (Chewi et al., 2020; Altschuler et al., 2021). The corresponding *analysis* problem for the specific case of Gaussian measures has only been recently studied in (Musolas et al., 2021) but limited to the case when  $p = 2$ . When specialized to this case, our results extend the model considered in (Musolas et al., 2021) to the  $p \geq 3$  case and characterize solutions to barycenter parameterized covariance estimation.

In Section 4, we use the proposed BCM framework to model and address several problems in image processing and natural language processing. In this context our work is related to a long lineage of methods in signal processing and machine learning that use coefficients in a basis or a dictionary for data modeling and processing (Mallat, 1999; Donoho, 2006; Tošić & Frossard, 2011). Our approach of representing a general measure in terms of its barycentric coordinates with respect to a fixed set of reference measures is a novel approach in this tradition, and our results in Section 4 suggest its utility for data modeled as probability measures.

## 2. Convex Optimization for the Analysis Problem

To solve the analysis problem (3), we propose to solve a convex optimization problem based on an alternative characterization of  $W_2$ -barycenters as minimizers of a certain

functional. We leverage this alternative characterization to show in Theorem 1 that, under mild assumptions on  $\{\mu_i\}_{i=1}^p$ , one can use the optimal value of a convex program to both check if  $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$  and if so, recover its barycentric coordinate. Before proceeding, we must introduce some background on the Riemannian geometry of the  $\mathcal{W}_2$  space. The geodesic curve from  $\nu_0$  to  $\nu_1$  in  $\mathcal{W}_2$  is given by  $[\nu_t]_{t=0}^1$  where  $\nu_t = (tT + (1-t)\text{Id})\#\nu_0$  where  $T$  is the optimal transport map from  $\nu_0$  to  $\nu_1$  in (1) (Santambrogio, 2015). These facts together with Brenier’s Theorem (see Supplementary Material Section 1) motivate the definition of the tangent space at  $\nu$ :

$$T_\nu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \triangleq \overline{\{\beta(\nabla\varphi - \text{Id}) : \beta > 0, \varphi \in C_c^\infty(\mathbb{R}^d), \varphi \text{ convex}\}},$$

where the closure is in  $L^2(\nu)$  (Ambrosio et al., 2005). For  $u, v \in T_\nu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ , the inner product is defined by  $\langle u, v \rangle_\nu \triangleq \int_{\mathbb{R}^d} \langle u(x), v(x) \rangle d\nu(x)$ , which also gives the standard norm of  $u \in T_\nu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ ,  $\|u\|_\nu \triangleq \sqrt{\langle u, u \rangle_\nu}$ .

### 2.1. Alternative Characterization of Barycenters

Given the reference measures  $\{\mu_i\}_{i=1}^p$  and barycentric coordinates  $\lambda \in \Delta^p$ , the *variance functional* is  $G_\lambda : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$  defined by

$$G_\lambda(\nu) \triangleq \sum_{i=1}^p \frac{\lambda_i}{2} W_2^2(\nu, \mu_i),$$

so that the barycenter for  $\{\mu_i\}_{i=1}^p$  and  $\lambda \in \Delta^p$  is

$$\nu_\lambda = \arg \min_{\nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)} G_\lambda(\nu).$$

In many unconstrained optimization problems, the most direct way of checking if a point is optimal is to use first-order optimality conditions: compute the gradient at the point and check if it is zero. Without further assumptions this is a necessary, but not sufficient, condition for optimality. Applying this to a general variance functional, one arrives at the definition of a *Karcher mean* (Karcher, 1977).

**Definition 1.** Let  $(X, \rho)$  be a metric space furnished with a tangent space at every  $x \in X$ . For a collection of points  $\{x_i\}_{i=1}^p \subset X$  and coordinates  $\lambda \in \Delta^p$ , a point  $y \in X$  is said to be a *Karcher mean* of the set for  $\lambda$  if  $G_\lambda = \sum_{i=1}^p \lambda_i \rho^2(\cdot, x_i)$  is differentiable at  $y$  and  $\|\nabla G_\lambda(y)\|_y = 0$  where  $\nabla G_\lambda(y)$  is understood as an element of the tangent space at  $y$ .

Any barycenter  $y$  is a global minimizer of  $G_\lambda$ . Using first order optimality conditions, this implies  $\|\nabla G_\lambda(y)\|_y = 0$ , and therefore  $y$  must be a Karcher mean. Special care must be made to ensure that being a Karcher mean is sufficient to

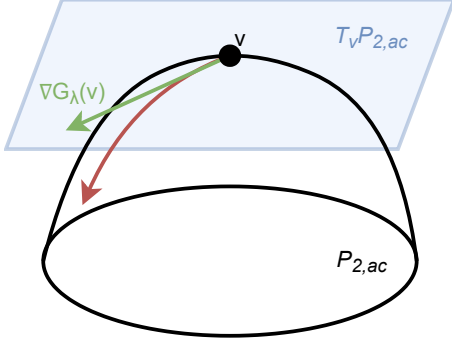


Figure 2: Visualization of the tangent space  $T_\nu P_{2,ac}(\mathbb{R}^d)$  and  $\nabla G_\lambda$ . The red arc is the path along which  $G_\lambda$  is maximized, starting at  $\nu$  and the green arrow points in that direction in the tangent space at  $\nu$ .

be a barycenter. When it is sufficient, it immediately gives an alternative test for  $y$  to be a barycenter for  $\lambda$ : simply check if  $\|\nabla G_\lambda(y)\|_y = 0$ .

An example of a Karcher mean which is not a barycenter is given in the Supplementary Material Section 7. To ensure that being a Karcher mean is sufficient to be a barycenter we require the following assumptions:

**A1:** The measures  $\{\mu_i\}_{i=0}^p$  are absolutely continuous and supported on either all of  $\mathbb{R}^d$  or a bounded open convex subset. Call this shared support set  $\Omega$ .

**A2:** The measures  $\{\mu_i\}_{i=0}^p$  have respective densities  $\{g_i\}_{i=0}^p$  which are bounded above and  $g_1, \dots, g_p$  are strictly positive on  $\Omega$ .

**A3:** If  $\Omega = \mathbb{R}^d$  then  $\{g_i\}_{i=0}^p$  are locally Hölder continuous. Otherwise  $\{g_i\}_{i=0}^p$  are bounded away from zero on  $\Omega$ .

Under these assumptions if  $\mu_0$  is a Karcher mean with coordinates  $\lambda$ , then  $\mu_0 = \nu_\lambda$  (Panaretos & Zemel, 2020); see the Supplementary Material Section 1 for a precise statement.

This characterization of barycenters can be much easier to work with and the assumptions required are mild in practice. To accompany the notion of a Karcher mean, the *Fréchet derivative* (Fréchet, 1948) of the variance functional  $G_\lambda$  is given by

$$\nabla G_\lambda(\nu) = - \sum_{i=1}^p \lambda_i (T_i - \text{Id}), \quad (4)$$

where  $T_i$  is the optimal transport map from  $\nu$  to  $\mu_i$  as in (1) (Ambrosio et al., 2005; Panaretos & Zemel, 2020).

We now have all the tools to begin building up to Theorem 1. The first step is to derive a formula for  $\|\nabla G_\lambda(\mu_0)\|_{\mu_0}^2$  in terms of  $\lambda$  and show its convexity.

**Proposition 1.** Let  $\{\mu_i\}_{i=0}^p \subset \mathcal{P}_{2,ac}(\mathbb{R}^d)$ . Then  $\|\nabla G_\lambda(\mu_0)\|_{\mu_0}^2 = \lambda^T A \lambda$  where  $A \in \mathbb{R}^{p \times p}$  is given by

$$A_{ij} = \int_{\mathbb{R}^d} \langle T_i(x) - \text{Id}(x), T_j(x) - \text{Id}(x) \rangle d\mu_0(x) \quad (5)$$

and  $T_i$  is the optimal transport map from  $\mu_0$  to  $\mu_i$  as in (1). Furthermore, this is a convex function in  $\lambda$ .

The proof is given in the Supplementary Material Section 2.1. The expression  $T_i(x) - \text{Id}(x)$  is the displacement of the point  $x$  when being transported from  $\mu_0$  to  $\mu_i$ . Under this interpretation,  $\langle T_i(x) - \text{Id}, T_j(x) - \text{Id} \rangle$  can be thought of as the angle between the displacement associated to the optimal transport map between  $\mu_0$  to  $\mu_i$  with that of  $\mu_0$  to  $\mu_j$ . Integrating this quantity with respect to  $\mu_0$  therefore quantifies the average angle between displacements and if the optimal displacement maps move in similar directions.

We now use this functional form as part of a convex program and show that it can be used to check if a  $\mu_0$  is a Karcher mean. Our first main result ties Propositions 1 to the conditions for equivalence between barycenters and Karcher means.

**Theorem 1.** Let  $\{\mu_i\}_{i=0}^p$  satisfy **A1-A3**. Then  $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$  if and only if

$$\min_{\lambda \in \Delta^p} \lambda^T A \lambda = 0$$

where  $A \in \mathbb{R}^{p \times p}$  is given by (5). Furthermore, if the minimum value is 0 and  $\lambda_*$  is an optimal argument, then  $\mu_0 = \nu_{\lambda_*}$ .

The proof is given in the Supplementary Material Section 2.3. Theorem 1 gives a geometric characterization of the solution to (3) which relies on the evaluation of inner-products in the tangent space  $T_{\mu_0} \mathcal{P}_{2,ac}(\mathbb{R}^d)$  (noting that  $A_{ij} = \langle T_i - \text{Id}, T_j - \text{Id} \rangle_{\mu_0}$ ) and solving a constrained quadratic program. Furthermore, this result suggests a method for finding an approximator of the  $\mu_0$  in the set of barycenters: simply minimize the objective and use  $\lambda_*$ , even if the minimum value is not zero.

Note that if  $A$  has rank less than  $p-1$  there may be multiple minimizers which correspond to redundancies in the set  $\text{Bary}(\{\mu_i\}_{i=1}^p)$ ; see Supplementary Material Section 6.

## 2.2. Theorem 1 For Gaussians

In the Gaussian case there are formulas for both the optimal transport maps and the integration involved in  $A_{ij}$ . Let  $\mathbb{S}_{++}^d$  denote the set of symmetric positive definite  $d \times d$  matrices. Let  $\mu_i = \mathcal{N}(0, S_i)$  with  $S_i \in \mathbb{S}_{++}^d$  for  $i = 0, \dots, p$ . Then the optimal transport map from  $\mu_0$  to  $\mu_i$  is given by  $T_i(x) = C_i x$  where

$$C_i = S_0^{-1/2} \left( S_0^{1/2} S_i S_0^{1/2} \right)^{1/2} S_0^{-1/2} \quad (6)$$

(Agueh & Carlier, 2011). Furthermore, **A1-A3** are satisfied for non-degenerate Gaussians. Combining Theorem 1 with (6) gives the following.

**Corollary 1.** *For  $i = 1, \dots, p$ , let  $\mu_i = \mathcal{N}(0, S_i)$  with  $S_i \in \mathbb{S}_{++}^d$ . Then  $\mu_0$  is a barycenter if and only if  $\mu_0 = \mathcal{N}(0, S_0)$  for some  $S_0 \in \mathbb{S}_{++}^d$ , and the convex program of Theorem 1 has minimum value 0, where the matrix  $A$  is given by  $A_{ij} = \text{Tr}((C_i - I)(C_j - I)S_0)$ . Furthermore, if the minimum value is zero and  $\lambda_*$  is an optimal argument, then  $\mu_0 = \nu_{\lambda_*}$ .*

The proof is deferred to the Supplementary Material Section 2.3. We note that in the Gaussian case, one can use  $\|\nabla G_\lambda(\mu_0)\|_{\mu_0}$  to upper bound  $W_2(\mu_0, \nu_\lambda)$  (Chewi et al., 2020; Altschuler et al., 2021). This further justifies the choice of optimizing  $\|\nabla G_\lambda(\mu_0)\|_{\mu_0}^2$ , since it can be seen as a convex, sharp upper bound on the distance to the set of barycenters.

### 2.3. Projections of Compatible Measures

If one imposes extra structure on the set  $\{\mu_i\}_{i=0}^p$ , then it is possible to substantially strengthen Theorem 1.

**Definition 2.** A family of measures  $\{\mu_i\}_{i=0}^p$  is said to be *compatible* if for any  $i, j, k \in \{0, 1, \dots, p\}$  the optimal transport map  $T_i^k$  from  $\mu_i$  to  $\mu_k$  can be expressed as

$$T_i^k = T_j^k \circ T_i^j$$

where  $T_i^j, T_j^k$  are the optimal transport maps from  $\mu_i$  to  $\mu_j$  and  $\mu_j$  to  $\mu_k$  respectively.

Note that whenever these maps exist, the map  $T_j^k \circ T_i^j$  is a valid map from  $\mu_i$  to  $\mu_k$ , but it need not be optimal. Under the condition of compatibility, we have the following result.

**Theorem 2.** *Suppose that  $\{\mu_i\}_{i=0}^p$  are compatible and let  $\lambda_*$  be the minimizer in Theorem 1. Then*

$$\lambda_* = \arg \min_{\lambda \in \Delta^p} W_2^2(\mu_0, \nu_\lambda).$$

The proof is deferred to the Supplementary Material Section 2.4. In other words, when the measures are compatible one can solve (3) and find the *exact projection* of  $\mu_0$  onto  $\text{Bary}(\{\mu_i\}_{i=1}^p)$ .

There are several known conditions under which the family of measures is compatible. These include any set of continuous measures on  $\mathbb{R}$ , any family of Gaussians with simultaneously diagonalizable covariance matrices, families of radial transformations of a base measure, or families with a "common dependence structure"; see (Panaretos & Zemel, 2020) Section 2.3 for further details.

## 3. Finite Sample Analysis and Rates of Convergence

The results of the previous section are all reliant upon having exact knowledge of both the underlying measure  $\mu_0$  as well as the optimal transport maps  $T_i$ . In practice, outside of specific parametric families (e.g., Gaussians), it is rare that one will have either of these available.

Much more common is the setting where the  $\mu_i$  are accessed through i.i.d. sampling. In this setting we need to estimate the maps  $T_i$ , as well as  $\mu_0$  and use these estimates to compute an  $\hat{A}$  whose entries are  $\hat{A}_{ij} = \langle \hat{T}_i - \text{Id}, \hat{T}_j - \text{Id} \rangle_{\hat{\mu}_0}$ , where  $\hat{T}_i$  and  $\hat{\mu}_0$  are our estimates of the transport maps and  $\mu_0$  respectively. The quality of the estimate  $\hat{A}$  will depend on the approximations used above, and ideally these would be accompanied by performance guarantees.

### 3.1. The Entropic Map Estimate

Let  $X_1, \dots, X_n \sim \mu_0$  and  $Y_1, \dots, Y_n \sim \mu_1$  be i.i.d. samples. In (Pooladian & Niles-Weed, 2021) the authors propose using the *entropic map* to estimate the transport maps. For  $\epsilon > 0$ , this map is defined by

$$\hat{T}_\epsilon(x) \triangleq \frac{\frac{1}{n} \sum_{i=1}^n Y_i \exp\left(\frac{1}{\epsilon}(g_\epsilon(Y_i) - \frac{1}{2}\|x - Y_i\|_2^2)\right)}{\frac{1}{n} \sum_{i=1}^n \exp\left(\frac{1}{\epsilon}(g_\epsilon(Y_i) - \frac{1}{2}\|x - Y_i\|_2^2)\right)} \quad (7)$$

where  $g_\epsilon$  is defined to be the optimal variable in the dual entropic optimal transport problem on the samples. See the Supplementary Material 1.1 for details on  $g_\epsilon$  and important properties of  $\hat{T}_\epsilon$ . Practically, the calculation of  $g_\epsilon$  has been well studied and there exists fast algorithms for its computation (Peyré & Cuturi, 2020). The estimate has also been shown to converge (with rates) to the optimal transport map in  $L_2(\mu_0)$  under the appropriate technical conditions.

To accompany this we also have the following result, stated informally below. In this result  $a \lesssim b$  means that there exists a constant  $C > 0$  such that  $a \leq Cb$ . We allow this constant  $C$  to depend on the size of the support of the measures, upper and lower bounds on their densities, and the regularity properties of the optimal transport maps  $T_i, T_j$ , but it is importantly independent of the sample size  $n$ .

**Theorem 3. (Informal)** *Let  $i, j \in \{1, \dots, p\}$  and suppose that  $\mu_i, \mu_j, \mu_0$  are supported on bounded domains and that the maps  $T_i$  and  $T_j$  are sufficiently regular. Let  $X_1, \dots, X_{2n} \sim \mu_0, Y_1, \dots, Y_n \sim \mu_i, Z_1, \dots, Z_n \sim \mu_j$ . For an appropriately chosen  $\epsilon$ , let  $\hat{T}_i$  and  $\hat{T}_j$  be the entropic maps computed using  $\{X_i\}_{i=1}^{2n}, \{Y_i\}_{i=1}^n, \{Z_i\}_{i=1}^n$ . Then*

we have

$$\begin{aligned} \mathbb{E} \left[ \left\| A_{ij} - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i(X_k) - X_k, \hat{T}_j(X_k) - X_k \rangle \right\| \right] \\ \lesssim \frac{1}{\sqrt{n}} + n^{-\frac{\alpha+1}{4(d'+\alpha+1)}} \sqrt{\log n} \end{aligned} \quad (8)$$

where  $d' = 2\lceil d/2 \rceil$ , and  $\alpha \leq 3$  depends on the regularity of optimal transport maps.

A precise statement and proof of this theorem is given in the Supplementary Material Section 2.5. Theorem 3 tells us that in expectation, by using the entropic map we can accurately approximate the entries of the matrix  $A_{ij}$ , and do so in a numerically feasible way. We chose the entropic map as it is computationally tractable and has competitive convergence rates when compared to the intractable but optimal estimator of (Hütter & Rigollet, 2021). We expect similar analysis holds for other recently proposed estimators (Deb et al., 2021; Gunsilius, 2021; Manole et al., 2021; Muzellec et al., 2021) under different assumptions.

Theorem 3 can be combined with a matrix perturbation analysis to demonstrate the consistency of the estimated barycentric coordinates  $\hat{\lambda}$  from Algorithm 1 for the true ones  $\lambda_*$  when  $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$ , that is when  $\lambda_*^T A \lambda_* = 0$ . This is stated more precisely as follows.

**Corollary 2.** *Let  $\hat{\lambda}$  be the random estimate obtained from Algorithm 1. Suppose that  $A$  has an eigenvalue of 0 with multiplicity 1 and that  $\lambda_* \in \Delta^p$  realizes  $\lambda_*^T A \lambda_* = 0$ . Then under the assumptions of Theorem 3,*

$$\mathbb{E}[\|\hat{\lambda} - \lambda_*\|_2^2] \lesssim \frac{1}{\sqrt{n}} + n^{-\frac{\alpha+1}{4(d'+\alpha+1)}} \sqrt{\log n}.$$

The proof is deferred to the Supplementary Material Section 2.6. Note that the implicit constant in the inequality of Corollary 2 depends on  $p$  in addition to the dependencies of listed above. Corollary 2 ensures that in the large sample limit, the entropic regularization parameter may be chosen to guarantee precise estimation of the true barycentric coordinates. Note that the rate of convergence in Corollary 2 depends crucially on the smoothness of the transport maps between the reference measures ( $\alpha$ ) and the dimensionality of the space in which they are supported ( $d'$ ).

In the next section we illustrate the utility of the BCM and propose novel approaches based on Theorem 1 and Algorithm 1 for several applications.

## 4. Applications

### 4.1. Barycenter Parameterized Covariance Estimation

Extending the set-up in (Musolas et al., 2021), we first consider the case where  $\mu_i = \mathcal{N}(0, S_i)$ ,  $S_i \in \mathbb{S}_{++}^d$  for

---

### Algorithm 1 Estimate $\lambda$

---

**Input:** i.i.d. samples  $\{X_1, \dots, X_{2n}\} \sim \mu_0$ ,  $\{Y_1^i, \dots, Y_n^i\} \sim \mu_i : i = 1, \dots, p\}$ , regularization parameter  $\epsilon > 0$ .

**for**  $i = 1, \dots, p$  **do**

Set  $M^i \in \mathbb{R}^{n \times n}$  with  $M_{jk}^i = \frac{1}{2} \|X_j - Y_k^i\|_2^2$ .

Solve for  $g^i$  as the optimal  $g$  in

$$\begin{aligned} \max_{f, g \in \mathbb{R}^n} \frac{1}{n} \sum_{j=1}^n f_j + \frac{1}{n} \sum_{k=1}^n g_k \\ - \frac{\epsilon}{n^2} \sum_{j,k} \exp((f_j + g_k - M_{jk}^i)/\epsilon). \end{aligned}$$

Define  $\hat{T}_i$  through 7 with  $g_\epsilon = g_i$  and  $\{Y_1^i, \dots, Y_n^i\}$ .

**end for**

Set  $\hat{A} \in \mathbb{R}^{p \times p}$  to be the matrix with entries

$$\hat{A}_{ij} = \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i(X_k) - X_k, \hat{T}_j(X_k) - X_k \rangle.$$

**Return**  $\hat{\lambda} = \arg \min_{\lambda \in \Delta^p} \lambda^T \hat{A} \lambda$ .

---

$i = 0, \dots, p$ . We will use  $S_i$  both as a matrix and to refer to  $\mu_i$ ; similarly, let  $S_\lambda$  denote  $\nu_\lambda$ . Let  $\{x_i\}_{i=1}^n$  be i.i.d. samples with empirical covariance  $\hat{S}$ . In this setting, we can use the formulas laid out in Corollary 1.

For comparison, we also consider maximum likelihood estimation (MLE) for the parameter  $\lambda$  in the BCM. Let  $P_\lambda$  denote the probability density function of  $S_\lambda$ . The MLE is  $\arg \max_{\lambda \in \Delta^p} \prod_{i=1}^n P_\lambda(x_i)$ , which is equivalent to minimizing the KL-divergence  $D_{KL}(\hat{S}, S_\lambda)$ . However, this problem may be non-convex and difficult to optimize. We solve it numerically using auto-differentiation (Bartholomew-Biggs et al., 2000) and use the coordinates recovered by this method, which may not be the true MLE.

**Experimental Setup:** We consider the problem of estimating the covariance matrix  $S_0$  from i.i.d. samples  $\{x_i\}_{i=1}^n \sim \mathcal{N}(0, S_0)$  when  $S_0$  is known to be a barycenter of  $S_1, \dots, S_p$ . To do so we use the following procedure:

1. **Choose  $\lambda$  and  $\{S_i\}_{i=1}^p$ :** Select  $\lambda \in \Delta^p$  and the reference measures  $\{S_i\}_{i=1}^p$  either by hand or at random.
2. **Generate  $S_\lambda$ :** Using Algorithm 1 of (Chewi et al., 2020), find the true barycenter,  $S_\lambda$ .
3. **Sample:** Sample  $\{x_i\}_{i=1}^n$  i.i.d. from  $\mathcal{N}(0, S_\lambda)$ . Compute the empirical covariance  $\hat{S} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$ .
4. **Estimate  $\hat{\lambda}$ :** Perform the optimization in Corollary 1, or use MLE with  $S_0 = \hat{S}$  to recover an estimate  $\hat{\lambda}$ .
5. **Compute  $S_{\hat{\lambda}}$ :** Again using Algorithm 1 of (Chewi et al., 2020) find  $S_{\hat{\lambda}}$  and use that as the estimate of  $S_\lambda$ .



The results are in Figure 3. These show that using our approach (denoted *gradient norm*) we can estimate the underlying covariance matrix much more accurately than the empirical covariance. Our optimization is competitive with the auto-differentiation approach to MLE, while being significantly more stable and much faster. Indeed, computing all the estimates  $\hat{\lambda}$  to generate Figure 3 takes our method less than 1 second compared to over 3 hours and 40 minutes for MLE. For reference it takes on average roughly 0.01 seconds to perform empirical covariance estimation<sup>2</sup>. Further details are in the Supplementary Material Section 3.

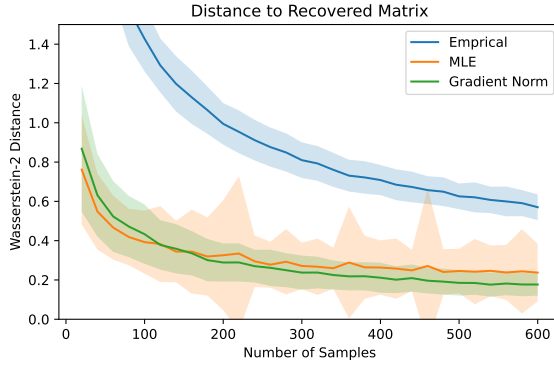


Figure 3:  $W_2$  between true covariance and estimates. We set  $p = 6$ ,  $d = 10$ ,  $\lambda \sim \text{Unif}(\Delta^p)$ , and  $\mu_i \sim \text{Wish}(I_d) + 0.5I_d$  with results averaged over 250 trials, and 1 standard deviation shaded.

We now consider applications where it is necessary to directly estimate the optimal transport map from samples (unlike the previous case where we estimate it via sample-covariance estimators) and between measures with non-uniform weights on the observed support. To that end, we first modify Algorithm 1 for practical purposes.

## 4.2. Adapting Algorithm 1 for Applications

In practice, Algorithm 1 has a few mild requirements: that we use the same number of samples for each reference measure; that the weight on each sample is the same; and access to an extra set of points for evaluating the inner products. Algorithm 2 relaxes these requirements and is in terms of matrix-vector operations, making it suitable for practical implementation. We note that if one uses Algorithm 2 with the same constraints as in Algorithm 1, then both produce the exact same  $\hat{T}_i$  on the samples  $X_1, \dots, X_n$  (Pooladian & Niles-Weed, 2021). The main difference between the two approaches is that the dual potentials in Algorithm 1 make an out-of-sample extension possible which may be of its own interest.

<sup>2</sup>All reported times are obtained using 20 Intel Xeon CPU E5-2660 V4 cores at 2.00 GHz.

## Algorithm 2 Estimate $\lambda$ on Point Clouds

**Input:** PMFs  $p \in \mathbb{R}^{n_0}$ ,  $q^i \in \mathbb{R}^{n_i}$ , support matrices  $X \in \mathbb{R}^{n_0 \times d}$ ,  $Y^i \in \mathbb{R}^{n_i \times d}$ , regularization parameter  $\epsilon > 0$ .

**for**  $i = 1, \dots, p$  **do**

Set  $M^i \in \mathbb{R}^{n_0 \times n_i}$  with  $M_{jk}^i = \|X_j - Y_k^i\|_2^2$ .

Solve for the entropic assignment matrix  $\pi^i$  as the optimal matrix in

$$\min_{\substack{\pi \in \mathbb{R}_+^{n_0 \times n_i} \\ \pi \mathbf{1} = p \\ \pi^T \mathbf{1} = q^i}} \sum_{j=1}^{n_0} \sum_{k=1}^{n_i} M_{jk}^i \pi_{jk} + \epsilon \pi_{jk} \log \pi_{jk}.$$

Compute the approximate transport matrix  $\hat{T}_i \in \mathbb{R}^{n_0 \times d}$  as

$$\hat{T}_i = \text{diag}(1/p) \pi^i Y^i.$$

**end for**

Set  $\hat{A} \in \mathbb{R}^{p \times p}$  to be the matrix with entries

$$\hat{A}_{ij} = \text{Tr} \left( \text{diag}(p) (\hat{T}_i - X) (\hat{T}_j - X)^T \right).$$

**Return**  $\hat{\lambda} = \arg \min_{\lambda \in \Delta^p} \lambda^T \hat{A} \lambda$ .

## 4.3. Image Inpainting and Denoising

We consider the problem of recovering an image in the presence of corruption, interpreting the image as a probability measure. We consider two specific models of corruption: **a.** additive *noise* and **b.** *occlusion* of a portion of the image. Our experimental procedure for both of these problems is outlined below taking as experimental data the MNIST dataset of  $28 \times 28$  pixel images of hand-written digits (LeCun, 1998). Additional details are in Supplementary Material Section 4. Note that after normalization, these can be treated as measures supported on a  $28 \times 28$  grid.

### Experimental Setup:

- Select  $\mu_0$ :** Select a digit  $\mu_0$  **a.** generate a white noise image  $\zeta$ . Set  $\tilde{\mu}_0$  to be  $\tilde{\mu}_0 = (1 - \alpha)\mu_0 + \alpha\zeta$ . **b.** Set  $\tilde{\mu}_0$  to be  $\mu_0$  with the central  $8 \times 8$  square removed, and renormalized.
- Select  $\{\mu_i\}_{i=1}^p$ :** Select a set of images of the same digit as  $\mu_0$  to serve as the reference measures. **a.** Let  $\tilde{\mu}_i = (1 - \alpha)\mu_0 + \alpha\mathbb{E}[\zeta]$ . **b.** Let  $\tilde{\mu}_i$  be  $\mu_i$  with the central  $8 \times 8$  square removed, and renormalized
- Estimate  $A$ ,  $\lambda$ :** Use Algorithm 2 to compute the approximate Gram matrix  $\hat{A}$  using  $\tilde{\mu}_0, \tilde{\mu}_1, \dots, \tilde{\mu}_p$  and then compute estimated coordinates  $\hat{\lambda}$ .
- Compute  $\hat{\nu}_\lambda$ :** Output  $\hat{\mu}_0 = \hat{\nu}_\lambda$ , where the barycenter is reconstructed from  $\mu_1, \dots, \mu_p$ .

An illustration of the procedure is shown in Figure 4. We

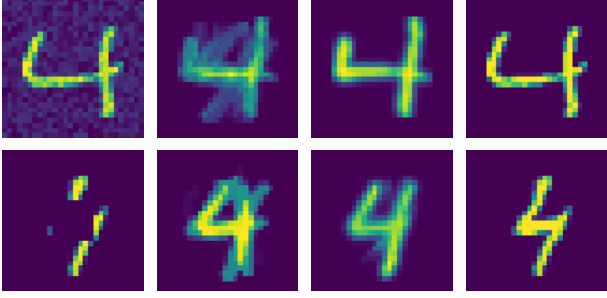


Figure 4: *Left to right*: corrupted image; recovery by linear projection; recovery using  $\hat{\lambda}$  from Algorithm 2; original image. *Top*: white noise added to the image as in **a**. *Bottom*: occlusion of the image as in **b**. We see the linear reconstruction fails, while Algorithm 2 recovers well, albeit with some blurring that can be interpreted as a consequence of entropic regularization.

also show for comparison digits reconstructed using a linear method, in which the corrupted image is recovered by computing the Euclidean projection onto the convex hull of the reference measures (see the Supplement 4.3 for further details). To recover the barycenter we use the method of (Benamou et al., 2015). As a competitive baseline we compare to a well-known existing method for histogram regression (Bonneel et al., 2016) in this setting. Our results are summarized in Tables 1 and 2. We note that our method is over an order of magnitude faster than (Bonneel et al., 2016) on this dataset and achieves competitive results. This is particularly remarkable because (Bonneel et al., 2016) is specifically adapted to measures with structured support such as grids and meshes, while our Algorithm 2 is much more general.

#### 4.4. Document Classification

Finally, we consider the task of identifying the topic of a document using its word embedding representation as the empirical distribution. We assume that there are  $t$  topics and that we have  $p$  reference documents about each topic. Each document can be represented as a high-dimensional empirical measure using a combination of a bag-of-words representation and a node2vec (Mikolov et al., 2013) word-embedding, and is paired with topic label. We use the publicly available dataset provided by (Huang et al., 2016). The task is to use a small number of labelled documents from each topic to predict the topic of a test document.

We consider four predictors: **(1) 1-Nearest Neighbor (1NN)** which classifies using the topic of the  $W_2^2$ -nearest reference document; **(2) Minimum Average Distance** which selects the topic with reference documents on average  $W_2^2$ -closest to the test document; **(3) Minimum Barycenter Loss** which runs Algorithm 2 using the references from each topic separately and selects the topic with the small-

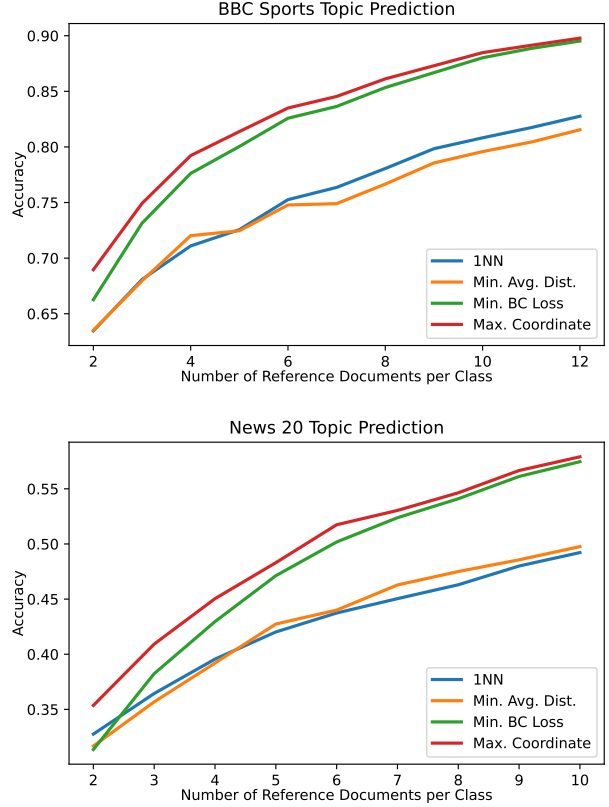


Figure 5: Document topic prediction accuracy as a function of number of reference documents in each class. *Top*: BBC Sport dataset (5 classes). *Bottom*: News20 dataset (20 classes).

est loss in the quadratic form; **(4) Maximum Coordinate** which runs Algorithm 2 using all the reference documents and chooses the topic which receives the most mass from  $\hat{\lambda}$ .

Our findings on the BBCSport (5 topics) and News20 (20 topics) datasets are reported in Figure 5. To generate these figures we randomly sample  $k$  reference documents per topic and a test set of 100 random documents. We apply the four methods listed above and compute the accuracy on the test set. This procedure is repeated 50 times for each choice of  $k$  and the average accuracy is plotted. See Supplementary Material Section 5.2 for details.

We see that even with a very small amount of training data, our BCM approaches perform well. Importantly, the BCM is able to represent unseen documents using these small training sets, giving it a clear advantage over simply classifying based on which document class it is  $W_2^2$ -nearest (1NN) or  $W_2^2$ -nearest in an aggregate sense (Min. Avg. Dist.). This suggests that even when  $p$  is small, the BCM has high modeling capacity and that the learned barycentric coordinates encode important information.



	Bon., $\epsilon = 10$	Alg. 2, $\epsilon = 10$	Alg. 2, $\epsilon = 0$
Occlusion	2.5101 (2228s)	2.5488 (3.371s)	2.5287 (1.062s)
Noise ( $\alpha = 0.5$ )	2.4058 (2391s)	2.6797 (86.32s)	2.3787 (50.67s)

Table 1: Average quality of recovery of  $\mu_0$  measured in  $W_2^2(\mu_0, \hat{\mu}_0)$  when reconstructing 500 random 4’s using a barycenter constructed from 10 random 4’s, as well as run times of each method. Algorithm 2 affords comparably accurate reconstructions in orders-of-magnitude faster time than the state-of-the-art (Bonneel et al., 2016).

	(Alg. 2, $\epsilon = 10$ ) - (Bon., $\epsilon = 10$ )	(Alg. 2, $\epsilon = 0$ ) - (Bon., $\epsilon = 10$ )	(Alg. 2, $\epsilon = 10$ ) - (Alg. 2, $\epsilon = 0$ )
Occlusion	$0.03872 \pm 0.3103$	$0.01868 \pm 0.2934$	$0.02004 \pm 0.2145$
Additive	$0.2739 \pm 0.3564$	$-0.02706 \pm 0.2558$	$0.3009 \pm 0.2486$

Table 2: Average and standard deviation of the pairwise-difference in reconstruction quality measured in  $W_2^2(\mu_0, \hat{\mu}_0)$  in the same setting as Table 1. The average pairwise-difference is smaller than the standard deviation in all cases involving (Bonneel et al., 2016), suggesting there is no statistically significant difference. Our approach should thus be preferred to (Bonneel et al., 2016), on the grounds that it is considerably more efficient while achieving essentially the same accuracy.

## 5. Conclusion

We have proposed a new method for computing coordinates under the BCM, together with guarantees on its solution via a convex program (Theorem 1) with closed-form coordinates in the Gaussian case (Corollary 1). We further developed an algorithm for estimating the coordinates under the BCM when all measures are accessible only via i.i.d. samples (Algorithm 1) which enjoys a natural smoothness and dimension-dependent rate of convergence to the true parameters (Theorem 3, Corollary 2). The BCM paradigm affords significant gains in runtime and robustness for barycenter parameterized Gaussian measure estimation, and provides an effective approach to image reconstruction and document classification when interpreting these data as measures. The results in this paper suggest the efficiency and effectiveness of the BCM as a broadly viable modelling tool.

**Future Work and Open Questions:** Section 3 provides a consistency analysis for  $A$  and  $\lambda$  by estimating the transport maps between each measure. But, all that is needed is an *estimate on the inner product* between the displacement maps, which in principle could admit more sample efficient approaches that avoid explicitly estimating the maps.

Section 4 suggests that unseen data can be well-represented in the BCM for randomly chosen reference measures. An interesting theoretical problem is to understand the representational capacity of the BCM using particular reference measures (e.g., random ones, optimally chosen ones, those in certain parametric families). This naturally touches on the question of the smoothness of the synthesis map  $\lambda \mapsto \nu_\lambda$ .

Computationally, all of our real data experiments use randomly chosen reference measures. It is of interest to develop efficient procedures as in (Schmitz et al., 2018) for *learning* reference measures that induce low reconstruction error. Beyond that, one could regularize the analysis problem (3) to encourage sparse or otherwise structured coordinates as is

well-studied in Euclidean settings (Aharon et al., 2006).

## Acknowledgements

MW is supported by NSF CCF-1553075. RJ is supported by NSF DRL 1931978. SA acknowledges support by NSF CCF-1553075, NSF DRL 1931978, NSF EEC 1937057, and AFOSR FA9550-18-1-0465. JMM acknowledges support from the Dreyfus Foundation and the NSF through grants DMS-1912737, DMS-1924513. All authors acknowledge support through the Tufts TRIPODS Institute, supported by the NSF under grant CCF-1934553. We acknowledge the reviewers for their helpful comments which improved the paper considerably.

## 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.
- Altschuler, J. M., Chewi, S., Gerber, P., and Stromme, A. J. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.
- Ambrosio, L., Gigli, N., and Savare, G. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*.

- Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2005.
- Bartholomew-Biggs, M., Brown, S., Christianson, B., and Dixon, L. Automatic differentiation of algorithms. *Journal of Computational and Applied Mathematics*, 124: 171–190, 12 2000.
- 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.
- Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.
- Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71:1–71:10, 2016.
- Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.
- Cheng, K., Aeron, S., Hughes, M. C., and Miller, E. Dynamical Wasserstein barycenters for time-series modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. J. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pp. 1276–1304. PMLR, 2020.
- Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster Wasserstein distance estimation with the Sinkhorn divergence. *arXiv preprint arXiv:2006.08172*, 2020.
- Claici, S., Chien, E., and Solomon, J. Stochastic Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 999–1008. PMLR, 2018.
- Colombo, P., Staerman, G., Clavel, C., and Piantanida, P. Automatic text evaluation through the lens of Wasserstein barycenters. *arXiv preprint arXiv:2108.12463*, 2021.
- Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.
- Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 685–693. PMLR, 2014.
- Deb, N., Ghosal, P., and Sen, B. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34, 2021.
- Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
- Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l’institut Henri Poincaré*, 10(4):215–310, 1948.
- Genevay, A. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE), 2019.
- Gunsilius, F. F. On the convergence rate of potentials of breiner maps. *Econometric Theory*, pp. 1–37, 2021.
- Huang, G., Quo, C., Kusner, M. J., Sun, Y., Weinberger, K. Q., and Sha, F. Supervised word mover’s distance. In *Advances in Neural Information Processing Systems*, pp. 4869–4877, 2016.
- Hütter, J.-C. and Rigollet, P. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2): 1166–1194, 2021.
- Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.
- Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., and Uribe, C. On the complexity of approximating Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 3530–3540. PMLR, 2019.
- Le Gouic, T. and Loubes, J.-M. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2017.
- LeCun, Y. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- Mallat, S. *A wavelet tour of signal processing*. Elsevier, 1999.
- Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.
- McCann, R. J. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

- Montesuma, E. F. and Mboula, F. M. N. Wasserstein barycenter for multi-source domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16785–16793, 2021.
- Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Musolas, A., Smith, S. T., and Marzouk, Y. Geodesically parameterized covariance estimation. *SIAM Journal on Matrix Analysis and Applications*, 42(2):528–556, 2021.
- Muzellec, B., Vacher, A., Bach, F., Vialard, F.-X., and Rudi, A. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint arXiv:2112.01907*, 2021.
- Panaretos, V. M. and Zemel, Y. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.
- Peyré, G. and Cuturi, M. Computational optimal transport, 2020.
- Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. *arXiv: 2109.12004*, 2021.
- Popoviciu, T. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9:129–145, 1935.
- 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.
- Redko, I., Courty, N., Flamary, R., and Tuia, D. Optimal transport for multi-source domain adaptation under target shift. In *International Conference on Artificial Intelligence and Statistics*, pp. 849–858. PMLR, 2019.
- Santambrogio, F. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 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.
- Schwerdtfeger, H. *Introduction to linear algebra and the theory of matrices*. P. Noordhoff, 1961.
- Singh, S. P., Hug, A., Dieuleveut, A., and Jaggi, M. Context mover’s distance & barycenters: Optimal transport of contexts for building representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3437–3449. PMLR, 2020.
- Tošić, I. and Frossard, P. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, 2011.
- Villani, C. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.
- Xu, H., Wang, W., Liu, W., and Carin, L. Distilled Wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems*, pp. 1716–1725, 2018.
- Yang, L., Li, J., Sun, D., and Toh, K.-C. A fast globally linearly convergent algorithm for the computation of Wasserstein barycenters. *Journal of Machine Learning Research*, 22(21):1–37, 2021.

## Supplementary Material: Measure Estimation in the Barycentric Coding Model

### 1. Precise Statements of Technical Results

For completeness, we replicate the theorem describing conditions under which Karcher means are guaranteed to be barycenters.

**Theorem 4.** ((Panaretos & Zemel, 2020), Theorem 3.1.15) *Let  $\lambda \in \Delta^p$  with  $\lambda_i \neq 0$  for all  $i$ . Let  $\Omega \subset \mathbb{R}^d$  be open and convex and let  $\mu_1, \dots, \mu_p \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  supported on  $\Omega$  with bounded, strictly positive densities  $g_1, \dots, g_p$ . Suppose that  $\mu_0$  is an absolutely continuous Karcher mean for  $\lambda$  and is supported on  $\Omega$  with bounded strictly positive density  $f$  there. Then  $\mu_0$  is the Wasserstein barycenter for coordinates  $\lambda$  if one of the following holds:*

- (1)  $\Omega = \mathbb{R}^d$  and the densities  $f, g_1, \dots, g_p$  are locally Hölder continuous.
- (2)  $\Omega$  is bounded and the densities  $f, g_1, \dots, g_p$  are bounded below on  $\Omega$ .

Next we include a technical theorem which we will leverage in our proof of the convergence of the entries  $A_{ij}$ . Before doing so, we introduce the necessary background.

**Theorem 5.** ((Brenier, 1991)) *Let  $\mu \in \mathcal{P}_{2,\text{ac}}(\Omega)$  and  $\nu \in \mathcal{P}(\Omega)$ . Then*

1. *There exists a solution  $T_0$  to (1), with  $T_0 = \nabla \varphi_0$ , for a convex function  $\varphi_0$  solving*

$$\inf_{\varphi \in L^1(\mu)} \int \varphi d\mu + \int \varphi^* d\nu$$

where  $\varphi^*$  is the convex conjugate of  $\varphi$ .

2. *If in addition  $\nu \in \mathcal{P}_{2,\text{ac}}(\Omega)$ , then  $\nabla \varphi_0^*$  is an admissible optimal transport map from  $\nu$  to  $\mu$ .*

Let  $\mathcal{C}^\alpha$  denote the space of functions possessing  $\lfloor \alpha \rfloor$  continuous derivatives and whose  $\lfloor \alpha \rfloor^{\text{th}}$  derivative is  $(\alpha - \lfloor \alpha \rfloor)$  Hölder smooth. Using these conventions, we require the technical conditions:

**A4:**  $\mu, \nu \in \mathcal{P}_{2,\text{ac}}(\Omega)$  for a compact convex set  $\Omega$ , with densities satisfying  $p(x), q(x) \leq M$  and  $q(x) \geq m > 0$  for all  $x \in \Omega$ .

**A5:**  $\varphi_0 \in \mathcal{C}^2(\Omega)$  and  $\varphi_0^* \in \mathcal{C}^{\alpha+1}(\Omega)$  for  $\alpha > 1$ .

**A6:** There exist  $l, L > 0$  such that  $T_0 = \nabla \varphi_0$ , with  $lI \preceq \nabla^2 \varphi_0(x) \preceq LI$  for all  $x \in \Omega$ .

This first assumption is a specialization of **A1** and **A2**, restricting to the case where  $\Omega$  is compact. The latter two can be thought of as regularity conditions on the optimal transport maps between base and reference measures. **A5** asserts that the optimal transport maps have smooth derivatives and **A6** controls the derivatives from both above and below.

#### 1.1. Properties of The Entropic Map

Let  $X_i \sim \mu_0, Y_i \sim \mu_1$  for  $i = 1, \dots, n$  be i.i.d. samples from their respective distributions. From these samples construct the empirical measures  $\hat{\mu}_{0,n} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$  and  $\hat{\mu}_{1,n} = \frac{1}{n} \sum_{i=1}^n \delta_{Y_i}$ .

For  $\epsilon > 0$  and  $n$  samples from each measures, the discrete entropically regularized OT problem can be written

$$\min_{\substack{\pi \in \mathbb{R}_+^{n \times n} \\ \pi \mathbf{1} = (1/n) \mathbf{1} \\ \pi^T \mathbf{1} = (1/n) \mathbf{1}}} \sum_{j,k=1}^n \pi_{jk} \|X_j - Y_k\|_2^2 + \epsilon \pi_{jk} \log \pi_{jk}$$

This problem has a dual formulation (Genevay, 2019) given by

$$\max_{f, g \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n f_i + g_i - \epsilon \langle e^{f/\epsilon}, K e^{g/\epsilon} \rangle$$

where  $K_{i,j} = \exp(-\|X_i - Y_j\|_2^2/\epsilon)$ .  $g_\epsilon$  is defined to be the optimal  $g$  in the maximization above, parameterized by the setting of  $\epsilon > 0$ . For the method proposed it is important that the  $f_\epsilon$  and  $g_\epsilon$  are chosen such that for all  $x, y \in \mathbb{R}^d$  we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \exp \left( \frac{1}{\epsilon} \left[ f_\epsilon(x) + g_\epsilon(Y_i) - \frac{1}{2} \|x - Y_i\|_2^2 \right] \right) &= 1 \\ \frac{1}{n} \sum_{i=1}^n \exp \left( \frac{1}{\epsilon} \left[ f_\epsilon(X_i) + g_\epsilon(y) - \frac{1}{2} \|X_i - y\|_2^2 \right] \right) &= 1. \end{aligned}$$

This is always possible, and can be easily done in practice (Pooladian & Niles-Weed, 2021).

One of the primary motivations for selecting the entropic map as our map estimate is that  $g_\epsilon$  can be efficiently computed (Peyré & Cuturi, 2020), and it comes with the performance guarantee stated below.

In the following theorem the constants may depend on the dimension  $d$ , the diameter of the support  $|\Omega|$ ,  $M, m, l, L$ , and  $\|\varphi_0^*\|_{C^{\alpha+1}}$ .

**Theorem 6.** ((Pooladian & Niles-Weed, 2021) Theorem 3) Under **A4,A5,A6**, the entropic map  $\hat{T} = T_{\epsilon, (n,n)}$  from  $\hat{\mu}_n$  to  $\hat{\nu}_n$  with regularization parameter  $\epsilon \asymp n^{-1/(d'+\bar{\alpha}+1)}$  satisfies

$$\mathbb{E} \|\hat{T} - T_0\|_{L^2(\mu)}^2 \lesssim (1 + I_0(\mu, \nu)) n^{-\frac{(\bar{\alpha}+1)}{2(d'+\bar{\alpha}+1)}} \log n \quad (9)$$

where  $d' = 2\lceil d/2 \rceil$  and  $\bar{\alpha} = \min(\alpha, 3)$  and  $I_0(\mu, \nu)$  is the integrated Fisher information between  $\mu$  and  $\nu$ .

For more information on  $I_0(\mu, \nu)$  see (Chizat et al., 2020). We remark that when  $\alpha \geq 2$  then under **A4,A5,A6**, it has been shown  $I_0(\mu, \nu) \leq C$  for a positive constant  $C$  (Chizat et al., 2020). The notation  $a \asymp b$  means that there exists positive constants  $c, C$  such that  $ca \leq b \leq Ca$  and  $a \lesssim b$  means that there is a positive constant  $C$  such that  $a \leq Cb$ . Note that this result still holds without the convexity condition in **A4**. We include it for ease of exposition below and to ensure that the interior of an  $\Omega$  that satisfies **A4** is convex and therefore satisfies **A3**.

## 2. Proofs of Theoretical Results

### 2.1. Proof of Proposition 1

*Proof.* We first apply the definitions given in Section 2 and leverage (4):

$$\begin{aligned} \|\nabla G_\lambda(\mu_0)\|_{\mu_0}^2 &= \left\| - \sum_{i=1}^p \lambda_i (T_{\mu_0 \rightarrow \mu_i} - \text{Id}) \right\|_{\mu_0}^2 \\ &= \int_{\mathbb{R}^d} \left\langle \sum_{i=1}^p \lambda_i (T_{\mu_0 \rightarrow \mu_i} - \text{Id}), \sum_{j=1}^p \lambda_j (T_{\mu_0 \rightarrow \mu_j} - \text{Id}) \right\rangle d\mu_0 \\ &= \int_{\mathbb{R}^d} \sum_{i,j=1}^p \lambda_i \lambda_j \langle T_{\mu_0 \rightarrow \mu_i} - \text{Id}, T_{\mu_0 \rightarrow \mu_j} - \text{Id} \rangle d\mu_0 \\ &= \sum_{i,j=1}^p \lambda_i \lambda_j A_{ij} = \lambda^T A \lambda. \end{aligned}$$

To establish the convexity of the function in terms of  $\lambda$ , it is sufficient to show that  $A$  is symmetric positive semi-definite. Observe that  $A$  is a Gram matrix of the set  $\{T_{\mu_0 \rightarrow \mu_i} - \text{Id}\}_{i=1}^p \subset T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  with the inner product in  $T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ . It is classical that all Gram matrices are symmetric positive semi-definite (Schwerdtfeger, 1961) and this completes the proof.  $\square$

### 2.2. Proof of Theorem 1

*Proof.* The convexity of the objective is established in Proposition 1, and the feasible set  $\Delta^p$  is trivially convex.

By Theorem 4, when assumptions **A1-A3** hold, all Karcher means are barycenters, and therefore  $\mu_0 = \nu_\lambda$ , and it is always the case that barycenters are Karcher means. Therefore it is sufficient to check that the minimum value is 0 if and only if  $\mu_0$  is a Karcher mean.

In the case that the minimum is zero, there is some  $\lambda_*$  such that  $(\lambda_*)^T A \lambda_* = 0$  and we can apply Proposition 1 to show that  $\mu_0$  is a Karcher mean for  $\lambda_*$ . Similarly if  $\mu_0$  is a Karcher mean for some  $\lambda_*$  by definition we have  $\|\nabla G_{\lambda_*}(\mu_0)\|_{\mu_0}^2 = 0$  and again by Proposition 1 we have that  $\lambda_*$  achieves the minimum of zero in the optimization.  $\square$

### 2.3. Proof of Corollary 1

*Proof.* **A1, A2, A3** are clearly satisfied in the Gaussian case with  $\Omega = \mathbb{R}^d$  since the pdf of a non-degenerate multivariate Gaussian is positive everywhere and Lipschitz continuous. Furthermore the barycenter of zero-mean Gaussians is itself a zero-mean Gaussian ((Agueh & Carlier, 2011) Theorem 6.1).

Therefore we can apply Theorem 1 and all that remains is to calculate  $A_{ij}$  as follows. Noting that the  $C_i$  are all symmetric (Agueh & Carlier, 2011) we have

$$\begin{aligned} A_{ij} &= \int_{\mathbb{R}^d} \langle (T_i(x) - \text{Id}(x)), (T_j(x) - \text{Id}(x)) \rangle dS_0(x) \\ &= \int_{\mathbb{R}^d} \langle (C_i - I)x, (C_j - I)x \rangle dS_0(x) \quad (\text{Eq. 6}) \\ &= \mathbb{E}_{X \sim \mathcal{N}(0, S_0)} \left[ X^T (C_i - I)^T (C_j - I) X \right] \\ &= \mathbb{E}_{X \sim \mathcal{N}(0, S_0)} \left[ X^T (C_i - I) (C_j - I) X \right] \quad (C_i \text{ all sym.}) \\ &= \text{Tr}((C_i - I)(C_j - I)S_0). \end{aligned}$$

The last equality comes from the identity  $\mathbb{E}_{X \sim \mathcal{N}(0, S)} [X^T B X] = \text{Tr}(BS)$  which holds for all  $B \in \mathbb{R}^{d \times d}$  and  $S \in \mathbb{S}_{++}^d$ .  $\square$

### 2.4. Proof of Theorem 2

*Proof.* In this setting, the barycenter  $\nu_\lambda$  can be expressed as

$$\nu_\lambda = \left[ \sum_{i=1}^p \lambda_i T_0^i \right] \# \mu_0$$

((Panaretos & Zemel, 2020) Theorem 3.1.9). Together with Proposition 1, this implies (recalling  $T_i = T_0^i$ )

$$W_2^2(\mu_0, \nu_\lambda) = \int \|x - \sum_{i=1}^p \lambda_i T_i(x)\|_2^2 d\mu_0(x) = \|\nabla G_\lambda(\mu_0)\|_{\mu_0}^2 = \lambda^T A \lambda$$

and therefore

$$\arg \min_{\lambda \in \Delta^p} W_2^2(\mu_0, \nu_\lambda) = \arg \min_{\lambda \in \Delta^p} \lambda^T A \lambda.$$

$\square$

### 2.5. Exact Statement and Proof of Theorem 3

**Theorem 7.** (Formal Version of Theorem 3, Convergence Rate of  $A_{ij}$ ) Let  $T_1, T_2$  be the optimal transport maps from  $\mu_0$  to  $\mu_1, \mu_2$  respectively. Suppose that **A4, A5, A6** are satisfied for both pairs  $(\mu_0, \mu_1)$  and  $(\mu_0, \mu_2)$ . Let  $X_1, \dots, X_{2n} \sim \mu_0, Y_1, \dots, Y_n \sim \mu_1, Z_1, \dots, Z_n \sim \mu_2$ . Let  $\hat{T}_1$  and  $\hat{T}_2$  be the entropic maps from  $\mu_0$  to  $\mu_1$  and  $\mu_2$  respectively, both with  $\epsilon \asymp n^{-1/(d' + \bar{\alpha} + 1)}$  and computed using  $X_1, \dots, X_n$ . Then we have

$$\begin{aligned} &\mathbb{E} \left[ \left| \int \langle T_1 - \text{Id}, T_2 - \text{Id} \rangle d\mu_0 - \frac{1}{n} \sum_{i=n+1}^{2n} \langle \hat{T}_1(X_i) - X_i, \hat{T}_2(X_i) - X_i \rangle \right| \right] \\ &\lesssim \frac{1}{\sqrt{n}} + n^{-\frac{\bar{\alpha}+1}{4(d'+\bar{\alpha}+1)}} \sqrt{\log n} \sqrt{1 + I_0(\mu_0, \mu_1) + I_0(\mu_0, \mu_2)} \end{aligned}$$

where  $d' = 2\lceil d/2 \rceil$ ,  $\bar{\alpha} = \alpha \wedge 3$  and  $I_0(\mu_0, \mu_1)$  is the integrated Fisher information along the Wasserstein geodesic between  $\mu$  and  $\mu_1$ .



Before we begin the proof, we remark that by writing  $\mu_i, \mu_j$  instead of  $\mu_1, \mu_2$ , the random variable in the expectation is just  $|A_{ij} - \hat{A}_{ij}|$  and therefore this result allows us to control the entry-wise deviations of  $\hat{A}_{ij}$  from their true values.

*Proof.* Throughout, unless otherwise noted, the expectation is with respect to all of  $X_1, \dots, X_{2n}, Y_1, \dots, Y_n$  and  $Z_1, \dots, Z_n$ . We also use the notations  $X^n$  to denote the set of variables  $(X_1, \dots, X_n)$ , and similarly for  $Y^n$  and  $Z^n$ .

First observe that we can assume without loss of generality that  $0 \in \Omega$ ; see (Peyré & Cuturi, 2020) Remark 2.9 for invariance of  $T_1, T_2$  under translation. For the translation invariance of  $\hat{T}_1, \hat{T}_2$ , observe that both obtaining  $g_\epsilon$ , and evaluating  $\hat{T}_1, \hat{T}_2$  at fixed points, requires only the distances  $\|X_i - Y_j\|_2^2 = \|(X_i - t) - (Y_j - t)\|_2^2$ . We calculate:

$$\begin{aligned} & \mathbb{E} \left[ \left| \int \langle T_1 - \text{Id}, T_2 - \text{Id} \rangle d\mu_0 - \frac{1}{n} \sum_{i=n+1}^{2n} \langle \hat{T}_1(X_i) - X_i, \hat{T}_2(X_i) - X_i \rangle \right| \right] \\ &= \mathbb{E} \left[ \left| \int \langle T_1 - \text{Id}, T_2 - \text{Id} \rangle d\mu_0 - \frac{1}{n} \sum_{i=n+1}^{2n} \langle T_1(X_i) - X_i, T_2(X_i) - X_i \rangle \right. \right. \\ & \quad \left. \left. + \frac{1}{n} \sum_{i=n+1}^{2n} \langle T_1(X_i) - X_i, T_2(X_i) - X_i \rangle - \frac{1}{n} \sum_{i=n+1}^{2n} \langle \hat{T}_1(X_i) - X_i, \hat{T}_2(X_i) - X_i \rangle \right| \right] \\ &\leq \mathbb{E} \left[ \left| \int \langle T_1 - \text{Id}, T_2 - \text{Id} \rangle d\mu_0 - \frac{1}{n} \sum_{i=n+1}^{2n} \langle T_1(X_i) - X_i, T_2(X_i) - X_i \rangle \right| \right] \\ & \quad + \mathbb{E} \left[ \left| \frac{1}{n} \sum_{i=n+1}^{2n} \langle T_1(X_i) - X_i, T_2(X_i) - X_i \rangle - \frac{1}{n} \sum_{i=n+1}^{2n} \langle \hat{T}_1(X_i) - X_i, \hat{T}_2(X_i) - X_i \rangle \right| \right] \end{aligned}$$

We consider the first and second terms separately. For brevity, let  $h$  be the function defined by

$$h(x) = \langle T_1(x) - x, T_2(x) - x \rangle.$$

Note that  $h$  is bounded on  $\Omega$  since

$$|h(x)| = |\langle T_1(x) - x, T_2(x) - x \rangle| \leq \|T_1(x) - x\|_2 \|T_2(x) - x\|_2 \leq (2|\Omega|)(2|\Omega|) = 4|\Omega|^2,$$

which implies by Popoviciu's inequality (Popoviciu, 1935) that  $\text{Var}[h(X)] \leq 16|\Omega|^4$ . From this it follows

$$\begin{aligned} & \mathbb{E} \left[ \left| \int \langle T_1 - \text{Id}, T_2 - \text{Id} \rangle d\mu_0 - \frac{1}{n} \sum_{i=n+1}^{2n} \langle T_1(X_i) - X_i, T_2(X_i) - X_i \rangle \right| \right] \\ &= \mathbb{E} \left[ \left| \mathbb{E}_{X \sim \mu_0}[h(X)] - \frac{1}{n} \sum_{i=n+1}^{2n} h(X_i) \right| \right] \\ &\leq \sqrt{\frac{\text{Var}[h(X)]}{n}} \leq \sqrt{\frac{16|\Omega|^4}{n}} \lesssim \frac{1}{\sqrt{n}} \end{aligned}$$

where we have used that for all i.i.d. random variables  $U, U_1, \dots, U_n$  with finite variance

$$\mathbb{E}[|\mathbb{E}[U] - \frac{1}{n} \sum_{i=1}^n U_i|] \leq \sqrt{\frac{\text{Var}[U]}{n}}.$$

We now handle the second term:

$$\begin{aligned}
 & \mathbb{E} \left[ \left| \frac{1}{n} \sum_{i=n+1}^{2n} \langle T_1(X_i) - X_i, T_2(X_i) - X_i \rangle - \frac{1}{n} \sum_{i=n+1}^{2n} \langle \hat{T}_1(X_i) - X_i, \hat{T}_2(X_i) - X_i \rangle \right| \right] \\
 & \leq \frac{1}{n} \sum_{i=n+1}^{2n} \mathbb{E} \left[ \left| \langle T_1(X_i) - X_i, T_2(X_i) - X_i \rangle - \langle \hat{T}_1(X_i) - X_i, \hat{T}_2(X_i) - X_i \rangle \right| \right] \\
 & = \mathbb{E} \left[ \left| \langle T_1(X_{n+1}) - X_{n+1}, T_2(X_{n+1}) - X_{n+1} \rangle - \langle \hat{T}_1(X_{n+1}) - X_{n+1}, \hat{T}_2(X_{n+1}) - X_{n+1} \rangle \right| \right] \\
 & = \mathbb{E} \left[ \left| \langle T_1(X_{n+1}), T_2(X_{n+1}) \rangle - \langle \hat{T}_1(X_{n+1}), \hat{T}_2(X_{n+1}) \rangle \right. \right. \\
 & \quad \left. \left. + \langle \hat{T}_1(X_{n+1}) - T_1(X_{n+1}), X_{n+1} \rangle + \langle \hat{T}_2(X_{n+1}) - T_2(X_{n+1}), X_{n+1} \rangle \right| \right] \\
 & \leq \mathbb{E} \left[ \left| \langle T_1(X_{n+1}), T_2(X_{n+1}) \rangle - \langle \hat{T}_1(X_{n+1}), \hat{T}_2(X_{n+1}) \rangle \right| \right] \\
 & \quad + \mathbb{E} \left[ \left| \langle \hat{T}_1(X_{n+1}) - T_1(X_{n+1}), X_{n+1} \rangle \right| \right] + \mathbb{E} \left[ \left| \langle \hat{T}_2(X_{n+1}) - T_2(X_{n+1}), X_{n+1} \rangle \right| \right]
 \end{aligned}$$

Here we can control the three terms separately. We first take care of the middle term and the last term follows by an identical argument.

$$\begin{aligned}
 & \mathbb{E} \left[ \left| \langle \hat{T}_1(X_{n+1}) - T_1(X_{n+1}), X_{n+1} \rangle \right| \right] \\
 & = \mathbb{E}_{X^n, Y^n, X_{n+1}} \left[ \left| \langle \hat{T}_1(X_{n+1}) - T_1(X_{n+1}), X_{n+1} \rangle \right| \right] \\
 & \leq \mathbb{E}_{X^n, Y^n, X_{n+1}} \left[ \|\hat{T}_1(X_{n+1}) - T_1(X_{n+1})\|_2 \|X_{n+1}\|_2 \right] \quad (\text{Cauchy-Schwarz}) \\
 & \leq |\Omega| \mathbb{E}_{X^n, Y^n, X_{n+1}} \left[ \|\hat{T}_1(X_{n+1}) - T_1(X_{n+1})\|_2 \right] \quad (\|X_{n+1}\| \leq |\Omega|) \quad (10) \\
 & = |\Omega| \mathbb{E}_{X^n, Y^n, X_{n+1}} \left[ \mathbb{E}_{X_{n+1}} \|\hat{T}_1(X_{n+1}) - T_1(X_{n+1})\|_2 \right] \\
 & = |\Omega| \mathbb{E}_{X^n, Y^n, X_{n+1}} \left[ \mathbb{E}_{X_{n+1}} \sqrt{\|\hat{T}_1(X_{n+1}) - T_1(X_{n+1})\|_2^2} \right] \\
 & \leq |\Omega| \mathbb{E}_{X^n, Y^n, X_{n+1}} \left[ \sqrt{\mathbb{E}_{X_{n+1}} \|\hat{T}_1(X_{n+1}) - T_1(X_{n+1})\|_2^2} \right] \quad (\text{Jensen}) \\
 & = |\Omega| \mathbb{E}_{X^n, Y^n} \left[ \|\hat{T}_1 - T_1\|_{L^2(\mu_0)} \right] \\
 & \leq |\Omega| \sqrt{\mathbb{E}_{X^n, Y^n} \left[ \|\hat{T}_1 - T_1\|_{L^2(\mu_0)}^2 \right]} \\
 & \lesssim \sqrt{(1 + I_0(\mu_0, \mu_1)) n^{-\frac{\bar{\alpha}+1}{2(d'+\bar{\alpha}+1)} \log(n)}} \quad (\text{Theorem 6})
 \end{aligned}$$

As mentioned above the third term is exactly the same except replacing  $T_1$  and  $\hat{T}_1$  with  $T_2$  and  $\hat{T}_2$  as well as  $Y^n$  with  $Z^n$ .

To control the first term we start with a different trick and then end up following the same pattern as in the bound above.

$$\begin{aligned}
 & \mathbb{E} \left[ \left| \langle T_1(X_{n+1}), T_2(X_{n+1}) \rangle - \langle \hat{T}_1(X_{n+1}), \hat{T}_2(X_{n+1}) \rangle \right| \right] \\
 & = \mathbb{E} \left[ \left| \langle T_1(X_{n+1}) - \hat{T}_1(X_{n+1}), T_2(X_{n+1}) \rangle + \langle \hat{T}_1(X_{n+1}), T_2(X_{n+1}) - \hat{T}_2(X_{n+1}) \rangle \right| \right] \\
 & \leq \mathbb{E} \left[ \left| \langle T_1(X_{n+1}) - \hat{T}_1(X_{n+1}), T_2(X_{n+1}) \rangle \right| \right] + \mathbb{E} \left[ \left| \langle \hat{T}_1(X_{n+1}), T_2(X_{n+1}) - \hat{T}_2(X_{n+1}) \rangle \right| \right] \\
 & \leq \mathbb{E} \left[ \|T_1(X_{n+1}) - \hat{T}_1(X_{n+1})\|_2 \|T_2(X_{n+1})\|_2 \right] + \mathbb{E} \left[ \|\hat{T}_1(X_{n+1})\|_2 \|T_2(X_{n+1}) - \hat{T}_2(X_{n+1})\|_2 \right] \\
 & \leq |\Omega| \mathbb{E} \left[ \|T_1(X_{n+1}) - \hat{T}_1(X_{n+1})\|_2 \right] + |\Omega| \mathbb{E} \left[ \|T_2(X_{n+1}) - \hat{T}_2(X_{n+1})\|_2 \right] \\
 & = |\Omega| \mathbb{E}_{X^n, Y^n, X_{n+1}} \left[ \|T_1(X_{n+1}) - \hat{T}_1(X_{n+1})\|_2 \right] + |\Omega| \mathbb{E}_{X^n, Z^n, X_{n+1}} \left[ \|T_2(X_{n+1}) - \hat{T}_2(X_{n+1})\|_2 \right]. \quad (11)
 \end{aligned}$$

Above we have made use of the fact that both  $T_2(X_1)$  and  $\hat{T}_1(X_1)$  are contained in  $\Omega$  which implies that  $\|T_2(X_1)\|_2 \leq |\Omega|$ .  $T_2(X_1) \in \Omega$  follows from the fact that  $T_2$  is a map from  $\mu_0$  to  $\mu_2$ . To see that  $\|\hat{T}_1(X_1)\| \leq |\Omega|$ , note that from equation (7)  $\hat{T}_1(X_1)$  is a convex combination of  $Y_1, \dots, Y_n \in \Omega$  and each of which satisfies  $\|Y_i\|_2 \leq |\Omega|$ .

Observe that both terms in (11) have already been controlled, starting from Equation (10) (the latter requires replacing  $T_1, \hat{T}_1$ , and  $Y^n$  with  $T_2, \hat{T}_2$ , and  $Z^n$  respectively). Applying the bound derived there and combining terms we have the following result

$$\begin{aligned} & \mathbb{E} \left[ \left| \int \langle T_1 - \text{Id}, T_2 - \text{Id} \rangle d\mu_0 - \frac{1}{n} \sum_{i=n+1}^{2n} \langle \hat{T}_1(X_i) - X_i, \hat{T}_2(X_i) - X_i \rangle \right| \right] \\ & \lesssim \frac{1}{\sqrt{n}} + 2\sqrt{(1 + I_0(\mu_0, \mu_1))n^{-\frac{\bar{\alpha}+1}{2(d'+\bar{\alpha}+1)} \log(n)}} + 2\sqrt{(1 + I_0(\mu_0, \mu_2))n^{-\frac{\bar{\alpha}+1}{2(d'+\bar{\alpha}+1)} \log(n)}} \\ & = \frac{1}{\sqrt{n}} + n^{-\frac{\bar{\alpha}+1}{4(d'+\bar{\alpha}+1)}} \sqrt{\log n} \left( \sqrt{1 + I_0(\mu_0, \mu_1)} + \sqrt{1 + I_0(\mu_0, \mu_2)} \right) \\ & \lesssim \frac{1}{\sqrt{n}} + n^{-\frac{\bar{\alpha}+1}{4(d'+\bar{\alpha}+1)}} \sqrt{\log n} \sqrt{1 + I_0(\mu_0, \mu_1) + I_0(\mu_0, \mu_2)}. \end{aligned}$$

□

## 2.6. Proof of Corollary 2

*Proof.* Let  $B_n$  denote the entrywise bound in Theorem 3. Noting that  $\hat{\lambda}^T \hat{A} \hat{\lambda} \leq \lambda_*^T \hat{A} \lambda_*$  by construction, we estimate

$$\begin{aligned} \mathbb{E}[\hat{\lambda}^T A \hat{\lambda}] &= \mathbb{E}[\hat{\lambda}^T (A - \hat{A}) \hat{\lambda}] + \mathbb{E}[\hat{\lambda}^T \hat{A} \hat{\lambda}] \\ &\leq \mathbb{E}[\hat{\lambda}^T (A - \hat{A}) \hat{\lambda}] + \mathbb{E}[\lambda_*^T \hat{A} \lambda_*] \\ &= \mathbb{E}[\hat{\lambda}^T (A - \hat{A}) \hat{\lambda}] + \mathbb{E}[\lambda_*^T (\hat{A} - A) \lambda_*] \quad (\lambda_*^T A \lambda_* = 0) \\ &= \mathbb{E} \left[ \left| \sum_{i,j=1}^p (\hat{\lambda})_i (\hat{\lambda})_j (A - \hat{A})_{ij} \right| \right] + \mathbb{E} \left[ \left| \sum_{i,j=1}^p (\lambda_*)_i (\lambda_*)_j (A - \hat{A})_{ij} \right| \right] \\ &\leq \mathbb{E} \left[ \sum_{i,j=1}^p (\hat{\lambda})_i (\hat{\lambda})_j |A_{ij} - \hat{A}_{ij}| \right] + \mathbb{E} \left[ \sum_{i,j=1}^p (\lambda_*)_i (\lambda_*)_j |A_{ij} - \hat{A}_{ij}| \right] \quad (\hat{\lambda}, \lambda_* \in \Delta^p, \text{ triangle ineq.}) \\ &\leq 2 \mathbb{E} \left[ \sum_{i,j=1}^p |A_{ij} - \hat{A}_{ij}| \right] \\ &= 2 \sum_{i,j=1}^p \mathbb{E}[|A_{ij} - \hat{A}_{ij}|] \\ &\lesssim 2p^2 B_n \quad (\text{Theorem 3}) \end{aligned}$$

Since  $A$  is positive semidefinite and by assumption  $\lambda_* \in \Delta^p$  satisfies  $\lambda_*^T A \lambda_* = 0$ , it follows that  $\lambda_*$  is an eigenvector of  $A$  with eigenvalue 0. Let  $0 < \alpha_2 \leq \dots \leq \alpha_p$  be the non-zero eigenvalues of  $A$  with associated orthonormal eigenvectors  $v_2, \dots, v_p$ . Orthogonally decompose  $\hat{\lambda} = \hat{\beta} \lambda_* + \hat{\lambda}_\perp$ , where  $\hat{\beta} \in \mathbb{R}$  and  $\hat{\lambda}_\perp$  is in the span of  $\{v_2, \dots, v_p\}$ . Note that  $\hat{\beta}$  and  $\hat{\lambda}_\perp$  are random. Then,

$$\begin{aligned}
 \mathbb{E}[\|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2] &= \mathbb{E}[\|\hat{\lambda}_\perp\|_2^2] \\
 &= \mathbb{E}\left[\sum_{i=2}^p |v_i^T \hat{\lambda}_\perp|^2\right] \\
 &\leq \frac{1}{\alpha_2} \mathbb{E}\left[\sum_{i=2}^p \alpha_i |v_i^T \hat{\lambda}_\perp|^2\right] \\
 &= \frac{1}{\alpha_2} \mathbb{E}[(\hat{\lambda}_\perp)^T A \hat{\lambda}_\perp] \\
 &= \frac{1}{\alpha_2} \mathbb{E}[\hat{\lambda}^T A \hat{\lambda}] \\
 &\lesssim \frac{2p^2}{\alpha_2} B_n.
 \end{aligned}$$

Summing both sides of the equation  $\hat{\lambda} = \hat{\beta}\lambda_* + \hat{\lambda}_\perp$  and recalling  $\lambda_*, \hat{\lambda} \in \Delta^p$  yields

$$\begin{aligned}
 1 &= \hat{\beta} + \sum_{j=1}^p (\hat{\lambda}_\perp)_j \\
 &\leq \hat{\beta} + \|\hat{\lambda}_\perp\|_1 \\
 &\leq \hat{\beta} + \sqrt{p} \|\hat{\lambda}_\perp\|_2,
 \end{aligned}$$

which implies that  $\mathbb{E}[(1 - \hat{\beta})^2] \leq p \mathbb{E}[\|\hat{\lambda}_\perp\|_2^2] \lesssim \frac{2p^3}{\alpha_2} B_n$ .

Finally, we use the fact that  $\hat{\lambda} - \hat{\beta}\lambda_* = \hat{\lambda}_\perp$  and  $(\hat{\beta} - 1)\lambda_*$  are orthogonal to bound:

$$\begin{aligned}
 \mathbb{E}[\|\hat{\lambda} - \lambda_*\|_2^2] &= \mathbb{E}[\|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2 + \|(\hat{\beta} - 1)\lambda_*\|_2^2] \\
 &= \mathbb{E}[\|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2] + \mathbb{E}[\|(\hat{\beta} - 1)\lambda_*\|_2^2] \\
 &\lesssim \frac{2p^2}{\alpha_2} B_n + \mathbb{E}[(\hat{\beta} - 1)^2 \|\lambda_*\|_2^2] \\
 &\leq \frac{2p^2}{\alpha_2} B_n + \mathbb{E}[(\hat{\beta} - 1)^2] \\
 &\lesssim \frac{2p^2}{\alpha_2} B_n + \frac{2p^3}{\alpha_2} B_n
 \end{aligned}$$

as desired. □

## 2.7. Karcher Means are Barycenters under A4-A6

Corollary 2 (which holds under **A4-A6**) can be combined with Theorem 1 (which holds under **A1-A3**) as follows. Suppose  $\|G_\lambda(\nu)\|_\nu^2 = 0$  for some  $\lambda \in \Delta^p$  and reference measures  $\{\mu_i\}_{i=1}^p$ , and assume that the collection  $\{\nu\} \cup \{\mu_i\}_{i=1}^p$  satisfy assumptions **A4-A6**. We aim to show that this implies that  $\nu$  is a barycenter for the measures  $\{\mu_i\}_{i=1}^p$ .

Let  $\tilde{\mu}_i$  be the measure with density

$$d\tilde{\mu}_i(x) = \begin{cases} d\mu_i(x) & x \in \Omega^\circ, \\ 0 & \text{otherwise} \end{cases}$$

where  $\Omega^\circ$  is the interior of  $\Omega$ . Define  $\tilde{\nu}$  similarly.

We first remark that

$$1 = \mu_i[\Omega] = \mu_i[\Omega^\circ] + \mu_i[\partial\Omega] = \tilde{\mu}_i[\Omega^\circ] + 0 = \tilde{\mu}_i[\Omega^\circ]$$

and therefore  $\tilde{\mu}_i$  (and analogously  $\nu$ ) are valid measures (note we have used that  $\mu_i[\partial\Omega] = 0$ , which follows from  $\mu_i$  being a.c. and  $\partial\Omega$  having Lebesgue measure 0). It follows from **A4** that the collection  $\{\nu\} \cup \{\mu_i\}_{i=1}^p$  satisfies **A2** and **A3**. Furthermore, the set  $\Omega^\circ$  is open and convex and therefore **A1** is also satisfied. Therefore if  $\tilde{\nu}$  is a Karcher mean for  $\{\tilde{\mu}_i\}$  then it is a barycenter for them as well.

We next show that  $\tilde{\nu}$  is a Karcher mean of  $\{\tilde{\mu}_i\}_{i=1}^p$  with coordinate  $\lambda$ . Consider the maps  $\tilde{T}_i : \Omega^\circ \rightarrow \Omega^\circ$  which are defined by  $\tilde{T}_i(x) = T_i(x)$  and are undefined outside  $\Omega^\circ$  where  $T_i$  is the optimal transport map from  $\nu \rightarrow \mu_i$ . We first require that  $\tilde{T}_i$  is in fact optimal for  $\tilde{\nu}$  and  $\tilde{\mu}_i$  and that it is well-defined  $\tilde{\nu}$ -a.e.

To show the latter, note that  $T_i$  is a map from  $\nu$  to  $\mu_i$  and therefore  $\nu[T_i^{-1}(\partial\Omega)] = \mu_i[\partial\Omega] = 0$  which implies that  $T_i(x) \in \Omega^\circ$  for  $\nu$ -a.e.  $x$ . Furthermore  $\tilde{\nu}[\Omega^\circ] = \nu[\Omega^\circ] = 1$  and therefore the map  $\tilde{T}_i$  is well-defined  $\tilde{\nu}$ -a.e.

To show that the map  $\tilde{T}_i$  is optimal, one only needs to remark that for any measurable  $B \subset \Omega$  that  $T_i$  transports mass optimally between  $B$  and  $T_i(B)$  (since otherwise  $T_i$  would fail to be optimal). Applying this fact to  $\Omega^\circ$  and using the considerations above, shows that  $\tilde{T}_i$  is optimal for transporting between  $\tilde{\nu}$  and  $\tilde{\mu}_i$ .

Having established that the  $\tilde{T}_i$  are optimal, we can now calculate  $\|G_\lambda(\tilde{\nu})\|_{\tilde{\nu}}^2$ :

$$\begin{aligned} \|G_\lambda(\tilde{\nu})\|_{\tilde{\nu}}^2 &= \sum_{i,j=1}^p \lambda_i \lambda_j \int_{\Omega^\circ} \langle \tilde{T}_i(x) - x, \tilde{T}_j(x) - x \rangle d\tilde{\nu}(x) \\ &= \sum_{i,j=1}^p \lambda_i \lambda_j \int_{\Omega^\circ} \langle T_i(x) - x, T_j(x) - x \rangle d\nu(x) \\ &= \sum_{i,j=1}^p \lambda_i \lambda_j \int_{\Omega^\circ} \langle T_i(x) - x, T_j(x) - x \rangle d\nu(x) + \lambda_i \lambda_j \int_{\partial\Omega} \langle T_i(x) - x, T_j(x) - x \rangle d\nu(x) \\ &= \sum_{i,j=1}^p \lambda_i \lambda_j \int_{\Omega} \langle T_i(x) - x, T_j(x) - x \rangle d\nu(x) \\ &= \|G_\lambda(\nu)\|_{\nu}^2 = 0 \end{aligned}$$

which shows that  $\tilde{\nu}$  is a Karcher mean for  $\{\tilde{\mu}_i\}$ , and is also a barycenter of these measures. To conclude, we only require the fact that if two measures  $\xi, \tilde{\xi} \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  differ only on a set of measure zero, then for any  $\gamma \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$  we have  $W_2(\xi, \gamma) = W_2(\tilde{\xi}, \gamma)$ . Applying this fact repeatedly we have

$$\begin{aligned} \min_{\xi} \frac{1}{2} \sum_{i=1}^n \lambda_i W_2^2(\xi, \mu_i) &= \min_{\xi} \frac{1}{2} \sum_{i=1}^n \lambda_i W_2^2(\xi, \tilde{\mu}_i) \\ &= \frac{1}{2} \sum_{i=1}^n \lambda_i W_2^2(\tilde{\nu}, \tilde{\mu}_i) \\ &= \frac{1}{2} \sum_{i=1}^n \lambda_i W_2^2(\nu, \tilde{\mu}_i) \\ &= \frac{1}{2} \sum_{i=1}^n \lambda_i W_2^2(\nu, \mu_i) \end{aligned}$$

which shows that  $\nu$  is indeed a barycenter of  $\mu_i$ , as desired.

### 3. Covariance Estimation Experiment Details

#### 3.1. Methods for Maximum Likelihood Estimation

In order to solve the maximum likelihood estimation problem we differentiate through a truncated version of (Chewi et al., 2020) Algorithm 1 and perform projected gradient descent. This is implemented using the auto-differentiation library

**Algorithm 3** MLE

---

**Input:**  $\{S_i\}_{i=0}^p$ ,  $\eta > 0$ ,  $\text{MaxIters} > 0$ ,  $\text{FPIters} > 0$ ,  $\text{SQIters} > 0$   
 $i \leftarrow 0$ ,  $\lambda \leftarrow (1/p)\mathbf{1}$   
**while** Not Converged and  $i < \text{MaxIters}$  **do**  
      $i \leftarrow i + 1$   
      $\nabla \mathcal{L}(\lambda) \leftarrow \text{BackPropLoss}(\{S_j\}_{j=0}^p, \lambda, \text{FPIters}, \text{SQIters})$   
      $\lambda \leftarrow \text{SimplexProject}(\lambda - \eta \nabla \mathcal{L}(\lambda))$   
**end while**  
**Return**  $\lambda$

---

Parameter	Value
$\eta$	0.0003
MaxIters	500
FPIters	10
SQIters	10

Table 3: Parameters used when performing the maximum likelihood estimation.

PyTorch. The procedure is summarized in Algorithm 3. Here  $\text{SimplexProject}(x)$  is defined as

$$\text{SimplexProject}(x) = \arg \min_{y \in \Delta^p} \|x - y\|_2.$$

which enforces the constraints on  $\lambda$  at each iteration.

To compute  $\nabla \mathcal{L}(\lambda)$  we use auto-differentiation to obtain a gradient of the procedure given in Algorithm 4. In this procedure the square root of a matrix is computed using SQIters number of Newton-Schulz iterations. The forward pass of the loss computation is given in Algorithm 4, and the gradient is obtained by back propagation through it.

**Algorithm 4** ComputeLoss

---

**Input:**  $\{S_i\}_{i=0}^p$ ,  $\lambda$ ,  $\text{FPIters} > 0$ ,  $\text{SQIters} > 0$   
 $\text{BC} \leftarrow S_0$   
**for**  $j = 1, \dots, \text{FPIters}$  **do**  
      $\text{BC\_root} = \text{SquareRoot}(\text{BC}, \text{SQIters})$   
      $\text{BC\_root\_inv} = \text{Invert}(\text{BC\_root})$   
      $\text{BC} = \text{BC\_root\_inv} (\sum_{i=1}^p \lambda_i \text{SquareRoot}(\text{BC\_root}, \text{SQIters}, S_i \text{BC\_root})) \text{BC\_root\_inv}$   
**end for**  
**Return**  $\text{Tr}(\text{BC}^{-1} S_0) + \log \det \text{BC}$

---

The parameters that we chose in our experiments are given in Table 3. These parameters were sufficient to ensure that both the matrix square roots and fixed point iterations converged, and  $\lambda$  always converged before the final iteration was reached.

### 3.2. Considerations for Instability of MLE

Due to the requirement of differentiating through a matrix inverse in Algorithm 3 it is possible that the procedure may fail to numerically converge. In our trials, this happened in less than 0.5% of all attempts. When it does happen we discard the result, and these numerical failures impact neither the results of Figure 3, nor the timings of the trials.

## 4. MNIST Experiment Details

### 4.1. The MNIST Dataset

The MNIST Dataset (LeCun, 1998) consists of a collection of  $28 \times 28$  black and white images of hand-written digits. The brightness (or intensity) of each pixel is an integer between 0 and 255 with 0 being black and 255 being white.



Each image is initially a matrix  $I^m \in \mathbb{R}_+^{28 \times 28}$ , which is normalized to sum to one,  $P_{ij}^m = I_{ij}^m / (\sum_{i',j'} I_{i',j'}^m)$  and then converted into a point cloud

$$\mu^m = \sum_{i,j=1}^{28} P_{ij}^m \delta_{(i,j)}.$$

#### 4.2. White Noise Model

In order to apply additive noise we must generate white-noise images  $\zeta$ . We generate  $\zeta$  as a random matrix  $\mathbb{R}^{28 \times 28}$  with each entry  $\zeta_{ij} \sim \text{Unif}([0, 1])$ . This random matrix is then converted into a point cloud using the same process as is used to turn the image matrix into a point cloud.

#### 4.3. Linear Recovery

The linear reconstruction in Figure 4 method is done by the following procedure.

1. Given a corrupted image (which we will treat as a matrix)  $\tilde{\mu}_0$  and a collection of references  $\tilde{\mu}_1, \dots, \tilde{\mu}_p$  (also treated like matrices) estimate  $\hat{\lambda}$  as

$$\hat{\lambda} = \arg \min_{\lambda \in \Delta^p} \|\tilde{\mu}_0 - \sum_{i=1}^p \lambda_i \tilde{\mu}_i\|_F^2.$$

This corresponds to projecting  $\tilde{\mu}_0$  onto the convex hull of  $\tilde{\mu}_1, \dots, \tilde{\mu}_p$  with respect to the Euclidean (equivalently Frobenius) norm.

2. Recover  $\hat{\mu}_0$  as

$$\hat{\mu}_0 = \sum_{i=1}^p \hat{\lambda}_i \mu_i.$$

### 5. Document Classification Experiment Details

#### 5.1. Datasets

Dataset	Description	B.O.W. Dimension	Average Words	$\mathcal{C}$
<b>BBCSPORT</b>	BBC sports articles labeled by sport	13243	117	5
<b>20NEWS</b>	canonical news article dataset	29671	72	20

Table 4: Dataset characteristics.

We show the characteristics of the datasets in Table 4. Bag of words (B.O.W.) dimension is the number of the unique words contained in each dataset. Average words is the average number of the unique words in a document and  $\mathcal{C}$  is the number of classes of each of the datasets.

Both the BBCSport and News20 datasets that we used are those made publicly available from (Huang et al., 2016). Both datasets consist of two components:

1. The datasets employ a word2vec (Mikolov et al., 2013) embedding to represent the words. Each word is represented as a point  $w$  on the unit sphere in  $\mathbb{R}^{300}$  and the semantic relationships between words are encoded in the geometry of the embeddings. Call the set of words  $\mathcal{W}$ , then the word2vec embedding is the matrix  $W \in \mathbb{R}^{|\mathcal{W}| \times 300}$ .
2. Each document  $D^m$  with label  $y^m$  is represented by a set of tuples  $\{(c_i^m, w_i^m)\}_{i=1}^n$ , where  $n$  is the number unique words in  $D^m$ ,  $c_i^m$  is the number of times  $w_i^m$  shows in  $D^m$ .

We use the following procedure to convert a document into an empirical measure.

1. Given a document  $D^m = \{(c_i^m, w_i^m)\}_{i=1}^n$ , let  $p^m \in \mathbb{R}^n$  be the probability vector with entries  $p_i^m = c_i^m / (\sum_k c_k^m)$ .

2. The document empirical measure is then

$$\mu^m = \sum_i p_i^m \delta_{w_i^m}.$$

For convenience we filter out the points on which  $p^i$  places no support.

## 5.2. Procedure for Making Figure 5

As discussed in the main body, for  $k \geq 2$  we select  $k$  documents from each topic and 100 test documents, and then repeat 50 times and average. In both plots in Figure 5 we iterate over different choices of  $k$ . To reduce computational load we reuse the test sets and reference documents within each trial across choices of  $k$ . For example, all of the documents for  $k = 4$  are used for  $k \geq 5$ . This allows the Wasserstein distances and estimates  $\hat{A}_{ij}$  to be cached avoiding further computation.

We note that this procedure produces plots in which each of the points individually has the same sampling pattern as if there was no re-use, it is only across samples that there is correlation. However, this correlation is shared across the prediction methods consider which preserves the fairness of the comparison of the methods.

## 6. Possible Rank Deficiency of $A$

In general, there may exist a set  $\Lambda \subset \Delta^p$  such that  $|\Lambda| > 1$  and for all  $\lambda \in \Lambda$  we have  $\mu_0 = \nu_\lambda$ . Indeed, suppose there are index sets  $I_1, I_2$  such that  $I_1 \cap I_2 = \emptyset$  and  $I_1 \cup I_2 = \{1, 2, \dots, p\}$  but  $\text{Bary}(\{\mu_i\}_{i \in I_1}) \cap \text{Bary}(\{\mu_i\}_{i \in I_2}) \neq \emptyset$ . If  $\mu_0 \in \text{Bary}(\{\mu_i\}_{i \in I_1}) \cap \text{Bary}(\{\mu_i\}_{i \in I_2})$  then it can be expressed as  $\nu_{\lambda^1} \in \text{Bary}(\{\mu_i\}_{i \in I_1})$  and  $\nu_{\lambda^2} \in \text{Bary}(\{\mu_i\}_{i \in I_2})$ . Note that we may interpret  $\lambda^1$  and  $\lambda^2$  as vectors in  $\Delta^p$  with supports on  $I_1, I_2$ , respectively. Now let  $\alpha \in [0, 1]$  and let  $\lambda = (1 - \alpha)\lambda^1 + \alpha\lambda^2$  (with the indices modified as needed). Then we have

$$\arg \min_{\nu \in \mathcal{P}_{2, \text{ac}}(\mathbb{R}^d)} \sum_{i=1}^p \frac{\lambda_i}{2} W_2^2(\nu, \mu_i) = \arg \min_{\nu \in \mathcal{P}_{2, \text{ac}}(\mathbb{R}^d)} (1 - \alpha) \left( \sum_{i \in I_1} \frac{\lambda_i^1}{2} W_2^2(\nu, \mu_i) \right) + \alpha \left( \sum_{i \in I_2} \frac{\lambda_i^2}{2} W_2^2(\nu, \mu_i) \right) = \mu_0$$

where the last equality follows from the fact that if  $x$  minimizes both  $f(x)$  and  $g(x)$  then it must also minimize  $\alpha f(x) + (1 - \alpha)g(x)$  for  $\alpha \in [0, 1]$ , and the assumption that  $\mu_0$  is the barycenter. This demonstrates that  $\mu_0$  actually corresponds to all  $\lambda = (1 - \alpha)\lambda^1 + \alpha\lambda^2$  for all  $\alpha \in [0, 1]$ .

Below, we will give a geometric characterization for the rank deficiency of  $A$  and how that relates to the optimality of the quadratic objective in Theorem 1. We first recall the quadratic problem over the probability simplex:  $\min_{\lambda \in \Delta^p} \lambda^T A \lambda$ . Since  $A$  is a symmetric and positive semidefinite matrix, a Rayleigh quotient analysis yields that the minimum possible value of the objective without simplex constraints is zero, which would be attained at an eigenvector corresponding to a zero eigenvalue. Given this fact, finding an exact solution to the above quadratic problem amounts to determining whether the probability simplex intersects the eigenspace of  $A$  associated with the zero eigenvalue. Hereafter, we will denote this space by  $E_0$ .

**Proposition 2.** *If  $E_0 \cap \Delta^p = \emptyset$ , there is no  $\lambda \in \Delta^p$  such that  $\lambda^T A \lambda = 0$ .*

*Proof.* Any  $\lambda \in \Delta^p$  realizing  $\lambda^T A \lambda = 0$  is an eigenvector corresponding to the zero eigenvalue. The result follows.  $\square$

Proposition 2 shows in particular that exact minimization is only possible when  $\text{rank}(A) < p$ . Next, we state conditions under which there is a unique solution to the minimization program.

**Proposition 3.** *If  $\text{rank}(A) = p - 1$  and  $E_0 \cap \Delta^p \neq \emptyset$ , then there is a unique  $\lambda \in \Delta^p$  such that  $\lambda^T A \lambda = 0$ .*

*Proof.* If  $\text{rank}(A) = p - 1$ ,  $E_0 = \{\alpha v \mid \alpha \in \mathbb{R}\}$  for an eigenvector  $v \neq 0$  corresponding to the zero eigenvalue. Let  $\lambda \in E_0 \cap \Delta^p$ . Then  $\lambda = \alpha v$  for some  $\alpha \in \mathbb{R}$ . The constraint that  $\lambda \in \Delta^p$  determines  $\alpha$  uniquely, which gives the result.  $\square$

Note that if  $E_0$  non-trivially intersects the probability simplex and  $\text{rank}(A) = p - 1$ , the condition for uniqueness is equivalent to the statement that the origin can be written as a convex combination of the columns of  $A$ . We next consider the remaining case when  $\text{rank}(A) < p - 1$ ; for simplicity of exposition, we focus on the specific case  $\text{rank}(A) = p - 2$ .

**Proposition 4.** *Let  $\text{rank}(A) = p - 2$  and  $v_0$  and  $v_1$  be distinct eigenvectors corresponding to the zero eigenvalue of  $A$ . If  $v_0 \in \Delta^p$  and  $v_1 \in \Delta^p$ , then there are infinitely many solutions to  $\lambda^T A \lambda = 0$  with  $\lambda \in \Delta^p$ .*

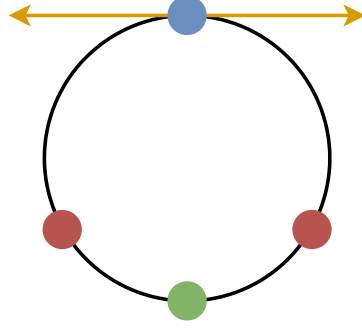


Figure 6: *Red*: The points  $z_1, z_2$ . *Green*: The true barycenter  $x_\lambda$ . *Blue*: A Karcher mean which is not the barycenter. *Gold*: The tangent space at the blue point.

*Proof.* By construction,  $Av_0 = 0$  and  $Av_1 = 0$ . Let  $\alpha \in [0, 1]$  and define  $v_\alpha = \alpha v_0 + (1 - \alpha)v_1$ . Since  $v_1$  and  $v_2$  are in  $\Delta^p$ , it follows that  $v_\alpha \in \Delta^p$ . In addition,  $Av_\alpha = \alpha Av_1 + (1 - \alpha)Av_2 = 0$  so that  $v_\alpha \in E_0$ . Since  $\alpha \in [0, 1]$  was arbitrary, the result follows.  $\square$

The main conclusion from the above Propositions is the following: if  $\text{rank}(A) < p - 1$ , the only way to mandate a unique solution is to require that only one eigenvector corresponding to the zero eigenvalue intersects with probability simplex. By convex geometry, if two eigenvectors corresponding to the zero eigenvalue intersect the probability simplex, there will be infinitely many solutions to the quadratic program.

## 7. Example of a Karcher Mean that is not a Barycenter

As mentioned in Section 2, special care must be made to ensure that any Karcher mean is also a barycenter. In this section we give a simple example which illustrates this point.

Consider  $S^1 = \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$  with distance given by  $\rho(x, y) = \cos^{-1}\langle x, y \rangle$ . Let  $z_1 \neq z_2 \in S^1$  with  $z_1 \neq -z_2$ , and  $\lambda = (1/2, 1/2)$ . Then the barycenter is given by

$$x_\lambda = \frac{z_1 + z_2}{\|z_1 + z_2\|_2}.$$

However, the point  $-\frac{z_1 + z_2}{\|z_1 + z_2\|_2}$  is not the barycenter but it is a Karcher mean. This is illustrated in Figure 6. To show this rigorously one need only compute the Euclidean gradient and demonstrate that it is orthogonal to the tangent space.