# Linearized Wasserstein Barycenters: Synthesis, Analysis, Representational Capacity, and Applications

**Matthew Werenski**[(†)], **Brendan Mallery**[(‡)], **Shuchin Aeron** [(□,*)], **James M. Murphy**[(‡,*)]
Tufts University

## Abstract

We propose the linear barycentric coding model (LBCM) which utilizes the linear optimal transport (LOT) metric for analysis and synthesis of probability measures. We provide a closed-form solution to the variational problem characterizing the probability measures in the LBCM and establish equivalence of the LBCM to the set of 2-Wasserstein barycenters in the special case of compatible measures. Computational methods for synthesizing and analyzing measures in the LBCM are developed with finite sample guarantees. One of our main theoretical contributions is to identify an LBCM, expressed in terms of a simple family, which is sufficient to express all probability measures on the closed unit interval. We show that a natural analogous construction of an LBCM in 2 dimensions fails, and we leave it as an open problem to identify the proper extension in more than 1 dimension. We conclude by demonstrating the utility of LBCM for covariance estimation and data imputation.

## 1 INTRODUCTION

Interpreting data as probability measures and embedding them into spaces equipped with an appropriate metric (or more generally, discrepancy) to allow for meaningful comparisons has emerged as an important paradigm in computer vision (Bonneel and Digne, 2023), natural language processing (Xu et al., 2018), high-energy physics (Komiske et al., 2019; Cai et al., 2022), time-series analysis (Cheng et al., 2021, 2023), and seismic imaging (Métivier et al., 2016) among other areas. In the space of probability measures, one can utilize the discrepancy used for comparisons to develop low-parameter and parsimonious models for data. In this paper, we consider a class of models that we term *barycentric coding models (BCM)* (Werenski et al., 2022). These models allow for *synthesis* and *analysis* of measures through variational problems with respect to a given set of *reference measures*, $\{\mu_i\}_{i=1}^m$, and non-negative coefficients collected in the vector $\lambda$ belonging to the $(m-1)$-dimensional simplex $\Delta^m$. Synthesis takes the form:

$$\mu_\lambda \triangleq \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^m \lambda_i D_s(\nu, \mu_i) + \mathcal{R}(\nu), \tag{1}$$

where $D_s$ is a discrepancy and $\mathcal{R}$ is a regularizer enforcing structural properties on $\mu_\lambda$. We call the collection of measures generated as in (1) a BCM. One can also project a measure $\eta$ onto the collection of measures of the form (1) via solving the *analysis problem*

$$\lambda_\eta = \operatorname*{arg\,min}_{\lambda \in \Delta^m} D_a(\eta, \mu_\lambda), \tag{2}$$

where $D_a$ is a discrepancy, not necessarily the same as $D_s$. We consider $D_s, D_a$ that are based on optimal transport (Santambrogio, 2015; Peyré and Cuturi, 2020) to impute the geometry of the space of probability measures onto our low-dimensional models.

Specifically, we consider the case when $D_s = D_a$ are given by a *linear optimal transport (LOT) metric* (Wang et al., 2013). Given an absolutely continuous *base measure* $\mu_0$, the (squared) LOT metric between measures $\mu, \nu$ is

$$\|T_{\mu_0}^\mu - T_{\mu_0}^\nu\|_{L^2(\mu_0)}^2 = \int \|T_{\mu_0}^\mu(x) - T_{\mu_0}^\nu(x)\|_2^2 d\mu_0(x),$$

where for any measure $\eta$, $T_{\mu_0}^\eta$ is the optimal transport map from $\mu_0$ to $\eta$ with respect to the squared-Euclidean ground cost (see Section 2). We call the set of all measures that can be represented this way the *linear barycentric coding model (LBCM):*

[(†)] Dept. of Computer Science, [(‡)] Dept. of Mathematics, [(□)] Dept. of Electrical and Computer Engineering. [(*)] co-last authors. Correspondence to: Matthew.Werenski@tufts.edu.
Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. 

**Definition 1.** Let $\{\mu_i\}_{i=1}^m$ be reference measures and $\mu_0$ a base measure. The associated *linear barycentric coding model (LBCM)* is

$$\text{LBCM}(\{\mu_i\}_{i=1}^m, \mu_0) \triangleq \{\mu_\lambda \mid \lambda \in \Delta^m\},$$
$$\mu_\lambda \triangleq \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^m \lambda_i \|T_{\mu_0}^\nu - T_{\mu_0}^{\mu_i}\|_{L^2(\mu_0)}^2. \tag{3}$$

### 1.1 Summary of Contributions

This paper focuses on modeling data with the LBCM.

1. Proposition 1 establishes that solutions to (3) are of the form:
$$\mu_\lambda = \left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right) \# \mu_0.$$
Proposition 2 proves that if the measures $\{\mu_0\}_{i=1}^m$ satisfy a particular compatibility condition, the LBCM is equivalent to the BCM with $D_s = D_a = W_2^2$ the squared 2-Wasserstein metric (Werenski et al., 2022).
2. We show that the analysis problem,
$$\lambda_\eta = \operatorname*{arg\,min}_{\lambda \in \Delta^m} \left\| \sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i} - T_{\mu_0}^{\eta} \right\|_{L^2(\mu_0)}^2,$$
can be efficiently solved via a quadratic program when $\eta \in \text{LBCM}(\{\mu_i\}_{i=1}^m, \mu_0)$. Leveraging recent results in entropy-regularized optimal transport, we establish rates of statistical estimation when the measures are available via samples.
3. In Section 5, we propose a novel problem of representational capacity in the space of probability measures. In Theorem 4 we show that the set of reference measures supported on $\{a, b\}$ with $\mu_0$ as the uniform measure on $[a, b]$ has dense LBCM in $\mathcal{P}([a, b])$ with respect to weak convergence. We then establish in Theorem 5 that a natural generalization of our 1-dimensional result fails in dimension 2, due to the increased complexity of the space of transport maps.
4. We provide experiments to show the utility of the LBCM model for covariance estimation and image reconstruction. We give both qualitative and quantitative comparisons with a $W_2$ variant of the BCM and linear mixtures, and show that reconstruction with the LBCM gives comparable results while being significantly faster than other methods. We also investigate how changing the base measure affects the quality of reconstruction.

All proofs are deferred to the Appendix.

## 2 BACKGROUND

For a vector $x \in \mathbb{R}^d$ (or random variable or function), we let $x_i$ denote its $i^{th}$ coordinate. For $t \in \mathbb{R}$, $\exp_\varepsilon(t) \triangleq \exp(t/\varepsilon)$. The set of probability measures in $\mathbb{R}^d$ will be denoted by $\mathcal{P}(\mathbb{R}^d)$. The set of measures which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ will be denoted $\mathcal{P}_{ac}(\mathbb{R}^d)$. For $\Omega \subseteq \mathbb{R}^d$ we let $\mathcal{P}(\Omega), \mathcal{P}_{ac}(\Omega)$ respectively be the set of measures $\mu$ in $\mathcal{P}(\mathbb{R}^d), \mathcal{P}_{ac}(\mathbb{R}^d)$ such that $\mu[\Omega] = 1$. Elements of $\mathcal{P}(\Omega)$ and $\mathcal{P}_{ac}(\Omega)$ with finite second moment are denoted $\mathcal{P}_2(\mathbb{R}^d)$ and $\mathcal{P}_{2,ac}(\mathbb{R}^d)$, respectively. For a convex body $C$ in $\mathbb{R}^d$ (i.e., $C$ is convex, compact, and has a non-empty interior) we will denote the uniform measure over $C$ by $U_C$. Note that we always have $U_C \in \mathcal{P}_{ac}(C)$. For any collection of elements $\{v_i\}_{i \in I}$ of a vector space with real scalars, we let $\text{conv}(\{v_i\}_{i \in I})$ denote the *convex hull* consisting of all convex combination (i.e., linear mixtures) of elements of $\{v_i\}_{i \in I}$. We use $\lesssim$ to specify inequalities which hold up to multiplicative constants.

**Definition 2.** Let $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. The *2-Wasserstein metric* between $\mu$ and $\nu$ is

$$W_2^2(\mu, \nu) \triangleq \min_{T \# \mu = \nu} \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 d\mu(x). \tag{4}$$

We call the $T_\mu^\nu$ realizing (4) the 2-Wasserstein *optimal transport map* from $\mu$ to $\nu$.

The use of a minimum rather than an infimum in (4) is justified in our setting by the following seminal result (Brenier, 1991; Gangbo and McCann, 1996; McCann, 1997).

**Theorem 1** (Brenier). *Let $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then there exists a unique (up to $\mu$-a.e. equivalence) map $T^*$ such $T^* \# \mu = \nu$, and*

$$\min_{T \# \mu = \nu} \int_{\mathbb{R}^d} ||T(x) - x||_2^2 d\mu(x) = \int_{\mathbb{R}^d} ||T^*(x) - x||_2^2 d\mu(x).$$

*Moreover, there exists a convex function $\varphi$ such that $\nabla \varphi = T^*$.*

Note, one can extend $W_2(\mu, \nu)$ to $\mu$ that are not absolutely continuous by optimizing over couplings rather than maps (Santambrogio, 2015).

### 2.1 Related Work

Many different barycentric models of the form (1) have been considered for synthesis; see Table 1. On the other hand, the analysis problem has received relatively little attention (Bonneel et al., 2016; Werenski et al., 2022; Blickhan, 2024; Gunsilius et al., 2024). The most closely related work to ours is (Werenski et al., 2022) where one works with the BCM model with $D_s = D_a = W_2^2$, which we refer to as the $W_2$BCM.

**Definition 3.** Let $\{\mu_i\}_{i=1}^m$ be reference measures. The associated *$W_2$-barycentric coding model ($W_2$BCM)* is

$$W_2\text{BCM}(\{\mu_i\}_{i=1}^m) \triangleq \{\mu_\lambda \mid \lambda \in \Delta^m\}, \tag{5}$$
$$\mu_\lambda \triangleq \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^m \lambda_i W_2^2(\nu, \mu_i).$$

Under some regularity assumptions on the reference measures and $\mu$, determining the coordinates $\lambda \in \Delta^m$ associated with $\mu \in W_2\text{BCM}(\{\mu_i\}_{i=1}^m)$ can be done via a quadratic program whose parameters depend on $\mu$ and $\{\mu_i\}_{i=1}^m$ (see Theorem 1 in (Werenski et al., 2022)). When measures are observed via samples, plug-in estimates of OT maps can be used for the QP formulation with guarantees on error rates (see Corollary 2 in (Werenski et al., 2022)). We prove similar results for the LBCM in Section 4.

In the context of high computational complexity and curse of dimensionality in statistical error rates for estimating quantities related to $W_2$, (Mallery et al., 2025) considers entropy-regularized optimal transport for analysis and synthesis of measures.

We note other parsimonious models in the space of probability measures are possible including geodesic PCA (Bigot et al., 2017; Cazelles et al., 2018) and log-PCA (Fletcher et al., 2004; Sommer et al., 2010). In this paper we are concerned with analysis and synthesis given the LBCM model, but the problem of identifying a BCM model given data assumed to be generated by the BCM have also been considered (Schmitz et al., 2018; Cheng et al., 2021; Mueller et al., 2023; Cheng et al., 2023).

The use of transport maps in the LBCM is motivated by *linearized optimal transport*, in which for a chosen base measure $\mu_0$, one approximates $W_2(\mu, \nu) \approx \|T_{\mu_0}^\mu - T_{\mu_0}^\nu\|_{L^2(\mu_0)}$ (Wang et al., 2013). The embedding of measures $\mu, \nu$ into $L^2(\mu_0)$ via $\mu \mapsto T_{\mu_0}^\mu$ is not an isometry in general. However, upper and lower bounds between $W_2(\mu, \nu)$ and $\|T_{\mu_0}^\mu - T_{\mu_0}^\nu\|_{L^2(\mu_0)}$ exist under suitable regularity conditions on the measures considered (Delalande and Merigot, 2023). Conditions guaranteeing an exact isometry are closely related to the notion of compatible measures described in Definition 4 of (Moosmüller and Cloninger, 2023).

The recent paper (Jiang et al., 2023) proposes the use of LBCM sets as a regularizer for variational inference. Therein is proved a quantitative approximation result for the number of reference measures required to estimate a sufficiently regular optimal transport map in a 1-dimensional LBCM. Our results in Section 5 may be compared to this, albeit our result does not require smoothness of the transport maps and shows that the reference measures can be taken to have singular support. Moreover, we investigate the multidimensional setting and identify interesting problems therein.

## 3 PROPERTIES OF THE LBCM

The following result asserting that the variational form of the LBCM has a closed form is stated informally in (Mérigot et al., 2020); we give a proof in Appendix A for completeness.

**Proposition 1.** *For $\mu_0, \{\mu_i\}_{i=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ and $\lambda \in \Delta^m$, the optimization problem*

$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^m \lambda_i \|T_{\mu_0}^\nu - T_{\mu_0}^{\mu_i}\|_{L^2(\mu_0)}^2 \tag{6}$$

*has solution $\left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right) \#\mu_0$.*

Proposition 1 demonstrates that the LBCM is similar in spirit to generative model for data associated to nonnegative matrix factorization (NMF) (Lee and Seung, 2000) and archetypal analysis (Cutler and Breiman, 1994), in which observations are modeled as non-negative (or convex) combinations of generators. In the case of the LBCM, the convex combination of maps are then used to pushforward the base measure $\mu_0$.

While they differ in general, the LBCM is equal to the $W_2$BCM for a common set of reference measures in the special case in which the underlying measures are *compatible* (Panaretos and Zemel, 2020).

**Definition 4.** Measures $\{\nu_i\}_{i=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ are *compatible* if for any $\nu_i, \nu_j, \nu_k$ the optimal transport maps satisfy $T_{\nu_i}^{\nu_j} = T_{\nu_k}^{\nu_j} \circ T_{\nu_i}^{\nu_k}$.

Compatibility of measures is a strong condition, but does hold in several important cases, including: (i) all absolutely continuous 1-dimensional probability measures; (ii) Gaussian measures with simultaneously diagonalizable covariance matrices ; (iii) measures that are translation-dilations of a common measure. We refer to (Panaretos and Zemel, 2020) for further discussion of compatibility. The following equivalence result follows from several known results on multimarginal optimal transport; a proof is given in Appendix A for completeness.

**Proposition 2.** *Suppose $\{\mu_i\}_{i=0}^m$ are compatible. Then*

$$\text{LBCM}(\mu_0; \{\mu_i\}_{i=1}^m) = W_2\text{BCM}(\{\mu_i\}_{i=1}^m),$$

*and $\nu_\lambda$ is the same in both sets.*

## 4 SYNTHESIS AND ANALYSIS IN THE LBCM

Fix base and reference measures $\{\mu_i\}_{i=0}^m$. The *synthesis problem* in the LBCM is, via Proposition 1, to compute

| Barycenter | $D(\nu, \mu_i)$ (Discrepancy) | $\mathcal{R}(\nu)$ (Regularizer) |
|---|---|---|
| Linear | $\int \left\lvert \frac{d\nu}{dx}(x) - \frac{d\mu_i}{dx}(x) \right\rvert^2 dx$ | none |
| 2-Wasserstein (Agueh and Carlier, 2011) | $W_2^2(\nu, \mu_i)$ | none |
| Inner-Regularized Entropic (Janati et al., 2020; Mallery et al., 2025) | $S_\varepsilon(\nu, \mu_i)$ | none |
| Outer-Regularized Entropic (Bigot et al., 2019; Carlier et al., 2021) | $W_2^2(\nu, \mu_i)$ | $\int \log\left(\frac{d\nu}{dx}(x)\right) d\nu(x)$ |
| Doubly-Regularized (Chizat, 2023; Vaskevicius and Chizat, 2024) | $S_\varepsilon(\nu, \mu_i)$ | $\int \log\left(\frac{d\nu}{dx}(x)\right) d\nu(x)$ |
| Sinkhorn (Janati et al., 2020; Mallery et al., 2025) | $S_\varepsilon(\nu, \mu_i) - \frac{1}{2}S_\varepsilon(\nu, \nu) - \frac{1}{2}S_\varepsilon(\mu_i, \mu_i)$ | none |
| Linearized 2-Wasserstein ((Mérigot et al., 2020), this paper) | $\int \|T_{\mu_0}^{\nu}(x) - T_{\mu_0}^{\mu_i}(x)\|_2^2 d\mu_0(x)$ | none |

Table 1: Summary of methods to synthesize probability measures in the framework of (1).

$\left(\sum_{i=1}^m T_{\mu_0}^{\mu_i}\right) \# \mu_0$. For a new measure $\eta$ the *analysis problem* in the LBCM is to find coefficients that optimally represent $T_{\mu_0}^{\eta}$ as a convex combination of the maps $\{T_{\mu_0}^{\mu_i}\}_{i=1}^m$. Specifically, it solves the optimization:

$$\min_{\lambda \in \Delta^m} \left\| \sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i} - T_{\mu_0}^{\eta} \right\|_{L^2(\mu_0)}^2.$$

One can equivalently write this as

$$\min_{\lambda \in \Delta^m} \lambda^T A^L \lambda, \tag{7}$$

where $A_{ij}^L = \int_{\mathbb{R}^d} \left\langle T_{\mu_0}^{\mu_i} - T_{\mu_0}^{\eta}, T_{\mu_0}^{\mu_j} - T_{\mu_0}^{\eta} \right\rangle d\mu_0$. This reduces the problem to solving a convex quadratic program.

**Remark 1.** We note that the analysis problem for the $W_2$BCM can be solved in an analogous manner *if the underlying measure is exactly a 2-Wasserstein barycenter of the reference measures*. More precisely, given $\eta$ and reference measures $\{\mu_i\}_{i=1}^m$, we formulate the analysis problem for the $W_2$BCM (Werenski et al., 2022) as

$$\min_{\lambda \in \Delta^m} W_2^2(\nu_\lambda, \eta), \quad \nu_\lambda = \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^m \lambda_i W_2^2(\nu, \mu_i).$$

If there exists a unique $\lambda^* \in \Delta^m$ such that $\eta = \nu_\lambda$, then (Werenski et al., 2022) proves that, under certain assumptions on $\eta$ and $\{\mu_i\}_{i=1}^m$ (Caffarelli, 1992; Panaretos and Zemel, 2020) the coefficients $\lambda^*$ satisfy the optimization problem

$$\lambda^* = \operatorname*{arg\,min}_{\lambda \in \Delta^m} \lambda^T A \lambda, \tag{8}$$

where $A_{ij} = \int_{\mathbb{R}^d} \left\langle T_\eta^{\mu_i} - \mathrm{Id}, T_\eta^{\mu_j} - \mathrm{Id} \right\rangle d\eta$.

### 4.1 Plug-in Estimation for the Synthesis and Analysis

In order to solve (7) or (8) when measures are only accessible via samples, the key idea is to use a sample-based plug-in estimator for the necessary optimal transport maps and estimate the inner products via Monte-Carlo. We employ the well-known entropy-regularized map (Pooladian and Niles-Weed, 2021; Rigollet and Stromme, 2022; Stromme, 2023; Masud et al., 2023; Werenski et al., 2023, 2024), denoted $T_\mu^{\nu,\varepsilon}$, obtained via entropy-regularized optimal transport (EOT) as the plug-in estimator of the OT maps due to computational efficiency of EOT (Sinkhorn and Knopp, 1967; Cuturi, 2013). We refer to Appendix B for necessary background on EOT and (Nutz, 2021) for more details. Its sample version $T_\mu^{\nu,n,\varepsilon}$ estimates $T_\mu^\nu$ with statistical rates that depend exponentially on the data dimension (Pooladian and Niles-Weed, 2021). Estimating each optimal transport map $T_{\mu_i}^\nu, i = 1, \ldots, m$, allows us to estimate $A_{ij}^L$, which requires estimating the inner product of two OT maps. To do so, we follow the framework of (Pooladian and Niles-Weed, 2021) and place Assumptions 1-3 on all measures $\{\mu_i\}_{i=0}^m$.

**Assumption 1.** The measures $\mu_i$ have compact convex support $\Omega$ and they are absolutely continuous on this set with densities $p_i$ which satisfy $0 < m < p_i < M$ for two constants $m, M$.

**Assumption 2.** For $i = 1, ..., m$ let $\varphi_i, \varphi_i^*$ be the optimal potentials between $\mu_0$ and $\mu_i$ as defined in Theorem 1. Then $\varphi_i \in \mathcal{C}^2(\Omega)$ and $\varphi_i^* \in \mathcal{C}^{\alpha+1}(\Omega)$ for some $\alpha > 1$.

**Assumption 3.** There exist $l, L > 0$ with $lI \preceq \nabla^2 \varphi_i(x) \preceq LI$ for all $x \in \Omega$.

**Theorem 2** ((Pooladian and Niles-Weed, 2021), Theorem 3)**.** *Suppose that $\mu$ and $\nu$ satisfy Assumptions 1, 2, and 3. Then the plug-in estimate of the entropy-regularized map $T_\mu^{\nu,n,\varepsilon}$, with regularization parameter $\varepsilon \asymp n^{-1/(d+\bar{\alpha}+1)}$, satisfies*

$$\mathbb{E}\left[ \left\| T_\mu^{\nu,n,\varepsilon} - T_\mu^\nu \right\|_{L^2(\mu)}^2 \right] \lesssim n^{-\frac{(\bar{\alpha}+1)}{2(d+\bar{\alpha}+1)}} \log n \tag{9}$$

*where $\bar{\alpha} = \min(\alpha, 3)$. The implicit constant may depend on $\mu$ and $\nu$, but does not depend on $n$.*

**Remark 2.** The use of Theorem 2 imposes strong assumptions on $\mu, \nu$ and requires that the integrated

Fisher information along the 2-Wasserstein geodesic between them be finite. Other estimators for $T_\mu^\nu$ exist that offer improved rates of convergence under smoothness assumptions (Deb et al., 2021; Manole et al., 2024) and may be used as plug-in estimators.

Once these maps are estimated, we can proceed to solve the synthesis and analysis problems for the LBCM as follows.

**Proposition 3.** *Assume that $\mu_0$ and $\{\mu_i\}_{i=1}^m$ satisfy Assumptions 1, 2, and 3. For each $i$, let $T_{\mu_0}^{\mu_i,n,\varepsilon}$ be the entropic map between $\mu_0^n$ and $\mu_i^n$, where $\mu_0^n$ and $\mu_i^n$ are $n$ sample empirical measures from $\mu_0$ and $\mu_i$, respectively. Define $\nu = \left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right)\#\mu_0$ and $\nu^n = \left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i,n,\varepsilon}\right)\#\tilde{\mu}_0^n$, where $\tilde{\mu}_0^n$ is an i.i.d. sample from $\mu_0$ which is independent from $\mu_0^n$.*

*Let*

$$r_{n,d} = \begin{cases} n^{-\frac{1}{4}} & \text{if } d = 1,2,3 \\ n^{-\frac{1}{4}}\sqrt{\log(n)} & \text{if } d = 4 \\ n^{-\frac{1}{d}} & \text{if } d \geq 5 \end{cases}$$

*Then*

$$\mathbb{E}W_2(\nu,\nu^n) \lesssim n^{-\frac{(\bar{\alpha}+1)}{4(d+\bar{\alpha}+1)}}\sqrt{\log n} + r_{n,d},$$

*where $\bar{\alpha} = \min(\alpha, 3)$ and where the implicit constant may depend on $\mu_0, ..., \mu_m$ but not $n$.*

**Algorithm 1** Estimate $\lambda$ in the LBCM

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

**for** $i = 1, ..., m$ **do**

Solve for $g_\varepsilon^i$ as the optimal $g$ in

$$\max_{f,g} \frac{1}{n}\sum_{j=1}^n f(X_j) + \frac{1}{n}\sum_{k=1}^n g(Y_k^i) - \frac{\varepsilon}{n^2}\sum_{j,k}^n \exp_\varepsilon\left(f(X_j) + g(Y_k^i) - \frac{1}{2}\left\|X_j - Y_k^i\right\|_2^2\right),$$

and similarly for $g_\varepsilon^\eta$.

Define $T_{\mu_0}^{\mu_i,n,\varepsilon}$ through (18) with $g_\varepsilon = g_\varepsilon^i$ and $\{Y_1^i, ..., Y_n^i\}$ and similarly for $T_{\mu_0}^{\eta,n,\varepsilon}$.

**end for**

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

$$\hat{A}_{ij}^L = \frac{1}{n}\sum_{k=n+1}^{2n} \langle T_{\mu_0}^{\mu_i,n,\varepsilon}(X_k) - T_{\mu_0}^{\eta,n,\varepsilon}(X_k), T_{\mu_0}^{\mu_j,n,\varepsilon}(X_k) - T_{\mu_0}^{\eta,n,\varepsilon}(X_k)\rangle.$$

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

**Theorem 3.** *Suppose that Assumptions 1, 2, and 3 are satisfied for the pairs $(\mu_0,\mu_1), (\mu_0,\mu_2)$, and $(\mu_0,\eta)$. Let $X_1, ..., X_{2n} \sim \mu_0, Y_1^1, ..., Y_n^1 \sim \mu_1, Y_1^2, ..., Y_n^2 \sim \mu_2$, and $Z_1, ..., Z_n \sim \eta$ be i.i.d. samples from the respective measures. Let $T_{\mu_0}^{\mu_1,n,\varepsilon}, T_{\mu_0}^{\mu_2,n,\varepsilon}$, and $T_{\mu_0}^{\eta,n,\varepsilon}$ be the plug-in estimate entropy-regularized maps from $\mu_0^n$ to $\mu_1^n, \mu_2^n$, and $\eta^n$ respectively, all with $\varepsilon \asymp n^{-1/(d+\bar{\alpha}+1)}$ and computed using $X_1, ..., X_n$. Then we have*

$$\mathbb{E}\left[\left|\int \langle T_{\mu_0}^{\mu_1} - T_{\mu_0}^{\eta}, T_{\mu_0}^{\mu_2} - T_{\mu_0}^{\eta}\rangle d\mu_0 - \frac{1}{n}\sum_{i=n+1}^{2n} \langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_i) - T_{\mu_0}^{\eta,n,\varepsilon}(X_i), T_{\mu_0}^{\mu_2,n,\varepsilon}(X_i) - T_{\mu_0}^{\eta,n,\varepsilon}(X_i)\rangle\right|\right] \lesssim \frac{1}{\sqrt{n}} + n^{-\frac{\bar{\alpha}+1}{4(d+\bar{\alpha}+1)}}\sqrt{\log n},$$

*where $\bar{\alpha} = \min(\alpha, 3)$. The implicit constant may depend on $\mu_0, \mu_1, \mu_2, \eta$ but not $n$.*

**Corollary 1.** *Let $\hat{\lambda}$ be the random estimate obtained from Algorithm 1. Suppose that $A^L$ has an eigenvalue of 0 with multiplicity 1 and that $\lambda_* \in \Delta^m$ realizes $\lambda_*^T A^L \lambda_* = 0$. 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 implicit constant may depend on $\mu_0, ..., \mu_m, \eta$ but not $n$.*

## 5 REPRESENTATION VIA LBCM

One of the most natural questions for using the $W_2$BCM and the LBCM is to describe their "representational capacity." This is intended to quantify in an appropriate sense how closely the classes of measures in $W_2\text{BCM}(\{\mu_i\}_{i=1}^m)$ and $\text{LBCM}(\mu_0;\{\mu_i\}_{i=1}^m)$ can represent either an arbitrary probability measure or probability measures from a predetermined class. For the following discussion we will slightly generalize the $W_2$BCM and LBCM to accommodate infinite families of reference measures. For a family of measures indexed by a set $I$ we define

$$\text{LBCM}(\mu_0;\{\mu_i\}_{i\in I}) \triangleq \{\mu_\lambda \mid \lambda \in \mathcal{P}(I)\} \text{ where } \mu_\lambda = \arg\min_{\nu\in\mathcal{P}(\mathbb{R}^d)} \int \left\|T_{\mu_0}^\nu - T_{\mu_0}^{\mu_i}\right\|_{L^2(\mu_0)}^2 d\lambda(i),$$

which extends the case of finitely many measures. In an analogous way we extend the $W_2$BCM by using

$$\mu_\lambda = \arg\min_{\nu\in\mathcal{P}(\mathbb{R}^d)} \int W_2^2(\nu,\mu_i) d\lambda(i).$$

Selecting a base measure along with a set of references such that the associated LBCM or $W_2$BCM well-approximates a given set of measures is a non-trivial

problem. Indeed some common families which are have exceptional representational capacity for linear mixtures have extremely limited capacity when used in either the LBCM or $W_2$BCM. In particular, we note the following Proposition.

**Proposition 4.** *Suppose that $\mu_{(x,\Sigma)} = \mathcal{N}(x;\Sigma)$ is a Gaussian with mean $x$ and covariance matrix $\Sigma$ and consider the index set $I = \mathbb{R}^d \times \mathbb{S}^d_{++}$. It holds that*

$$W_2\text{BCM}(\{\mu_i\}_{i\in I}) \subset \{\mathcal{N}(x;\Sigma) \mid x \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++}\},$$
$$\text{LBCM}(\mu_0;\{\mu_i\}_{i\in I}) \subset \{\mathcal{N}(x;\Sigma) \mid x \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++}\}.$$

*In contrast, the set*

$$\left\{\int \mu_{(x,\Sigma)} d\lambda(x,\Sigma) \mid \lambda \in \mathcal{P}(\mathbb{R}^d \times \mathbb{S}^d_{++})\right\}$$

*is dense (in the sense of weak convergence) in $\mathcal{P}(\mathbb{R}^d)$.*

**Remark 3.** Proposition 4 holds not just for Gaussian reference measures, but for any scale-translation family, including Dirac masses (Bonneel et al., 2015).

Proposition 4 tells us that a simple family which is dense in $\mathcal{P}(\mathbb{R}^d)$ when combined in the linear mixture sense is far from being dense using either the $W_2$BCM or LBCM. This leads to the question of whether or not a family of measures $\mu_0, \{\mu_i\}_{i\in I}$ can be found where the opposite result holds true, that is a family of measures where the $W_2$BCM or LBCM is dense but the collection of linear mixtures is not. The result in the next section exhibits such a case in one dimension.

### 5.1 Density of the LBCM in $\mathcal{P}([a,b])$

For convenience we restrict ourselves to the interval $[0,1]$ but the construction can be extended to all of $\mathbb{R}$.

**Theorem 4.** *Let $U([0,1])$ denote the uniform measure over the interval $[0,1]$. The sets*

$$W_2\text{BCM}(\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\})$$
$$\text{LBCM}(U([0,1]);\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\})$$

*are both dense in $\mathcal{P}([0,1])$ with respect to weak convergence. The set*

$$\text{conv}(\{U([0,1])\} \cup \{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\})$$

*is not dense in $\mathcal{P}([0,1])$.*

The strategy of the proof is to start by representing any measure $\mu \in \mathcal{P}([0,1])$ by its transport map $T$ from $U([0,1])$. One then constructs a map, which is almost everywhere equal to $T$, as a convex combination of the maps $T_a$, the optimal transport map from $U([0,1])$ to $a\delta_0 + (1-a)\delta_1$. The existence of such a convex combination immediately establishes that $\mu \in \text{LBCM}(U([0,1]);\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\})$.

From a convex analysis perspective, the proof is done by identifying the *extreme* optimal transport maps in the set of transport maps from $U([0,1])$ to measures supported on $[0,1]$ as well as identify the measures that these extreme maps are associated with. In the next section we investigate if analogous results hold in higher dimensions.

### 5.2 Capacity in Higher Dimensions

In this section we will require the following definition.

**Definition 5.** Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a convex function. The sub-differential $\partial\phi$ is a set-valued function

$$\partial\phi(x) \triangleq \left\{y \in \mathbb{R}^d \mid \phi(z) \geq \phi(x) + \langle y, z-x\rangle \ \forall z \in \mathbb{R}^d\right\}.$$

It is well-known by Rockafellar's theorem (Rockafellar, 1997) that convex functions are almost everywhere differentiable. The sub-differential is useful for handling those points where a convex function fails to be differentiable.

One particular point of interest in the proof of Theorem 4 is that the reference measure are themselves supported on the extreme points of the interval $[0,1]$ (namely $\{0,1\}$). A natural attempt to generalize the above example to $\mathbb{R}^d$, $d > 1$, is to replace $[0,1]$ with a convex polytope $C \subset \mathbb{R}^d$, that is a compact convex set that can be realized as $C = \text{conv}(\{v_i\}_{i=1}^\ell)$ where $\ell < \infty$ [1], replace $U([0,1])$ with $U(C)$, replace $\mathcal{P}([0,1])$ with $\mathcal{P}(C)$ and replace $\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\}$ with $\mathcal{P}(\{v_i\}_{i=1}^\ell)$, the set of probability measures on the extreme points.

The following result shows that maps with image contained in $\{v_i\}_{i=1}^\ell$ are extreme, extending the approach used in Theorem 4 in one dimension.

**Proposition 5.** *Define the sets*

$$\mathcal{T}(C) \triangleq \{T : C \to C \mid \exists \varphi \text{ cvx s.t. } T(x) \in \partial\varphi(x) \ \forall x \in C\},$$
$$\mathcal{V}(C) \triangleq \left\{T \in \mathcal{T}(C) \mid T(x) \in \{v_i\}_{i=1}^\ell \text{ for a.e. } x \in C\right\}.$$

*The set $\mathcal{V}(C)$ consists of the optimal transport maps from $U(C)$ to $\mathcal{P}(\{v_i\}_{i=1}^\ell)$ and additionally $\mathcal{V}(C)$ are extreme points in $\mathcal{T}(C)$ up to almost everywhere equality in the sense that if $T \in \mathcal{V}(C)$ can be expressed as*

$$T = \int_{\mathcal{T}(C)} T_i d\lambda(T_i)$$

*for some $\lambda \in \mathcal{P}(\mathcal{T}(C))$ then $\mathbb{P}(T = T_i \text{ a.e.}) = 1$ where $T_i \sim \lambda$.*

The choice of $\mathcal{T}$ is to refer to the transport maps while $\mathcal{V}$ is to refer to the vertices of the polytope $C$ which define

[1] We may and do assume without loss of generality that every vertex $v_i$ of $C$ is extreme.

the set $\mathcal{V}(C)$. Importantly, note that in Proposition 5 it states that $\mathcal{V}(C)$ consists of points which are extreme in $\mathcal{T}(C)$, but it does *not* claim that $\mathcal{V}(C)$ contains all of the extreme points. However this is the case in one dimension when $C = [0, 1]$ and the set $\mathcal{V}([0, 1])$ is not only a subset of the extreme points of $\mathcal{T}([0, 1])$ but in fact consists of the *entire* set of extreme points.

A natural question in higher dimensions is if the set of measures on the extreme points is sufficient to generate a dense LBCM. Of course the measures on the extreme points $\mathcal{P}(\{v_i\}_{i=1}^{\ell})$ can be associated the set of $\mathcal{V}(C)$ through the relation $\mu[\{v_i\}] = U(C)[T^{-1}(v_i)]$ for every $T \in \mathcal{V}(C)$, or vice versa by observing that for every $\mu \in \mathcal{P}(\{v_i\}_{i=1}^{\ell})$ the map $T^{\mu}_{U(C)} \in \mathcal{V}(C)$. This leads to the following two equivalent questions

**Question 1.** Is the set $\text{LBCM}(U[C]; \mathcal{P}(\{v_i\}_{i=1}^{\ell}))$ dense in $\mathcal{P}(C)$?

**Question 2.** Is it the case that $\text{conv}(\mathcal{V}(C))$ is dense in $\mathcal{T}(C)$?

The following result provides a negative answer to Question 2, and thus a negative answer to Question 1.

**Theorem 5.** *The answer to Question 2 is false, that is* $\text{conv}(\mathcal{V}(C))$ *is not dense in* $\mathcal{T}(C)$*. As a consequence the answer to Question 1 is also false.*

Theorem 5 invites the following open question.

**Question 3.** What is the smallest set $\{\mu_i\}_{i\in I}$ such that $\text{LBCM}(\text{U}(C); \{\mu_i\}_{i\in I})$ is dense in $\mathcal{P}(C)$? Equivalently, what are the extreme elements of the convex set $\mathcal{T}(C)$?

## 6 APPLICATIONS

We consider applications of the LBCM to two problems: estimation of covariance matrices and recovery of corrupted MNIST digits[2]. We compare against standard baselines as well as the $W_2$BCM approach of (Werenski et al., 2022). Details of our approach are in Appendices E-I.

### 6.1 Covariance Estimation

In covariance estimation one is given a finite set of samples $X_1, ..., X_n \sim \mathcal{N}(0, \Sigma)$ and is tasked with constructing an estimate $\hat{\Sigma} \approx \Sigma$ (Pourahmadi, 2013; Srivastava and Vershynin, 2013; Friedman et al., 2008). We consider the experimental set up in Algorithm 2.

Algorithm 2 requires two procedures: `RandomCovariance`, which returns $m$ positive definite matrices representing the covariances of the reference measures, and `RandomCoordinate` which returns a random vector in $\Delta^m$. Throughout we only consider sampling from $\Delta^m$ uniformly. The estimates of $\lambda$ and $\Sigma$ are constructed using a covariance estimation procedure which is specified in Algorithm 3.

[2]Code to recreate the experiments is publicly available at `https://github.com/MattWerenski/LBCM`

**Algorithm 2** Covariance Experimental Set Up

1: **Input:** Number of references ($m$), random covariance procedure (`RandomCovariances`), the number of samples ($n$), random coordinate procedure (`RandomCoordinate`).
2: Set $(\Sigma_i, ..., \Sigma_m) =$ `RandomCovariances`$(m)$.
3: Set $\lambda =$ `RandomCoordinate`$(m)$.
4: Compute $\Sigma$ by using Algorithm 1 in (Chewi et al., 2020) with parameters $\lambda, \Sigma_1, ..., \Sigma_m$.
5: Draw $n$ samples $X_1, ..., X_n \sim \mathcal{N}(0, \Sigma)$.
6: Construct estimates $\hat{\lambda}, \hat{\Sigma}$ using Algorithm 3 with the samples $X_1, ..., X_n$ and references $\Sigma_1, ..., \Sigma_m$.
7: **Return** $\hat{\lambda}, \hat{\Sigma}$.

**Algorithm 3** Covariance Estimation with References

1: **Input:** i.i.d. samples $(X_1, ..., X_n \sim \mathcal{N}(0, \Sigma))$, reference measures $(\mathcal{N}(0, \Sigma_1), ..., \mathcal{N}(0, \Sigma_m))$, coordinate estimation procedure (`EstimateCoordinate`), covariance estimation procedure (`EstimateCovariance`).
2: Set $\hat{\Sigma}_{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$.
3: Set $\hat{\lambda} =$ `EstimateCoordinate`$(\hat{\Sigma}_{\text{emp}}; \Sigma_1, ..., \Sigma_m)$.
4: Set $\hat{\Sigma} =$ `EstimateCovariance`$(\hat{\lambda}; \Sigma_1, ..., \Sigma_m)$.
5: **Return** $\hat{\lambda}, \hat{\Sigma}$.

Algorithm 3 requires two further procedures to be provided. The first is `EstimateCoordinate` which takes as arguments the empirical covariance matrix $\hat{\Sigma}_{\text{emp}}$ and the collection of reference covariances $\Sigma_1, ..., \Sigma_m$ and returns an estimate of the coordinate $\hat{\lambda}$. The second procedure is `EstimateCovariance` which takes as arguments the estimated coordinate and the set of reference covariances and returns the estimate of the covariance.

We consider three versions of Algorithm 3: LBCM, BCM, and maximum likelihood estimation (MLE); details are in Appendix E.

**Remark 4.** In order to effectively compare the three methods above we must consider a setting where the $W_2$BCM and LBCM coincide. This is the case when the covariance matrices $\Sigma_1, ..., \Sigma_m$ are *simultaneously diagonalizable* which is equivalent to $\Sigma_1, ..., \Sigma_m$ having the same eigenvectors. The `RandomCovariances` achieves this using a two step process. First, sample an orthogonal matrix $O \sim \text{Ortho}(d)$ and then for each $i = 1, ..., m$ set $\Sigma_i = O^T D_i O$ where $D_i$ is a random diagonal matrix with strictly positive entries on its

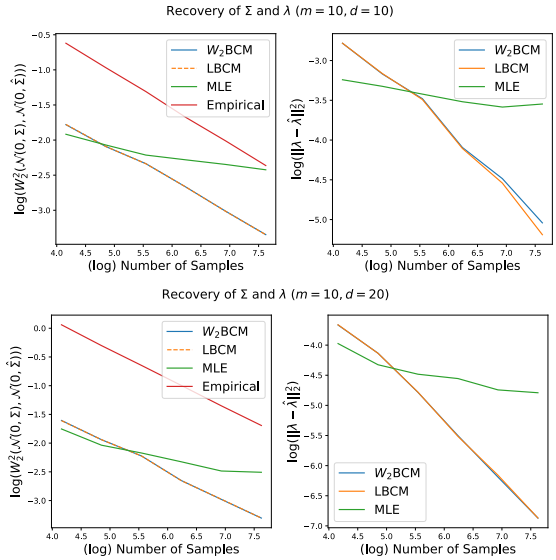


Figure 1: Recovery of the covariance matrix and coordinate in logarithmic scale.

diagonal. Throughout we let $D_i$ have independent diagonal entries distributed according to the half-normal distribution.[3]

Results are shown in Figure 1. We plot both the discrepancy between the recovered covariance matrix and the discrepancy in the estimate of the coordinate $\lambda$. In addition we plot the distance from $\mathcal{N}(0, \hat{\Sigma}_{\text{emp}})$ to $\mathcal{N}(0, \Sigma)$. The LBCM uses $\mu_0 = \mathcal{N}(0, I)$. The first setting we consider is when $m = 10$ and $d = 10$. In this case, incorporating the $W_2$BCM or LBCM leads to approximately one quarter the error obtained by the empirical covariance. The asymptotic behavior for the $W_2$BCM, LBCM, and empirical covariance are all essentially of order $n^{-1/2}$. The quality of the estimate of the covariance is extremely similar for both the $W_2$BCM and LBCM in this setting. The behavior of the MLE is interesting but we do not have a precise explanation of its behavior which may be either theoretical or connected with the numerical methods involved in obtaining it. Next we follow the same procedure except with $d = 20$ and observe similar results.

### 6.2 MNIST Reconstruction

Another application we consider is in digit reconstruction. This is done using the MNIST image set which consists of 28×28 grayscale images of handwritten digits 0 through 9. Sample digits are given as the first column of Figures 2 and 6. The problem we consider is, given an occluded digit which is missing the central block of pixels, to recover the digit (see Figure 2). We do so using the $W_2$BCM, LBCM, and a simple linear method. Each of these methods falls into the framework of Algorithm 4 by specifying `EstimateCoordinate` and `SynthesizeImage`.

**Algorithm 4** Digit Reconstruction Procedure

1: **Input:** Corrupted digit to reconstruct ($\mathtt{C}_\eta$), Reference digits ($\mathtt{D}_1, ..., \mathtt{D}_m$), Coordinate estimation procedure (`EstimateCoordinate`), Image synthesis procedure (`SynthesizeImage`).
2: Set $\mathtt{C}_i = \mathtt{Occlude}(\mathtt{D}_i)$ for $i = 1, ..., m$ # Removes the central block of pixels
3: Set $\lambda = \mathtt{EstimateCoordinate}(\mathtt{C}_\eta, \mathtt{C}_1, ..., \mathtt{C}_m)$
4: **Return** $\mathtt{SynthesizeImage}(\lambda, [\mathtt{D}_1, ..., \mathtt{D}_m])$

For the LBCM we specify a base digit $\mathtt{D}_0$ and an occluded version $\mathtt{C}_0$. `EstimateCoordinate` is then performed by converting $\mathtt{C}_\eta, \mathtt{C}_0, ..., \mathtt{C}_m$ to measures using Algorithm 8 and then obtain $\lambda$ by solving the minimization in (7). For `SynthesizeImage` we convert $\mathtt{D}_0, ..., \mathtt{D}_m$ to measures, again using Algorithm 8, and construct a measure $\rho_\lambda$ by pushing $\mathtt{D}_0$ through the weighted combination of the transport maps. We then convert the resulting measure back to an image using Algorithm 9 in the Appendix which is akin to kernel density estimation.

The linear method we consider performs `EstimateCoordinate` by normalizing the occluded digits $\mathtt{C}_\eta, \mathtt{C}_1, ..., \mathtt{C}_m$ to each have total mass one and then treats them as vectors in $\mathbb{R}^{784}$ with each entry corresponding to the color at a pixel position and projects $\mathtt{C}_\eta$ onto the convex hull $\text{conv}(\mathtt{C}_1, ..., \mathtt{C}_m)$ and returns $\lambda$ corresponding to the closest point. `SynthesizeImage` returns the weighted sum of the digits $\sum_{i=1}^m \lambda_i \mathtt{D}_i$.

Finally, for the $W_2$BCM, `EstimateCoordinate` converts $\mathtt{C}_\eta, \mathtt{C}_1, ..., \mathtt{C}_m$ into measures using Algorithm 8 in the Appendix and then obtains $\lambda$ by solving the minimization in (8). To recover an image, we first synthesize a (approximate) 2-Wasserstein barycenter using Algorithm 7 (see Subsection G for details), with $\alpha = 0.05$, $k = 200$ and initialized at $\mu_0 \triangleq \sum_{i=1}^m \lambda_i \mathtt{D}_i$, the weighted sum of digits computed via the linear method. We recover an image by applying `SynthesizeImage` to the output.

Examples of running this procedure with 10 given reference 4's are shown in Figure 2. The first two columns correspond to the original digit before and after being occluded. The second and third columns are the reconstruction using the LBCM and two different base digits (see Appendix I). The fifth column uses the $W_2$BCM and the sixth column is the linear reconstruction. Above each column, we report the (average) run time to produce each image, as well as the $W_2^2$ cost be-

[3] If $X \sim \mathcal{N}(0, 1)$ then $|X|$ has a half-normal distribution.

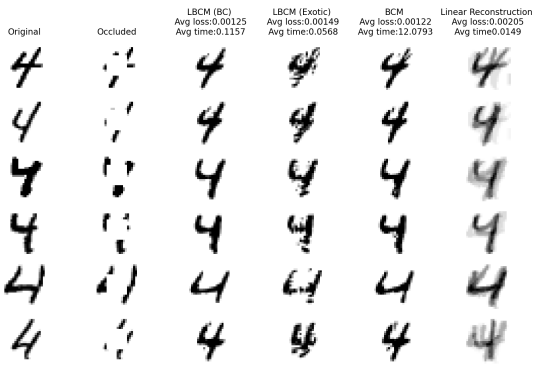


Figure 2: Reconstruction of occluded digits using the $W_2$BCM, LBCM, and a linear method. The base measure in Column 4 is the Double Checker in Figure 5.

tween the reconstructed image and the original image as probability measures. We note that both the $W_2$BCM and LBCM tend to do a reasonable job recovering the digit while the linear method has significant difficulties. This indicates that the non-linear methods employed in the $W_2$BCM and LBCM may be better suited to this task than the naive linear method. We also note that the choice of initialization for the LBCM has an impact on the quality of the reconstruction, which we explore further in Appendix I. Finally, we remark that while the $W_2$BCM reconstruction has slightly lower average loss than the LBCM, the average run time is significantly greater than for the LBCM method.

## 7 CONCLUSION

While Theorem 5 establishes that it is not possible to represent an arbitrary optimal transport map as a convex combination of transport maps onto the extreme points, is is interesting to consider whether maps that push onto richer sets (e.g., the boundary) are sufficient.

In (Jiang et al., 2023), quantitative rates of approximation are given for representing 1-dimensional optimal transport maps as linear combinations of polynomials. Developing similar results in our setting and in higher dimensions is of interest. This is related to the problem of developing function classes well-suited to approximation of transport maps.

**Acknowledgments:** Matthew Werenski was supported by NSF CCF 1553075, NSF DMS 2309519, and the Camille & Henry Dreyfus Foundation. Brendan Mallery was supported by NSF CCF 1553075, NSF DMS 2309519, and NSF DMS 2318894. Shuchin Aeron would like to acknowledge partial support by NSF CCF 1553075, NSF DMS 2309519, and from NSF PHY 2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). James M. Murphy would like to acknowledge partial support from NSF DMS 2309519, NSF DMS 2318894, and support from the Camille & Henry Dreyfus Foundation.

### References

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

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., and Matrán, C. (2018). Wide consensus aggregation in the wasserstein space. application to location-scatter families. *Bernoulli*, 24(4A):3147–3179.

Bigot, J., Cazelles, E., and Papadakis, N. (2019). Penalization of barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 51(3):2261–2285.

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

Blickhan, T. (2024). A registration method for reduced basis problems using linear optimal transport. *SIAM Journal on Scientific Computing*, 46(5):A3177–A3204.

Bonneel, N. and Digne, J. (2023). A survey of optimal transport for computer graphics and computer vision. *Computer Graphics Forum*, 42(2):439–460.

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

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

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

Caffarelli, L. A. (1992). The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104.

Cai, T., Cheng, J., Craig, K., and Craig, N. (2022). Which metric on the space of collider events? *Physical Review D*, 105(7):076003.

Carlier, G., Eichinger, K., and Kroshnin, A. (2021). Entropic-Wasserstein barycenters: PDE characteri-

zation, regularity, and CLT. *SIAM Journal on Mathematical Analysis*, 53(5):5880–5914.

Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., and Papadakis, N. (2018). Geodesic PCA versus log-PCA of histograms in the wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456.

Cheng, K., Aeron, S., Hughes, M. C., and Miller, E. L. (2021). Dynamical Wasserstein barycenters for time-series modeling. *Advances in Neural Information Processing Systems*, 34:27991–28003.

Cheng, K. C., Miller, E. L., Hughes, M. C., and Aeron, S. (2023). Non-parametric and regularized dynamical Wasserstein barycenters for sequential observations. *IEEE Transactions on Signal Processing*, 71:3164–3178.

Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. J. (2020). Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR.

Chizat, L. (2023). Doubly regularized entropic Wasserstein barycenters. *arXiv preprint arXiv:2303.11844*.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158.

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

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

Deb, N., Ghosal, P., and Sen, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753.

Delalande, A. and Merigot, Q. (2023). Quantitative stability of optimal transport maps under variations of the target measure. *Duke Mathematical Journal*, 172(17):3321–3357.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gangbo, W. and McCann, R. J. (1996). The geometry of optimal transportation. *Acta Mathematica*, 177:113–161.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gunsilius, F., Hsieh, M. H., and Lee, M. J. (2024). Tangential Wasserstein projections. *Journal of Machine Learning Research*, 25(69):1–41.

Janati, H., Cuturi, M., and Gramfort, A. (2020). Debiased sinkhorn barycenters. In *International Conference on Machine Learning*, pages 4692–4701. PMLR.

Jiang, Y., Chewi, S., and Pooladian, A.-A. (2023). Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *arXiv preprint arXiv:2312.02849*.

Komiske, P. T., Metodiev, E. M., and Thaler, J. (2019). Metric space of collider events. *Physical review letters*, 123(4):041801.

Lee, D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13.

Mallery, B., Murphy, J. M., and Aeron, S. (2025). Synthesis and analysis of data as probability measures with entropy-regularized optimal transport. *arXiv preprint arXiv:2501.07446*.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2024). Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998.

Masud, S. B., Werenski, M., Murphy, J. M., and Aeron, S. (2023). Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling. *Journal of Machine Learning Research*, 24(160):1–65.

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

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32.

Mérigot, Q., Delalande, A., and Chazal, F. (2020). Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pages 3186–3196. PMLR.

Métivier, L., Brossier, R., Merigot, Q., Oudet, E., and Virieux, J. (2016). An optimal transport approach for seismic tomography: Application to 3D full waveform inversion. *Inverse Problems*, 32(11):115008.

Moosmüller, C. and Cloninger, A. (2023). Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389.

Mueller, M., Aeron, S., Murphy, J. M., and Tasissa, A. (2023). Geometrically regularized wasserstein dictionary learning. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 384–403. PMLR.

Nutz, M. (2021). Introduction to entropic optimal transport. *Lecture notes, Columbia University*.

Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. Springer Nature.

Peyré, G. and Cuturi, M. (2020). Computational optimal transport.

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

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*, volume 882. John Wiley & Sons.

Rigollet, P. and Stromme, A. J. (2022). On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*.

Rockafellar, R. T. (1997). *Convex Analysis*, volume 18. Princeton University Press.

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

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

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

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

Sommer, S., Lauze, F., Hauberg, S., and Nielsen, M. (2010). Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In *European Conference on Computer Vision*, pages 43–56. Springer.

Srivastava, N. and Vershynin, R. (2013). Covariance estimation for distributions with $2+\varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111.

Stromme, A. (2023). Sampling from a Schrödinger bridge. In *International Conference on Artificial Intelligence and Statistics*, pages 4058–4067. PMLR.

Vaskevicius, T. and Chizat, L. (2024). Computational guarantees for doubly entropic wasserstein barycenters. *Advances in Neural Information Processing Systems*, 36.

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

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

Werenski, M., Masud, S. B., Murphy, J. M., and Aeron, S. (2024). On rank energy statistics via optimal transport: Continuity, convergence, and change point detection. *IEEE Transactions on Information Theory*, 70(7):5168–5190.

Werenski, M., Murphy, J. M., and Aeron, S. (2023). Estimation of entropy-regularized optimal transport maps between non-compactly supported measures. *arXiv preprint arXiv:2311.11934*.

Xu, H., Wang, W., Liu, W., and Carin, L. (2018). Distilled wasserstein learning for word embedding and topic modeling. *Advances in Neural Information Processing Systems*, 31.

Zemel, Y. and Panaretos, V. M. (2019). Fréchet means and procrustes analysis in wasserstein space. *Bernoulli*, 25(2):932–976.

## CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/**No**/Not Applicable], **All code should be able to run on a modest personal computer.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

# A PROOFS FROM SECTION 3

## A.1 Proof of Proposition 1

For any $x$ such that $T_{\mu_0}^{\mu_i}(x)$ exists for each $i = 1, \dots, m$, the pointwise variational problem

$$\arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^m \lambda_i \|y - T_{\mu_0}^{\mu_i}(x)\|^2$$

has unique solution $\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}(x)$. In particular, since the transport maps exist $\mu_0-$almost everywhere and are necessarily in $L^2(\mu_0)$,

$$\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i} = \arg\min_{f \in L^2(\mu_0)} \sum_{i=1}^m \lambda_i \|f - T_{\mu_0}^{\mu_i}\|_{L^2(\mu_0)}^2.$$

We now argue that in fact $\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}$ is equal to $T_{\mu_0}^{\left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right)\#\mu_0}$. Notice that $\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}$ is the convex combination of 2-Wasserstein optimal transport maps, each of which is by Brenier's theorem the gradient of a convex function. It follows that $\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}$ is itself the gradient of a convex function. Moreover, $\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}$ pushes $\mu_0$ onto $\left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right)\#\mu_0$. Again by Brenier's theorem, it follows that

$$\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i} = T_{\mu_0}^{\left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right)\#\mu_0}.$$

It follows that $\left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right)\#\mu_0$ solves (6).

## A.2 Proof of Proposition 2

This result relies on the equivalence of the 2-Wasserstein barycenter and the solution to the *multimarginal optimal transport (MMOT)* (Agueh and Carlier, 2011). Let $\{\mu_i\}_{i=1}^m$ be reference measures and $\lambda \in \Delta^m$. The MMOT problem solves

$$\arg\min_{\pi \in \Pi(\mu_1, \dots, \mu_m)} \int_{\mathbb{R}^d \times \cdots \times \mathbb{R}^d} \sum_{i<j} \lambda_i \lambda_j \|x_i - x_j\|^2 d\pi(x_1, \dots, x_m) \tag{10}$$

where $\Pi(\mu_1, \dots, \mu_m)$ is the space of measures in $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$ with $i^{th}$ $d$-dimensional marginal equal to $\mu_i$. The MMOT problem can be related to 2-Wasserstein barycenters via the averaging operator $M_\lambda : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}^d$ given by $M_\lambda(x_1, \dots, x_m) \triangleq \sum_{i=1}^m \lambda_i x_i$. Indeed,

$$\begin{aligned} &\int_{\mathbb{R}^d \times \cdots \times \mathbb{R}^d} \sum_{i<j} \lambda_i \lambda_j \|x_i - x_j\|^2 d\pi(x_1, \dots, x_m) \\ &= \int_{\mathbb{R}^d \times \cdots \times \mathbb{R}^d} \sum_{i=1}^m \lambda_i \|x_i - M_\lambda(x_1, \dots, x_m)\|^2 d\pi(x_1, \dots, x_m). \end{aligned}$$

**Proposition 6.** *(Proposition 4.2 in (Agueh and Carlier, 2011) and Proposition 3.1.2 in (Panaretos and Zemel, 2020)) Let $\{\mu_i\}_{i=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ and $\lambda \in \Delta^m$. Then $\mu_*$ is a 2-Wasserstein barycenter for $\{\mu_i\}_{i=1}^m$ with coefficieints $\lambda$ if and only if there exists a solution $\pi_*$ to (10) such that $\mu_* = M_\lambda \# \pi_*$.*

*Proof.* Let $\pi \in \Pi(\mu_1, \dots, \mu_m)$ be arbitrary and let $\mu = M_\lambda \# \pi$. Then the map $S_i(x_1, \dots, x_m) = (x_i, M(x_1, \dots, x_m))$ generates a coupling between $\mu_i$ and $\mu$ via $S_i \# \pi$. Thus

$$\int_{(\mathbb{R}^d)^m} \|x_i - M_\lambda(x_1, \dots, x_m)\|^2 d\pi(x_1, \dots, x_m) \geq W_2^2(\mu, \mu_i).$$

Summing over $\lambda$, we see

$$\min_{\pi\in\Pi(\mu_1,\dots,\mu_m)}\int_{(\mathbb{R}^d)^m}\sum_{i=1}^m\lambda_i\|x_i-M_\lambda(x_1,\dots,x_m)\|^2d\pi(x_1,\dots,x_m)\geq\min_{\mu\in\mathcal{P}(\mathbb{R}^d)}\sum_{i=1}^m\lambda_iW_2^2(\mu,\mu_i). \tag{11}$$

On the other hand, let $\mu\in\mathcal{P}(\mathbb{R}^d)$ be arbitrary and for each $i=1,\dots,m$ let $\pi^i$ be the optimal 2-Wasserstein coupling between $\mu$ and $\mu_i$. We may glue the $\pi^i$ via their common marginal to get a measure $\eta$ on $(\mathbb{R}^d)^{m+1}$ with marginals $\mu_1,\dots,\mu_m,\mu$ and its projection $\pi$ onto the first $m$ $d$-dimensional marginals is an element of $\Pi(\mu_1,\dots,\mu_m)$. We now note

$$\begin{aligned}
&\sum_{i=1}^m\lambda_iW_2^2(\mu,\mu_i)\\
=&\sum_{i=1}^m\lambda_i\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x_i-y\|^2d\pi^i(y,x_i)\\
=&\int_{(\mathbb{R}^d)^{m+1}}\sum_{i=1}^m\lambda_i\|x_i-y\|^2d\eta(x_1,\dots,x_m,y)\\
\geq&\int_{(\mathbb{R}^d)^{m+1}}\sum_{i=1}^m\lambda_i\|x_i-M_\lambda(x_1,\dots,x_m)\|^2d\eta(x_1,\dots,x_m,y)\\
=&\int_{(\mathbb{R}^d)^m}\sum_{i=1}^m\lambda_i\|x_i-M_\lambda(x_1,\dots,x_m)\|^2d\pi(x_1,\dots,x_m),
\end{aligned}$$

with equality if and only if $y=M_\lambda(x)$ $\eta$-a.e., so that $\mu=M_\lambda\#\pi$. This establishes

$$\min_{\pi\in\Pi(\mu_1,\dots,\mu_m)}\int_{(\mathbb{R}^d)^m}\sum_{i=1}^m\lambda_i\|x_i-M_\lambda(x_1,\dots,x_m)\|^2d\pi(x_1,\dots,x_m)=\min_{\mu\in\mathcal{P}(\mathbb{R}^d)}\sum_{i=1}^m\lambda_iW_2^2(\mu,\mu_i). \tag{12}$$

Now, if a barycenter $\mu_*$ does not equal $M_\lambda\#\pi$ with $\pi$ as above, then

$$\begin{aligned}
&\sum_{i=1}^m\lambda_iW_2^2(\mu_*,\mu_i)\\
>&\int_{(\mathbb{R}^d)^m}\sum_{i=1}^m\lambda_i\|x_i-M_\lambda(x_1,\dots,x_m)\|^2d\pi(x_1,\dots,x_m)\\
\geq&\sum_{i=1}^m\lambda_iW_2^2(M_\lambda\#\pi,\mu_i),
\end{aligned}$$

which contradicts optimality of $\mu_*$. Now, let $\pi_*$ denote this choice such that $M_\lambda\#\pi_*=\mu_*$. Then by (12) $\pi_*$ must be optimal for the MMOT problem:

$$\begin{aligned}
&\int_{(\mathbb{R}^d)^m}\sum_{i=1}^m\lambda_i\|x_i-M_\lambda(x_1,\dots,x_m)\|^2d\pi_*(x_1,\dots,x_m)\\
=&\sum_{i=1}^m\lambda_iW_2^2(\mu_*,\mu_i)\\
=&\min_{\mu\in\mathcal{P}(\mathbb{R}^d)}\sum_{i=1}^m\lambda_iW_2^2(\mu,\mu_i)\\
=&\min_{\pi\in\Pi(\mu_1,\dots,\mu_m)}\int_{(\mathbb{R}^d)^m}\sum_{i=1}^m\lambda_i\|x_i-M_\lambda(x_1,\dots,x_m)\|^2d\pi(x_1,\dots,x_m)
\end{aligned}$$

Conversely, if $\pi_*$ is an optimal coupling for the MMOT problem, then

$$\begin{aligned}
&\sum_{i=1}^m \lambda_i W_2^2(M_\lambda \# \pi_*, \mu_i) \\
\leq &\sum_{i=1}^m \lambda_i \int_{(\mathbb{R}^d)^m} \sum_{i=1}^m \|x_i - M_\lambda(x_1,\dots,x_m)\|^2 d\pi_*(x_1,\dots,x_m) \\
= &\min_{\pi \in \Pi(\mu_1,\dots,\pi_m)} \int_{(\mathbb{R}^d)^m} \sum_{i=1}^m \|x_i - M_\lambda(x_1,\dots,x_m)\|^2 d\pi(x_1,\dots,x_m) \\
= &\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^m \lambda_i W_2^2(\mu, \mu_i),
\end{aligned}$$

so that $M_\lambda \# \pi_*$ is necessarily a 2-Wasserstein barycenter.

□

With this result, we can now establish the equivalence of the $W_2$BCM and LBCM for compatible measures.

***Proof of Proposition 2.*** It is enough to show that if $\{\mu_i\}_{i=0}^m$ are compatible, then the 2-Wasserstein barycenter of $\{\mu_i\}_{i=1}^m$ with coefficients $\lambda \in \Delta^m$ is $\left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right) \# \mu_0$. Note that $\pi_* = (T_{\mu_0}^{\mu_1}, T_{\mu_0}^{\mu_2}, \dots, T_{\mu_0}^{\mu_m}) \# \mu_0 \in \Pi(\mu_1, \mu_2, \dots, \mu_m)$ is such that

$$\begin{aligned}
&\int_{(\mathbb{R}^d)^m} \|x_i - x_j\|^2 d\pi_*(x_1,\dots,x_m) \\
=&\int_{\mathbb{R}^d} \|T_{\mu_0}^{\mu_i}(x_0) - T_{\mu_0}^{\mu_j}(x_0)\|^2 d\mu_0(x_0) \\
=&\int_{\mathbb{R}^d} \|T_{\mu_0}^{\mu_i} \circ T_{\mu_j}^{\mu_0}(x_j) - T_{\mu_0}^{\mu_j} \circ T_{\mu_j}^{\mu_0}(x_j)\|^2 d\mu_j(x_j) \\
=&\int_{\mathbb{R}^d} \|T_{\mu_j}^{\mu_i}(x_j) - x_j\|^2 d\mu_j(x_j) \\
=&W_2^2(\mu_i, \mu_j).
\end{aligned}$$

Noting that

$$\min_{\pi \in \Pi(\mu_1,\dots,\mu_m)} \int_{(\mathbb{R}^d)^m} \sum_{i<j} \lambda_i \lambda_j \|x_i - x_j\|^2 d\pi(x_1,\dots,x_m) \geq \sum_{i<j} \lambda_i \lambda_j W_2^2(\mu_i, \mu_j),$$

we see that $\pi_*$ is optimal for the MMOT problem, and thus $M_\lambda \# \pi_*$ is the 2-Wasserstein barycenter of $\{\mu_i\}_{i=1}^m$. The result follows by noting $M_\lambda \# \pi_* = \left(\sum_{i=1}^m \lambda_i T_{\mu_0}^{\mu_i}\right) \# \mu_0$. □

## B ESTIMATION OF OT MAPS VIA ENTROPY-REGULARIZED OT

The *entropy-regularized optimal transport (EOT)* problem is:

$$S_\varepsilon(\mu,\nu) \triangleq \min_{\pi \in \Pi(\mu,\nu)} \int \frac{1}{2} \|x-y\|_2^2 \, d\pi(x,y) + \varepsilon \int \log\left(\frac{d\pi}{d\mu \otimes d\nu}\right)(x,y) d\pi(x,y) \tag{13}$$

where $\varepsilon > 0$.

**Theorem 6.** *((Csiszár, 1975; Mena and Niles-Weed, 2019)) Assume that $\mu$ and $\nu$ have finite second moments. Then the following hold.*

1. *There exists a measure $\pi_\varepsilon$ which achieves the minimum in (13).*
2. *The dual problem to (13) is given by*

$$S_\varepsilon(\mu,\nu) = \max_{f,g\in L^1(\mu)\times L^1(\nu)} \Psi_\varepsilon^{\mu,\nu}(f,g), \tag{14}$$

$$\Psi_\varepsilon^{\mu,\nu}(f,g) \triangleq \int f(x)d\mu(x) + \int g(y)d\nu(y) - \varepsilon \iint \exp_\varepsilon\left(f(x)+g(y)-\frac{1}{2}\|x-y\|_2^2\right) d\mu(x)d\nu(y) + \varepsilon. \tag{15}$$

3. *There exists an optimal pair $f_\varepsilon, g_\varepsilon$ which are unique up to a.e. equality and translation by an additive constant, that is $(f_\varepsilon + a, g_\varepsilon - a)$ is also optimal for any $a \in \mathbb{R}$.*
4. *The primal-dual optimality relation is given by the following holding for $\mu\otimes\nu$ almost every $(x,y)$:*

$$\frac{d\pi_\varepsilon}{d\mu\otimes d\nu}(x,y) = \exp_\varepsilon(f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|_2^2). \tag{16}$$

5. *The optimal pair $f_\varepsilon, g_\varepsilon$ can be taken such that*

$$f_\varepsilon(x) = -\varepsilon\log\left(\int \exp_\varepsilon(g_\varepsilon(y)-\frac{1}{2}\|x-y\|_2^2)d\nu(y)\right) \ \forall x\in\mathbb{R}^d,$$
$$g_\varepsilon(y) = -\varepsilon\log\left(\int \exp_\varepsilon(f_\varepsilon(x)-\frac{1}{2}\|x-y\|_2^2)d\mu(x)\right) \ \forall y\in\mathbb{R}^d.$$

Point 1-4 are well-known and can be found in (Csiszár, 1975). Point 5 can be found in (Mena and Niles-Weed, 2019). We will refer to $f_\varepsilon, g_\varepsilon$, the maximizers in the dual problem (14), as the optimal potentials. Importantly in point 5, the equality is for all $x$ (or $y$) instead of just almost every $x$ (or $y$). We will only consider potentials which satisfy point 5.

Unlike in the unregularized OT problem in which an optimal transport map realizing (4) exists when $\mu$ is absolutely continuous, there is no transport map on whose graph $\pi_\varepsilon$ concentrates. However, we can construct one from $f_\varepsilon, g_\varepsilon$:

**Definition 6.** The *entropy-regularized map* $T_\mu^{\nu,\varepsilon}$ for $\mu$ and $\nu$ is defined as

$$T_\mu^{\nu,\varepsilon}(x) \triangleq \int y \exp_\varepsilon(f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|_2^2)d\nu(y). \tag{17}$$

When $\mu^n$ and $\nu^n$ are empirical measures of $\mu$ and $\nu$ supported on points $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$ respectively, we get the plug-in estimate of $T_\mu^{\nu,\varepsilon}$

$$T_\mu^{\nu,n,\varepsilon}(x) \triangleq \sum_{j=1}^n \exp_\varepsilon(f_{n,\varepsilon}(x)+g_{n,\varepsilon}(Y_j)-\frac{1}{2}\|x-Y_j\|_2^2)Y_j, \tag{18}$$

where $f_{n,\varepsilon}, g_{n,\varepsilon}$ are the (extended) entropic dual potentials for the measures $\mu^n, \nu^n$.

# C PROOFS FROM SECTION 4

## C.1 Proof of Proposition 3

We will require the following bound for estimating $\mathbb{E}[W_2(\mu,\mu^n)]$ on a compact set:

**Theorem 7.** *(Theorem 1, Fournier and Guillin (2015)) Let $\mu\in\mathcal{P}(\Omega)$, where $\Omega\subset\mathbb{R}^d$ is a compact set, and let $\mu^n$ denote the empirical measure on $n$ samples from $\mu$. Then:*

$$\mathbb{E}[W_2(\mu,\mu^n)] \le C_{\Omega,d}\begin{cases} n^{-\frac{1}{4}} & \text{if } d=1,2,3 \\ n^{-\frac{1}{4}}\sqrt{\log(n)} & \text{if } d=4 \\ n^{-\frac{1}{d}} & \text{if } d\ge 5 \end{cases}$$

*where $C_{\Omega,d}$ is a constant depending on the dimension and diameter of $\Omega$.*

*Proof of Proposition 3.* Recall that for any measures $\mu_0, \mu, \nu$, $W_2(\mu,\nu) \leq \|T^{\mu}_{\mu_0} - T^{\nu}_{\mu_0}\|_{L^2(\mu_0)}$. We now calculate and apply Theorem 2:

$$
\begin{aligned}
\mathbb{E}[W_2(\nu,\nu^n)] =& \mathbb{E}\left(W_2\left(\left(\sum_{i=1}^m \lambda_i T^{\mu_i}_{\mu_0}\right)\#\mu_0, \left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\tilde{\mu}^n_0\right)\right)\\
\leq& \mathbb{E}\left(W_2\left(\left(\sum_{i=1}^m \lambda_i T^{\mu_i}_{\mu_0}\right)\#\mu_0, \left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\mu_0\right)\right)\\
&+ \mathbb{E}\left[\mathbb{E}\left(W_2\left(\left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\mu_0, \left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\tilde{\mu}^n_0\right)\middle|\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\right]\\
\leq& \mathbb{E}\left(W_2\left(\left(\sum_{i=1}^m \lambda_i T^{\mu_i}_{\mu_0}\right)\#\mu_0, \left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\mu_0\right)\right) + \mathbb{E}\left[\mathbb{E}\left(C_{\Omega,d} r_{n,d}\middle|\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\right] \qquad (19)\\
\leq& \mathbb{E}\left(\left\|\sum_{i=1}^m \lambda_i T^{\mu_i}_{\mu_0} - \sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right\|_{L^2(\mu_0)}\right) + C_{\Omega,d} r_{n,d}\\
\leq& \sum_{i=1}^m \lambda_i \mathbb{E}\|T^{\mu_i}_{\mu_0} - T^{\mu_i,n,\varepsilon}_{\mu_0}\|_{L^2(\mu_0)} + C_{\Omega,d} r_{n,d}\\
\lesssim& \max_{1\leq i\leq m} \sqrt{(1+I_0(\mu_0,\mu_i)) n^{-\frac{(\bar{\alpha}+1)}{2(d+\bar{\alpha}+1)}} \log n} + r_{n,d},
\end{aligned}
$$

where in (19), we applied Theorem 7 to $\mathbb{E}\left(W_2\left(\left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\mu_0, \left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\tilde{\mu}^n_0\right)\middle|\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)$, which applies as $\tilde{\mu}^n_0$ and $T^{\mu_i,n,\varepsilon}_{\mu_0}$ are independent, and hence $\left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\tilde{\mu}^n_0$ is supported on an i.i.d. sample from $\left(\sum_{i=1}^m \lambda_i T^{\mu_i,n,\varepsilon}_{\mu_0}\right)\#\mu_0$. □

### C.2 Proof of Theorem 3

*Proof.* Throughout, unless otherwise noted, the expectation is with respect to all of $X_1,...,X_{2n}, Y^1_1,...,Y^1_n, Y^2_1,...,Y^2_n$ and $Z_1,...,Z_n$. We also use the notations $X^n$ to denote the set of variables $(X_1,...,X_n)$, and similarly for $Y^n$ and $Z^n$.

We can assume without loss of generality that $0 \in \Omega$; see (Peyré and Cuturi, 2020) Remark 2.19 for invariance of $T^{\mu_1}_{\eta}, T^{\mu_2}_{\eta}$ under translation. For the translation invariance of $T^{\mu_1,n,\varepsilon}_{\eta}, T^{\mu_2,n,\varepsilon}_{\eta}$, observe that both obtaining $g_\varepsilon$, and evaluating $T^{\mu_1,n,\varepsilon}_{\eta}, T^{\mu_2,n,\varepsilon}_{\eta}$ at fixed points, requires only the distances $\|X_i - Y_j\|_2^2 = \|(X_i - t) - (Y_j - t)\|_2^2$. We calculate:

$$
\begin{aligned}
&\mathbb{E}\left[\left|\int\langle T_{\mu_0}^{\mu_1}-T_{\mu_0}^{\eta},T_{\mu_0}^{\mu_2}-T_{\mu_0}^{\eta}\rangle d\mu_0-\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i),T_{\mu_0}^{\mu_2,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i)\rangle\right|\right]\\
=&\mathbb{E}\Bigg[\Bigg|\int\langle T_{\mu_0}^{\mu_1}-T_{\mu_0}^{\eta},T_{\mu_0}^{\mu_2}-T_{\mu_0}^{\eta}\rangle d\mu_0-\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1}(X_i)-T_{\mu_0}^{\eta}(X_i),T_{\mu_0}^{\mu_2}(X_i)-T_{\mu_0}^{\eta}(X_i)\rangle\\
&+\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1}(X_i)-T_{\mu_0}^{\eta}(X_i),T_{\mu_0}^{\mu_2}(X_i)-T_{\mu_0}^{\eta}(X_i)\rangle\\
&-\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i),T_{\mu_0}^{\mu_2,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i)\rangle\Bigg|\Bigg]\\
\le&\mathbb{E}\left[\left|\int\langle T_{\mu_0}^{\mu_1}-T_{\mu_0}^{\eta},T_{\mu_0}^{\mu_2}-T_{\mu_0}^{\mu_2}\rangle d\mu_0-\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1}(X_i)-T_{\mu_0}^{\eta}(X_i),T_{\mu_0}^{\mu_2}(X_i)-T_{\mu_0}^{\eta}(X_i)\rangle\right|\right] \qquad (20)\\
&+\mathbb{E}\Bigg[\Bigg|\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1}(X_i)-T_{\mu_0}^{\eta}(X_i),T_{\mu_0}^{\mu_2}(X_i)-T_{\mu_0}^{\eta}(X_i)\rangle\\
&-\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i),T_{\mu_0}^{\mu_2,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i)\rangle\Bigg|\Bigg]
\end{aligned}
$$

So, it suffices to bound two terms in (20) separately. To bound the first term, let $h(x)\triangleq\langle T_{\mu_0}^{\mu_1}(x)-T_{\mu_0}^{\eta}(x),T_{\mu_0}^{\mu_2}(x)-T_{\mu_0}^{\eta}(x)\rangle$. Note that for all $x\in\Omega$, $|h(x)|\le 4|\Omega|^2$, and thus $Var(h(x))\le 16|\Omega|^4$. From this it follows

$$
\begin{aligned}
&\mathbb{E}\left[\left|\int\langle T_{\mu_0}^{\mu_1}-T_{\mu_0}^{\eta},T_{\mu_0}^{\mu_2}-T_{\mu_0}^{\eta}\rangle d\eta-\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1}(X_i)-T_{\mu_0}^{\eta}(X_i),T_{\mu_0}^{\mu_2}(X_i)-T_{\mu_0}^{\eta}(X_i)\rangle\right|\right]\\
=&\mathbb{E}\left[\left|\mathbb{E}_{X\sim\eta}[h(X)]-\frac{1}{n}\sum_{i=n+1}^{2n}h(X_i)\right|\right]\\
\le&\sqrt{\frac{\mathrm{Var}[h(X)]}{n}}\le\sqrt{\frac{16|\Omega|^4}{n}}\lesssim\frac{1}{\sqrt{n}}
\end{aligned}
$$

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

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

We now proceed to the second term in (20).

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Bigg|\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1}(X_i)-T_{\mu_0}^{\eta}(X_i),T_{\mu_0}^{\mu_2}(X_i)-T_{\mu_0}^{\eta}(X_i)\rangle-\\
&\frac{1}{n}\sum_{i=n+1}^{2n}\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i),T_{\mu_0}^{\mu_2,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i)\rangle\Bigg|\Bigg]\\
\leq&\frac{1}{n}\sum_{i=n+1}^{2n}\mathbb{E}\Bigg[\Bigg|\langle T_{\mu_0}^{\mu_1}(X_i)-T_{\mu_0}^{\eta}(X_i),T_{\mu_0}^{\mu_2}(X_i)-T_{\mu_0}^{\eta}(X_i)\rangle-\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i),T_{\mu_0}^{\mu_2,n,\varepsilon}(X_i)-T_{\mu_0}^{\eta,n,\varepsilon}(X_i)\rangle\Bigg|\Bigg]\\
=&\mathbb{E}\Bigg[\Bigg|\langle T_{\mu_0}^{\mu_1}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1}),T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1})\rangle-\\
&\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}),T_{\mu_0}^{\mu_2,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1})\rangle\Bigg|\Bigg]\\
=&\mathbb{E}\Bigg[\Bigg|\langle(T_{\mu_0}^{\mu_1}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1}))-(T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1})),T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1})\rangle\\
&-\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}),(T_{\mu_0}^{\mu_2,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}))-(T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1}))\rangle\Bigg|\Bigg]\\
\leq&\mathbb{E}\Bigg[\Bigg|\langle(T_{\mu_0}^{\mu_1}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1}))-(T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1})),T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1})\rangle\Bigg|\Bigg]\\
+&\mathbb{E}\Bigg[\Bigg|\langle T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}),(T_{\mu_0}^{\mu_2,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}))-(T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1}))\rangle\Bigg|\Bigg]\\
\leq&\mathbb{E}\Big[\|T_{\mu_0}^{\mu_1}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1})-(T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}))\|\Big]\mathbb{E}\Big[\|T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1})\|\Big]\\
+&\mathbb{E}\Big[\|T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1})\|\Big]\mathbb{E}\Big[\|(T_{\mu_0}^{\mu_2,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}))-(T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1}))\|\Big]\\
\leq&|\Omega|\mathbb{E}\Big[\|T_{\mu_0}^{\mu_1}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1})-(T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}))\|\Big]\\
+&|\Omega|\mathbb{E}\Big[\|(T_{\mu_0}^{\mu_2,n,\varepsilon}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1}))-(T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\eta}(X_{n+1}))\|\Big]\\
\leq&|\Omega|\Big(\mathbb{E}\Big[\|T_{\mu_0}^{\mu_1}(X_{n+1})-T_{\mu_0}^{\mu_1,n,\varepsilon}(X_{n+1})\|\Big]+\mathbb{E}\Big[\|T_{\mu_0}^{\eta}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1})\|\Big]\\
+&\mathbb{E}\Big[\|T_{\mu_0}^{\mu_2}(X_{n+1})-T_{\mu_0}^{\mu_2,n,\varepsilon}(X_{n+1})\|\Big]+\mathbb{E}\Big[\|T_{\mu_0}^{\eta}(X_{n+1})-T_{\mu_0}^{\eta,n,\varepsilon}(X_{n+1})\|\Big]\Big).
\end{aligned}
$$

Each of these four expectations can be controlled by applying Jensen's inequality followed by Theorem 2 to give the bound $\lesssim\sqrt{n^{-\frac{\bar{\alpha}+1}{2(d+\bar{\alpha}+1)}}\log(n)}=n^{-\frac{\bar{\alpha}+1}{4(d+\bar{\alpha}+1)}}\sqrt{\log(n)}$. The result follows. □

### C.3 Proof of Corollary 1

*Proof.* The proof is similar to one found in (Werenski et al., 2022) and we include it here for completeness. Let $B_n$ denote the entrywise bound in Theorem 3. Noting that $\hat{\lambda}^T \hat{A}^L \hat{\lambda} \leq \lambda_*^T \hat{A}^L \lambda_*$ by construction, we estimate

$$\begin{aligned}
\mathbb{E}[\hat{\lambda}^T A^L \hat{\lambda}] &= \mathbb{E}\left[\hat{\lambda}^T (A^L - \hat{A}^L)\hat{\lambda}\right] + \mathbb{E}[\hat{\lambda}^T \hat{A}^L \hat{\lambda}] \\
&\leq \mathbb{E}\left[|\hat{\lambda}^T (A^L - \hat{A}^L)\hat{\lambda}|\right] + \mathbb{E}[\lambda_*^T \hat{A}^L \lambda_*] \\
&= \mathbb{E}\left[|\hat{\lambda}^T (A^L - \hat{A}^L)\hat{\lambda}|\right] + \mathbb{E}[\lambda_*^T (\hat{A}^L - A^L) \lambda_*] \\
&= \mathbb{E}\left[\left|\sum_{i,j=1}^m (\hat{\lambda})_i (\hat{\lambda})_j (A^L - \hat{A}^L)_{ij}\right|\right] + \mathbb{E}\left[\left|\sum_{i,j=1}^m (\lambda_*)_i (\lambda_*)_j (A^L - \hat{A^L})_{ij}\right|\right] \\
&\leq \mathbb{E}\left[\sum_{i,j=1}^m (\hat{\lambda})_i (\hat{\lambda})_j |A^L_{ij} - \hat{A}^L_{ij}|\right] + \mathbb{E}\left[\sum_{i,j=1}^m (\lambda_*)_i (\lambda_*)_j |A^L_{ij} - \hat{A}^L_{ij}|\right] \\
&\leq 2\mathbb{E}\left[\sum_{i,j=1}^m |A^L_{ij} - \hat{A}^L_{ij}|\right] = 2\sum_{i,j=1}^m \mathbb{E}[|A^L_{ij} - \hat{A}^L_{ij}|] \lesssim 2m^2 B_n
\end{aligned}$$

In the second line we have used the assumption that $\lambda_*^T A^L \lambda_* = 0$. The second to last line uses the triangle inequality. The last line uses $\lambda_*, \hat{\lambda} \in \Delta^m$ so their entries in $[0, 1]$ and then concludes with Theorem 3.

Since $A^L$ is positive semidefinite and by assumption $\lambda_* \in \Delta^m$ satisfies $\lambda_*^T A^L \lambda_* = 0$, it follows that $\lambda_*$ is an eigenvector of $A^L$ with eigenvalue 0. Let $0 < a_2 \leq \cdots \leq a_m$ be the non-zero eigenvalues of $A^L$ with associated orthonormal eigenvectors $v_2, \ldots, v_m$. 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 $\{v_2, \ldots, v_m\}$. Note that $\hat{\beta}$ and $\hat{\lambda}_\perp$ are random. Then,

$$\begin{aligned}
&\mathbb{E}[\|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2] \\
=&\mathbb{E}[\|\hat{\lambda}_\perp\|_2^2] \\
=&\mathbb{E}\left[\sum_{i=2}^m |v_i^T \hat{\lambda}_\perp|^2\right] \\
\leq&\frac{1}{a_2}\mathbb{E}\left[\sum_{i=2}^m a_i |v_i^T \hat{\lambda}_\perp|^2\right] \\
=&\frac{1}{a_2}\mathbb{E}[|(\hat{\lambda}_\perp)^T A^L \hat{\lambda}_\perp|] \\
=&\frac{1}{a_2}\mathbb{E}[|\hat{\lambda}^T A^L \hat{\lambda}|] \\
\lesssim&\frac{2m^2}{a_2} B_n.
\end{aligned}$$

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

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

which implies that $\mathbb{E}[(1 - \hat{\beta})^2] \leq m\mathbb{E}[\|\hat{\lambda}_\perp\|_2^2] \lesssim \frac{2m^3}{a_2} B_n$.

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

$$\begin{aligned}
\mathbb{E}[\|\hat{\lambda} - \lambda_*\|_2^2] &= \mathbb{E}[\|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2 + \|(\hat{\beta} - 1)\lambda_*\|_2^2] \\
&= \mathbb{E}[\|\hat{\lambda} - \hat{\beta}\lambda_*\|_2^2] + \mathbb{E}[\|(\hat{\beta} - 1)\lambda_*\|_2^2] \\
&\lesssim \frac{2m^2}{a_2} B_n + \mathbb{E}[(\hat{\beta} - 1)^2 \|\lambda_*\|_2^2] \\
&\leq \frac{2m^2}{a_2} B_n + \mathbb{E}[(\hat{\beta} - 1)^2] \\
&\lesssim \frac{2m^2}{a_2} B_n + \frac{2m^3}{a_2} B_n
\end{aligned}$$

as desired. □

**Remark 5.** As shown in the proof, the implicit constant in Corollary 1 depends on the second eigenvalue $a_2$ of $A^L$ as $1/a_2$. The choice of $\mu_0$ may be especially important to ensuring that the second eigenvalue is not too small.

# D PROOFS FROM SECTION 5

## D.1 Proof of Proposition 4

*Proof.* That an element of $W_2\text{BCM}(\{\mu_i\}_{i\in I})$ is a Gaussian is a consequence of Theorem 3.10 in (Álvarez-Esteban et al., 2018). That $\text{LBCM}(\{\mu_i\}_{i\in I}; \mu_0)$ is a Gaussian follows from the fact $T_{\mu_0}^{\mu_i}$ is an affine transformation and therefore $\int \lambda T_{\mu_0}^{\mu_i} d\lambda(i)$ is also affine, and the family of Gaussian measures is closed under affine transformations. The latter fact is well-known, see (Goodfellow et al., 2016). □

## D.2 Proof of Theorem 4

We begin by establishing a preliminary result. This serves the role of a Choquet-type theorem in that it identifies the set of extreme points in a convex set and shows that every point in the set is a convex combination of the extreme points.

**Proposition 7.** *Let $\mathcal{T}([0,1])$ denote the set of functions*

$$\mathcal{T}([0,1]) \triangleq \{T : [0,1] \to [0,1] \mid T \text{ is increasing}\}.$$

*The set $\mathcal{T}([0,1])$ is convex. Let $T_a(x) \triangleq \mathbf{1}[x \geq a]$. Then for every $T \in \mathcal{T}([0,1])$ there is a measure $\lambda \in \mathcal{P}([0,1])$ such that for $U([0,1])$-almost every $x$ it holds*

$$T(x) = \int T_a(x) d\lambda(a).$$

*Proof.* Let $T \in \mathcal{T}([0,1])$. Since $T$ is monotonic at each point in the domain of $T$ there exists well-defined left and right limit. We can define a right continuous version of $T$, say $T_+$ for every $x \in [0,1)$ by

$$T_+(x) = \lim_{y \to x^+} T(y)$$

and set $T_+(1) = T(1)$.

Since $T$ is monotonic it has only countably many discontinuities and $T_+$ agrees with $T$, except possibly at these discontinuities, the functions $T$ and $T_+$ disagree on a countable set which is of measure zero. Now define the measure $\lambda$ by

$$\lambda([0,x]) = T_+(x), \ x \in [0,1) \qquad \lambda(\{1\}) = 1 - T_+(1).$$

One can show that this satisfies all the required properties to be a probability measure over $[0,1]$.

For $x \in [0,1)$ we have

$$\int_{[0,1]} \mathbf{1}[a \leq x] d\lambda(a) = \int_{[0,x]} 1 d\lambda(a) = \lambda([0,x]) = T_+(x).$$

Since $T_+$ agrees with $T$ for almost every $x$ and the set $\{1\}$ is of measure zero we have established that for almost every $x$

$$\int_{[0,1]} \mathbf{1}[a \leq x] d\lambda(a) = T(x)$$

and therefore the proposed $\lambda$ satisfies the claim.

Since $T$ was chosen arbitrarily in $\mathcal{T}([0,1])$, the result holds over the entire class of functions. □

We will also leverage the following two known results from (Panaretos and Zemel, 2020).

**Theorem 8** ((Panaretos and Zemel, 2020) Theorem 3.1.5)**.** *Let $m \in \mathbb{N}$ and let $\lambda = (1/m, ..., 1/m)$. Suppose that $W_2(\mu_i^k, \mu_i) \to 0$ for $i = 1, ..., m$ and let $\mu_\lambda^k$ be the Wasserstein barycenter with parameter $\lambda$ for $(\mu_1^k, ..., \mu_m^k)$. Then any limit point of the sequence $\{\mu_\lambda^k\}_{k=1}^\infty$ is the Wasserstein barycenter of $\mu_1, ..., \mu_m$ with coordinate $\lambda$.*

**Corollary 2.** *Let $m \in \mathbb{N}$ and let $\lambda \in \mathbb{Q}^m \cap \Delta^m$. Suppose that $W_2(\mu_i^k, \mu_i) \to 0$ for $i = 1, ..., m$ and let $\mu_\lambda^k$ be the Wasserstein barycenter with parameter $\lambda$. Then any limit point of the sequence $\{\mu_\lambda^k\}_{k=1}^\infty$ is the Wasserstein barycenter of $\mu_1, ..., \mu_m$ with coordinate $\lambda$.*

*Proof.* This follows from Theorem 8 by repeating the measures $\mu_i$ as necessary. □

**Theorem 9** ((Panaretos and Zemel, 2020) Theorem 3.1.9)**.** *Let $m \in \mathbb{N}$ and let $\lambda = (1/m, ..., 1/m)$. Suppose that $\mu_0, ..., \mu_m$ are compatible. Then the Wasserstein barycenter $\mu_\lambda$ is given by*

$$\mu_\lambda = \left( \frac{1}{m} \sum_{i=1}^m T_i \right) \# \mu_0,$$

*where $T_i$ is the optimal transport map from $\mu_0$ to $\mu_i$.*

**Corollary 3.** *Let $m \in \mathbb{N}$ and let $\lambda \in \mathbb{Q}^m \cap \Delta^m$. Suppose that $\mu_0, ..., \mu_m$ are compatible. Then the Wasserstein barycenter $\mu_\lambda$ is given by*

$$\mu_\lambda = \left( \sum_{i=1}^m \lambda_i T_i \right) \# \mu_0,$$

*where $T_i$ is the optimal transport map from $\mu_0$ to $\mu_i$.*

*Proof.* This follows from Theorem 9 by repeating the measures $\mu_i$ as necessary. □

We can now proceed to the main proof.

*Proof of Theorem 4.* To each measure $\mu \in \mathcal{P}([0,1])$ associate to it the optimal transport map from $U([0,1])$ to $\mu$, denoted by $T$. The map $T$ is almost everywhere uniquely defined and is contained in the sub-differential of a convex function. In addition $T$ is an increasing function and the image of $[0,1]$ under $T$ is contained in $[0,1]$.

By Proposition 7 there exists a measure $\lambda \in \mathcal{P}([0,1])$ such that for $U([0,1])$ almost every $x$

$$T(x) = \int_{[0,1]} \mathbf{1}[a \leq x] d\lambda(a).$$

Note that the measures $\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\}$ have optimal transport maps given precisely by $\{x \mapsto \mathbf{1}[x \geq a] \mid a \in [0,1]\}$. Denote $a\delta_0 + (1-a)\delta_1$ by $\mu_a$ and $T_a$ the corresponding transport map from $U([0,1])$ to $\mu_a$. With this notation it holds for almost every $x$ that

$$T(x) = \int_{[0,1]} T_a(x) d\lambda(a). \tag{21}$$

This implies that for every bounded continuous function $f$ we have

$$\begin{aligned}\int f(x)d\mu(x) &= \int f(T(x))dU([0,1])(x)\\ &= \int f\left(\int_{[0,1]} T_a(x)d\lambda(a)\right)dU([0,1])(x)\\ &= \int f(x)d\left[\left(\int_{[0,1]} T_a(x)d\lambda(a)\right)\#U([0,1])\right](x)\end{aligned}$$

where the first equality holds because $T\#U([0,1]) = \mu$, the second holds because of (21), and the third holds by definition of the push-forward of a measure. This shows that with respect to weak convergence the measure $\mu = \left(\int_{[0,1]} T_a(x)d\lambda(a)\right)\#U([0,1])$ and therefore is contained in $\mathrm{LBCM}(U([0,1]);\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\})$. Since this $\mu$ was arbitrary we see that every measure in $\mathcal{P}([0,1])$ is indeed contained in $\mathrm{LBCM}(U([0,1]);\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\})$.

For the weak convergence of $W_2\mathrm{BCM}(\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\})$ we would like to appeal to compatibility but cannot immediately do so because the measures do not even have optimal transport maps. However, one can subvert this problem by an approximation argument and using Corollaries 2 and 3. We will establish that the set

$$\left\{\sum_{i=1}^m a_i\delta_{b_i} \mid m \in \mathbb{N}, a_i, b_i \in \mathbb{Q}\cap[0,1], \sum_{i=1}^m a_i = 1\right\} \subseteq W_2\mathrm{BCM}(\{a\delta_0 + (1-a)\delta_1 \mid a \in [0,1]\}) \tag{22}$$

and the former set is clearly dense with respect to weak convergence.

We introduce the family of measures parameterized by $a \in [0,1], b \in (0,1/2)$:

$$\nu_{a,b} = aU([0,b]) + (1-a)U([1-b,b]).$$

The measures $\nu_{a,b}$ are absolutely continuous and satisfy

$$\lim_{b\to 0^+} W_2\left(a\delta_0 + (1-a)\delta_1, \nu_{a,b}\right) = 0. \tag{23}$$

In addition, the optimal transport map from $U([0,1])$ to $\nu_{a,b}$, which we denote by $T_{a,b}$ for $a \in (0,1)$, is given by

$$T_{a,b}(x) = \begin{cases}\frac{b}{a}x & x \le a\\ (1-b) + \frac{b}{1-a}(x-a) & x > a\end{cases}$$

while $T_{1,b} = bx$ and $T_{0,b} = 1 - b + bx$. These maps satisfy for $a > 0$

$$\lim_{b\to 0^+} T_{a,b} = \mathbf{1}[x > a]$$

as well as the bound

$$\max_{x\in[0,1]} |T_{a,b}(x) - \mathbf{1}[x > a]| \le b.$$

For $a = 0$ they satisfy

$$\lim_{b\to 0^+} T_{0,b} = 1$$

as well as the bound

$$\max_{x\in[0,1]} |T_{0,b}(x) - 1| \le b.$$

This bound further implies for any $(\lambda_0, ..., \lambda_m) \in \Delta^{m+1}$

$$\begin{aligned}\max_{x\in[0,1]}\left|\left(\sum_{i=0}^m \lambda_i T_{i/m,b}(x)\right) - \left(\sum_{i=1}^m \lambda_i\mathbf{1}[x > i/m] + \lambda_0\cdot 1\right)\right| &\le \max_{x\in[0,1]} \lambda_0|T_{0,b}(x) - 1| + \sum_{i=1}^m \lambda_i|T_{i/m,b}(x) - \mathbf{1}[x > i/m]|\\ &\le \max_{x\in[0,1]} \lambda_0 b + \sum_{i=1}^m \lambda_i b = b.\end{aligned}$$

This bound further implies for any $(\lambda_0, ..., \lambda_m) \in \Delta^{m+1}$ that

$$\begin{aligned} &W_2^2\left(\left(\sum_{i=0}^m \lambda_i T_{i/m,b}\right)\#U([0,1]), \left(\sum_{i=1}^m \lambda_i \mathbf{1}[x > i/m] + \lambda_0 \cdot 1\right)\#U([0,1])\right) \\ &\leq \int \left(\left(\sum_{i=0}^m \lambda_i T_{i/m,b}\right) - \left(\sum_{i=1}^m \lambda_i \mathbf{1}[x > i/m] + \lambda_0 \cdot 1\right)\right)^2 dU([0,1])(x) \\ &\leq \int b^2 dU([0,1]) = b^2 \end{aligned} \tag{24}$$

where the first inequality is realized by the choose of coupling given by

$$\pi = \left(\left(\sum_{i=0}^m \lambda_i T_{i/m,b}\right), \left(\sum_{i=1}^m \lambda_i \mathbf{1}[x > i/m] + \lambda_0 \cdot 1\right)\right)\#U([0,1]).$$

Next observe that because $U([0,1])$ and $\nu_{a,b}$ for $b > 0$ are absolutely continuous and one-dimensional they are compatible and therefore by Corollary 3 for $(\lambda_0, ..., \lambda_m) \in \mathbb{Q}^{m+1} \cap \Delta^{m+1}$ it holds that the Wasserstein barycenter with coordinate $\lambda$ of $(\nu_{0,b}, \nu_{1/m,b}, ..., \nu_{(m-1)/m,b}, \nu_{1,b})$ is given by

$$\left(\sum_{i=0}^m \lambda_i T_{i/m,b}\right)\#U([0,1]).$$

We are now able to apply Corollary 2 to $\lambda \in \mathbb{Q}^{m+1} \cap \Delta^{m+1}$. This is done with $\mu_i^k = \nu_{i/m,1/k}$ and $\mu_i = (i/m)\delta_0 + (1 - i/m)\delta_1$. In (23) we have established that

$$\lim_{k\to\infty} W_2(\mu_i^k, \mu_i) = \lim_{k\to\infty} W_2(\nu_{i/m,1/k}, (i/m)\delta_0 + (1 - i/m)\delta_1) = 0.$$

We have also established that the barycenter of $(\mu_0^k, ..., \mu_m^k) = (\nu_{0/m,1/k}, ..., \nu_{m/m,1/k})$ is given by

$$\left(\sum_{i=0}^m \lambda_i T_{i/m,1/k}\right)\#U([0,1]).$$

and by (24)

$$\lim_{k\to\infty} W_2\left(\left(\sum_{i=0}^m \lambda_i T_{i/m,1/k}\right)\#U([0,1]), \left(\sum_{i=1}^m \lambda_i \mathbf{1}[x > i/m] + \lambda_0 \cdot 1\right)\#U([0,1])\right) \leq \lim_{k\to\infty} \frac{1}{k} = 0.$$

Therefore by Corollary 2 the barycenter of $(\mu_0, ..., \mu_m) = (\delta_1, (1/m)\delta_0 + (1 - 1/m)\delta_1, ..., (1 - 1/m)\delta_0 + (1/m)\delta_1, \delta_0)$ with coordinate $\lambda$ is given by

$$\left(\sum_{i=1}^m \lambda_i \mathbf{1}[x > i/m] + \lambda_0 \cdot 1\right)\#U([0,1])$$

which we can expand as

$$\left(\sum_{i=1}^m \lambda_i \mathbf{1}[x > i/m] + \lambda_0 \cdot 1\right)\#U([0,1]) = \sum_{i=1}^m \frac{1}{m}\delta_{\sum_{j=0}^{i-1}\lambda_j}. \tag{25}$$

It is a straightforward exercise to show that every set in (22) can be expressed in the form of (25) by taking $m$ sufficiently large and appropriately setting $\lambda$.

That $\text{conv}(\{U([0,1])\} \cup \{a\delta_0 + (1 - a)\delta_1 \mid a \in [0,1]\})$ is not dense follows from the fact that it only consists of measures of the form

$$\lambda_1\delta_0 + \lambda_2\delta_1 + (1 - \lambda_1 - \lambda_2)U([0,1]),$$

where $(\lambda_1, \lambda_2, \lambda_3) \in \Delta^3$. □

### D.3 Proof of Proposition 5

*Proof.* That $\mathcal{V}(C)$ consists of the optimal transport maps from $U(C)$ to $\mathcal{P}(\{v_i\}_{i=1}^\ell)$ follows from the fact that for every $T \in \mathcal{V}(C)$ the measure $T\#U(C) \in \mathcal{P}(\{v_i\}_{i=1}^\ell)$ since by definition of $\mathcal{V}(C)$ it holds that $T(x) \in \{v_i\}_{i=1}^\ell$ almost surely. In addition for every $\mu \in \mathcal{P}(\{v_i\}_{i=1}^\ell)$ the optimal transport map $T$ (which must be in $\mathcal{T}(C)$ by Theorem 1) from $U(C)$ to $\mu$ must be in $\mathcal{V}(C)$ since it must hold with probability 1 that $T(x) \in \{v_i\}_{i=1}^\ell$ in order for $T$ to transport $U(C)$ to $\mu$.

To establish that $\mathcal{V}(C)$ are extreme in $\mathcal{T}(C)$, for each $T \in \mathcal{V}(C)$ we will construct a continuous linear functional $F_T$ on $\mathcal{T}(C)$ such that

$$T = \operatorname*{arg\,max}_{T' \in \mathcal{T}(C)} F_T(T').$$

This will demonstrate that $T$ is not only extreme but *exposed* in $\mathcal{T}(C)$.

In order to give the function $F_T$ we will use that in a convex polytope $\text{conv}(\{v_i\}_{i=1}^\ell)$ in $\mathbb{R}^d$ for every extreme point there exists $v_i$ there exists at least one vector $u_i \in \mathbb{R}^d$ such that

$$v_i = \operatorname*{arg\,max}_{v \in \text{conv}(\{v_i\}_{i=1}^\ell)} \langle v, u_i \rangle$$

and the maximum is achieved uniquely.

Now let $\{u_i\}_{i=1}^\ell$ be such a set of vectors for $\{v_i\}_{i=1}^\ell$. In addition define the sets $R_1, ..., R_\ell \subset C$ according to

$$R_i \triangleq T^{-1}(v_i).$$

These sets are disjoint and since $T(x) \in \{v_i\}_{i=1}^\ell$ almost surely it holds that

$$\sum_{i=1}^\ell U(C)[R_i] = 1.$$

Finally define the function $M_T : \mathbb{R}^d \to \mathbb{R}^d$ by

$$M_T(x) = \sum_{i=1}^\ell \mathbf{1}[x \in R_i] u_i.$$

The linear functional we will use is given by

$$F_T(T') = \int \langle T'(x), M_T(x) \rangle d[U(C)](x).$$

This is clearly linear by the linearity of inner products and integration. We must show that $T$ is the unique maximizer of this functional.

To see this, we have for every $T' \in \mathcal{T}(C)$

$$\begin{aligned}
F_T(T') &= \int \langle T'(x), M_T(x) \rangle d[U(C)](x) \\
&= \sum_{i=1}^\ell \int_{R_i} \langle T'(x), M_T(x) \rangle d[U(C)](x) \\
&= \sum_{i=1}^\ell \int_{R_i} \langle T'(x), u_i \rangle d[U(C)](x) \\
&\leq \sum_{i=1}^\ell \int_{R_i} \langle v_i, u_i \rangle d[U(C)](x) \\
&= \sum_{i=1}^\ell \langle v_i, u_i \rangle U(C)[R_i]
\end{aligned}$$

The inequality follows from the definition of $u_i$ and the fact that $T' \in \mathcal{T}(C)$ and therefore

$$T'(x) \in C \implies \langle T'(x), u_i \rangle \leq \max_{x \in C} \langle x, u_i \rangle = \langle v_i, u_i \rangle.$$

The equality case happens if and only if for every $i = 1, ..., \ell$ almost every $x \in R_i$ it holds that $T'(x) = v_i = T(x)$ which demonstrates that indeed $T$ is the unique (up to almost everywhere equality) maximizer of $F_T$ over $\mathcal{T}(C)$ and is therefore exposed which implies it is extreme.

Since $T$ was chosen arbitrarily in $\mathcal{V}(C)$ we can conclude that the entire set $\mathcal{V}(C)$ must be extreme. □

### D.4 Proof of Theorem 5

The polytope that we use is the set $C_0 = \text{conv}\left(\{(0,0),(0,1),(1,0)\}\right)$ and the convex function is given by $\phi_0(x,y) = \frac{1}{4}\frac{x^2}{(2-y)}$. One can directly compute the Hessian of this function and check that it is positive definite and also show that $\nabla\phi_0(C_0) \subset C_0$ so that $\nabla\phi_0 \in \mathcal{T}(C_0)$. To establish that $\nabla\phi_0$ is not contained in $\text{conv}(\mathcal{V}(C_0))$, we must characterize the structure of the functions which are contained in the set $\mathcal{V}(C_0)$.

**Lemma 1.** *A function $T$ is contained in $\mathcal{V}(C_0)$ if and only if there exist a vector $b \in \mathbb{R}^3$ such that $T(x) \in \partial v(x;b)$ where*

$$v(x;b) = \max_{i \in \{1,2,3\}} \langle v_i, x \rangle + b_i$$

*with $v_1 = (0,0), v_2 = (0,1), v_3 = (1,0)$.*

*Proof.* Clearly if $T(x) \in \partial v(x;b)$ for some $b \in \mathbb{R}^d$ then $T \in \mathcal{V}(C_0)$. For the reverse direction let $\mu = T\#U(C_0)$ so that $\mu = a_1\delta_{v_1} + a_2\delta_{v_2} + a_3\delta_{v_3}$, which is the case because $T \in \{v_1, v_2, v_3\}$ for $U(C_0)$ by definition of $\mathcal{V}(C_0)$. Since it is also the case that $T$ is contained in the *sub-differential* of a convex function by Theorem 1 it must be the case that $T$ is the optimal transport map from $U(C_0)$ to $\mu$. We refer to (Peyré and Cuturi, 2020) Section 5.2 for proof that this implies $T$ has the claimed form. □

**Lemma 2.** *Let $T \in \mathcal{V}(C_0)$. The following holds for all but at most one $a_1 \in [0,1]$. For all $0 \leq a_2 < a_2' \leq 1 - a_1$ we have*

$$\langle e_1, T((a_1, a_2)) \rangle \geq \langle e_1, T((a_1, a_2')) \rangle.$$

*Proof.* By Lemma 1 there exists some $b \in \mathbb{R}^3$ such that $T(x) \in \partial v(x;b)$. Define the sets

$$R_1 = \{x \in \mathbb{R}^2 \mid \langle x, v_1 \rangle + b_1 = v(x;b)\}$$
$$R_2 = \{x \in \mathbb{R}^2 \mid \langle x, v_2 \rangle + b_2 = v(x;b)\}$$
$$R_3 = \{x \in \mathbb{R}^2 \mid \langle x, v_3 \rangle + b_3 = v(x;b)\}$$

These are illustrated in Figure 3. With these sets the sub-differential of $v(x;b)$ over all of $\mathbb{R}^2$ can be explicitly written as

$$\partial v(x;b) = \begin{cases} \{(0,0)\} & x \in \text{int}(R_1) \\ \{(0,1)\} & x \in \text{int}(R_2) \\ \{(1,0)\} & x \in \text{int}(R_3) \\ \text{conv}(\{(0,0),(0,1)\}) & x \in (R_1 \cap R_2) \setminus R_3 \\ \text{conv}(\{(0,0),(1,0)\}) & x \in (R_1 \cap R_3) \setminus R_2 \\ \text{conv}(\{(0,1),(1,0)\}) & x \in (R_2 \cap R_3) \setminus R_1 \\ \text{conv}(\{(0,0),(0,1),(1,0)\}) & x \in R_1 \cap R_2 \cap R_3 \end{cases}$$

The set $R_1 \cap R_2 \cap R_3$ consists of the single point $\{(b_1 - b_3, b_1 - b_2)\}$. If $b_1 - b_3 \in [0,1)$ then $a_1 = b_1 - b_3$ is the single value where it is possible that there are $0 \leq a_2' < a_2 \leq 1 - (b_1 - b_3)$

$$\langle e_1, T((a_1, a_2)) \rangle \geq \langle e_1, T((a_1, a_2')) \rangle.$$

Otherwise if $a_1 < b_1 - b_3$ and $0 \leq a_2 < a_2' \leq 1 - a_1$ then $(a_1, a_2), (a_1, a_2') \notin R_3$ and there are five possible configurations:

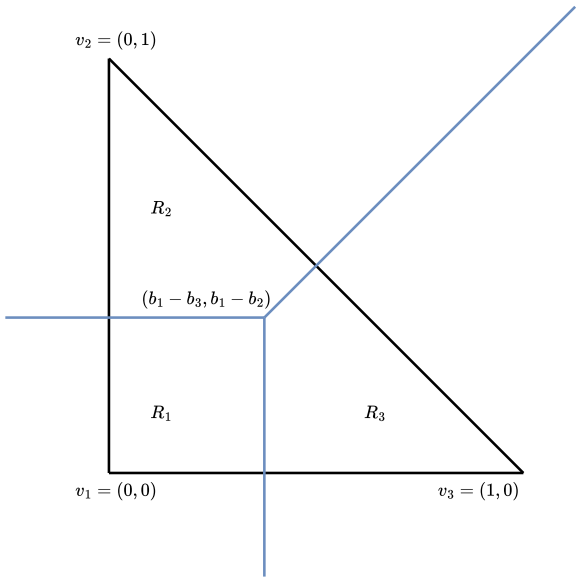


Figure 3: Structure of $C_0$ and the sets $R_1, R_2, R_3$. Each set contains its boundary (shown in blue) and the intersection point is determined by $b$.

1. $(a_1, a_2) \in \text{int}(R_1)$ and $(a_1, a_2') \in \text{int}(R_1)$,
2. $(a_1, a_2) \in \text{int}(R_1)$ and $(a_1, a_2') \in R_1 \cap R_2$,
3. $(a_1, a_2) \in \text{int}(R_1)$ and $(a_1, a_2') \in \text{int}(R_2)$,
4. $(a_1, a_2) \in R_1 \cap R_2$ and $(a_1, a_2') \in \text{int}(R_2)$,
5. $(a_1, a_2) \in \text{int}(R_2)$ and $(a_1, a_2') \in \text{int}(R_2)$.

One can check explicitly using the formula for the sub-differential that in each of these cases that $\langle e_1, T((a_1, a_2))\rangle = \langle e_1, T((a_1, a_2'))\rangle = 0$. If $a_1 > b_1 - b_3$ then $(a_1, a_2), (a_1, a_2') \notin R_1$ and there are again five cases.

1. $(a_1, a_2) \in \text{int}(R_3)$ and $(a_1, a_2') \in \text{int}(R_3)$
2. $(a_1, a_2) \in \text{int}(R_3)$ and $(a_1, a_2') \in R_2 \cap R_3$
3. $(a_1, a_2) \in \text{int}(R_3)$ and $(a_1, a_2') \in \text{int}(R_2)$
4. $(a_1, a_2) \in R_2 \cap R_3$ and $(a_1, a_2') \in \text{int}(R_2)$
5. $(a_1, a_2) \in \text{int}(R_2)$ and $(a_1, a_2') \in \text{int}(R_2)$

In the first three cases

$$\langle e_1, T((a_1, a_2))\rangle = 1 \geq \langle e_1, T((a_1, a_2'))\rangle,$$

in the fourth

$$\langle e_1, T((a_1, a_2))\rangle \geq 0 = \langle e_1, T((a_1, a_2'))\rangle,$$

and in the fifth

$$\langle e_1, T((a_1, a_2))\rangle = 0 = \langle e_1, T((a_1, a_2'))\rangle.$$

This completes the proof. □

Lemma 2 and the following lemma are the two main tools used to show that the function $\phi_0$ is a counterexample.

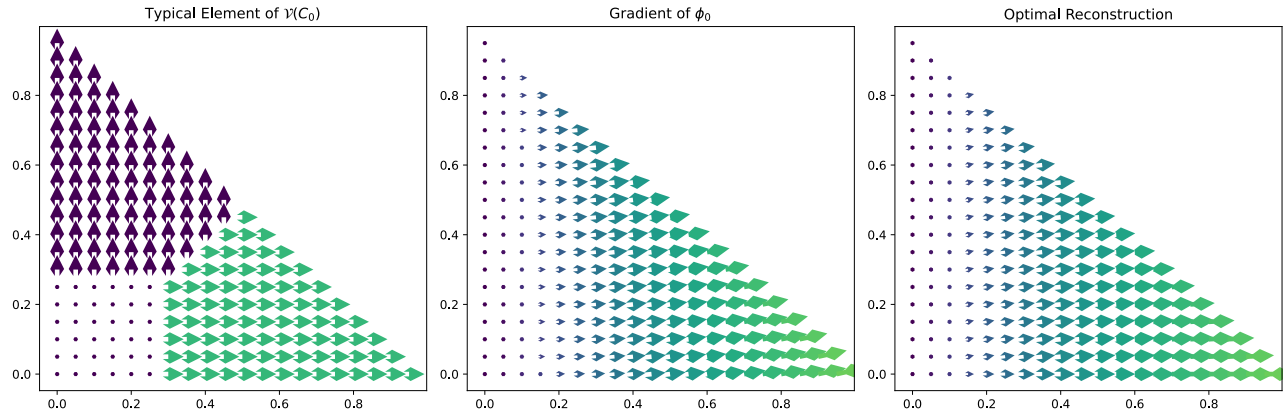


Figure 4: Counter example in Theorem 5. On the left we have a typical element of $\mathcal{V}(C_0)$, in the middle the gradient of $\phi_0$, and on the right a numerically obtained optimal reconstruction of $\phi_0$ using a convex combination of elements in $\mathcal{V}(C_0)$ represented as vector fields. Arrow color corresponds to the magnitude of the first coordinate.

**Lemma 3.** *For $x \in [0,1]$ and $0 \leq y_1 < y_2 \leq 1 - x$ it holds that*

$$\frac{x}{8}(y_2 - y_1) \leq \langle e_1, \nabla\phi_0(x, y_2) - \nabla\phi_0(x, y_1)\rangle.$$

Before proceeding to the proof we remark that the maps in Lemma 2 behave in an *opposite way* to $\nabla\phi_0$. Specifically, fix a map $T$ as in Lemma 2 and let $x \in [0,1]$ take any value besides $x = 1$ and the single point of issue in Lemma 2 (if it exists). Along the segment $\{(x, y) \mid 0 \leq y \leq 1 - x\}$ the inner product with $e_1$ *decreases* for $T$ as $y$ increases, while the inner product of $e_1$ and $\nabla\phi_0$ *increases* as $y$ increases. These properties will be leveraged in the proof of Theorem 5 below.

*Proof.* Fix an $x \in [0,1]$ and a $y \in [0, 1-x]$ and compute the partial derivative

$$\frac{\partial}{\partial y}\langle e_1, \nabla\phi_0(x, y)\rangle = \frac{x}{2(2-y)^2}.$$

Since $y \in [0,1]$ we have

$$\frac{x}{8} \leq \frac{x}{2(2-y)^2}.$$

From this it follows

$$\begin{aligned}
\langle e_1, \nabla\phi_0(x, y_2)\rangle &= \langle e_1, \nabla\phi_0(x, y_1)\rangle + \int_{y_1}^{y_2} \frac{\partial}{\partial y}\langle e_1, \nabla\phi_0(x, y)\rangle dy \\
&\geq \langle e_1, \nabla\phi_0(x, y_1)\rangle + \int_{y_1}^{y_2} \frac{\partial}{\partial y}\frac{x}{8} dy \\
&= \langle e_1, \nabla\phi_0(x, y_1)\rangle + \frac{x}{8}(y_2 - y_1).
\end{aligned}$$

Re-arranging proves the bound. □

The essence of Lemmas 2 and 3 are visualized in Figure 4. Along any vertical strip, when moving from bottom to top, the first component cannot increase for an element of $\mathcal{V}(C_0)$ while for $\nabla\phi_0$ it is strictly increasing. The optimal reconstruction fails to capture this behavior and has constant first component along any vertical line. The proof of Theorem 5 makes this precise and gives a lower bound on the amount of error. Before showing it we must establish a technical lemma which will dispose of many edge cases in the proof of Theorem 5.

**Lemma 4.** *Let $T \in \text{conv}(\mathcal{V}(C_0))$. Then for almost every $a_1 \in [0,1]$ it holds that for all $0 \leq a_2 < a_2' \leq 1 - a_1$ we have*

$$\langle e_1, T((a_1, a_2))\rangle \geq \langle e_1, T((a_1, a_2'))\rangle.$$

*Proof.* Let $T \in \text{conv}(\mathcal{V}(C_0))$ and let $\lambda \in \mathcal{P}(\mathcal{V}(C_0))$ be such that

$$T = \int_{\mathcal{V}(C_0)} T_i d\lambda(T_i).$$

The choice of $T_i$ can be interpreted as the optimal transport map to the corresponding measure $\mu_i$, or simply as the integrated transport map.

Define the set

$$A = \{a_1 \in [0,1] \mid \exists\ 0 \leq a_2 < a_2' \leq 1 - a_1 \text{ s.t. } \langle e_1, T((a_1, a_2))\rangle < \langle e_1, T((a_1, a_2'))\rangle\}$$

which corresponds to the set where the property in Lemma 2 fails. Now for $a_1 \in A$ define the set

$$\mathcal{T}_{a_1} = \{T' \in \mathcal{V}(C_0) \mid \exists\ 0 \leq a_2 < a_2' \leq 1 - a_1 \text{ s.t. } \langle e_1, T'((a_1, a_2))\rangle < \langle e_1, T'((a_1, a_2'))\rangle\}.$$

This corresponds to the set of maps where the property in Lemma 2 fails at the point $a_1$.

Observe that by Lemma 2 it must be the case that for $a_1 \neq a_1'$ that

$$\mathcal{T}_{a_1} \cap \mathcal{T}_{a_1'} = \varnothing$$

since each function $T' \in \mathcal{V}(C_0)$ fails to have the property in Lemma 2 at at most a single point.

Now let $a_1 \in A$ and let $0 \leq a_2 < a_2' \leq 1 - a_1$ be such that $\langle e_1, T((a_1, a_2))\rangle < \langle e_1, T((a_1, a_2'))\rangle$ which implies $\langle e_1, T((a_1, a_2))\rangle - \langle e_1, T((a_1, a_2'))\rangle < 0$. It holds that

$$\begin{aligned}
0 &> T((a_1, a_2))\rangle - \langle e_1, T((a_1, a_2'))\rangle \\
&= \int_{\mathcal{V}(C_0)} \langle e_1, T_i((a_1, a_2))\rangle - \langle e_1, T_i((a_1, a_2'))\rangle d\lambda(T_i) \\
&= \int_{\mathcal{T}_{a_1}} \langle e_1, T_i((a_1, a_2))\rangle - \langle e_1, T_i((a_1, a_2'))\rangle d\lambda(T_i) + \int_{\mathcal{V}(C_0)\setminus\mathcal{T}_{a_1}} \langle e_1, T_i((a_1, a_2))\rangle - \langle e_1, T_i((a_1, a_2'))\rangle d\lambda(T_i) \\
&\geq \int_{\mathcal{T}_{a_1}} \langle e_1, T_i((a_1, a_2))\rangle - \langle e_1, T_i((a_1, a_2'))\rangle d\lambda(T_i) \\
&\geq \int_{\mathcal{T}_{a_1}} -1 d\lambda(T_i) = -\lambda[\mathcal{T}_{a_1}]
\end{aligned}$$

which implies that $\lambda[\mathcal{T}_{a_1}] > 0$. In particular this holds for every $a_1 \in A$. Since the sets $\mathcal{T}_{a_1}$ are disjoint and for every $a_1 \in A$ the set $\mathcal{T}_{a_1}$ receives strictly positive mass from $\lambda$ it must be the case that $A$ is countable since for any probability measure any disjoint family of sets with positive probability is at most countable.

In particular since countable sets have measure 0 we can conclude that the set $A$ has measure 0 with respect to the Lebesgue measure on $[0,1]$. Therefore almost every $a_1 \in [0,1]$ is not contained in the set $A$ and therefore has the claimed property. □

***Proof of Theorem 5.*** Let $T \in \text{conv}(\mathcal{V}(C_0))$. By Lemma 4 we have for almost every $x$ the function $y \mapsto \langle e_1, T(x,y)\rangle$ is monotonically decreasing. Let $S$ denote the set of measure zero where this fails. Since $S$ has measure 0 we can ignore it while integrating. Now let $x \in [0,1) \setminus S$. Define the point

$$y_x = \sup\left\{y \in [0, 1-x] \mid \langle e_1, \nabla\phi_0(x,y)\rangle \leq \langle e_1, T(x,y)\rangle\right\}.$$

with the convention that $y_x = 0$ if the supremum is over an empty set. Along the segment $\{(x, y) \mid y \in [0, 1-x]\}$ we have

$$\begin{aligned}
&\int_0^{1-x} \|T(x,y) - \nabla\phi_0(x,y)\|_2 \, dy \\
\geq &\int_0^{1-x} |\langle e_1, T(x,y) - \nabla\phi_0(x,y)\rangle| \, dy \\
= &\int_0^{y_x} |\langle e_1, T(x,y) - \nabla\phi_0(x,y)\rangle| \, dy \\
&+ \int_{y_x}^{1-x} |\langle e_1, T(x,y) - \nabla\phi_0(x,y)\rangle| \, dy \\
= &\int_0^{y_x} \langle e_1, T(x,y) - \nabla\phi_0(x,y)\rangle \, dy \\
&+ \int_{y_x}^{1-x} \langle e_1, \nabla\phi_0(x,y) - T(x,y)\rangle \, dy \\
\geq &\int_0^{y_x} \langle e_1, \nabla\phi_0(x,y_x) - \nabla\phi_0(x,y)\rangle \, dy + \int_{y_x}^{1-x} \langle e_1, \nabla\phi_0(x,y) - \nabla\phi_0(x,y_x)\rangle \, dy \\
\geq &\int_0^{y_x} \frac{x}{8}(y_x - y) dy + \int_{y_x}^{1-x} \frac{x}{8}(y - y_x) dy \\
= &\frac{x}{8}\left[\left(y_x^2 - \frac{y_x^2}{2}\right) + \left(\frac{(1-x)^2}{2} - \frac{y_x^2}{2} - y_x(1-x) + y_x^2\right)\right] \\
= &\frac{x}{8}\left[\frac{(1-x)^2}{2} + y_x^2 - y_x(1-x)\right] \geq \frac{x}{8} \cdot \frac{(1-x)^2}{4}.
\end{aligned}$$

The first inequality is a consequence of $\|z\|_2^2 = \langle e_1, z\rangle^2 + \langle e_2, z\rangle^2 \geq \langle e_1, z\rangle^2$. For the second inequality we have that since $x \in S$ the function $T(x, \cdot)$ is monotonically decreasing and by continuity of $\nabla\phi_0(x, \cdot)$ (at least on $C_0$) that for all $0 \leq y < y_x$

$$\begin{aligned}
\langle e_1, T(x,y)\rangle &\geq \lim_{y_0^+ \to y_x} \langle e_1, T(x,y_0)\rangle \\
&\geq \lim_{y_0^+ \to y_x} \langle e_1, \nabla\phi_0(x,y_0)\rangle = \langle e_1, \nabla\phi_0(x,y_x)\rangle
\end{aligned}$$

which implies

$$\int_0^{y_x} \langle e_1, T(x,y) - \nabla\phi_0(x,y)\rangle \, dy \geq \int_0^{y_x} \langle e_1, \nabla\phi_0(x,y_x) - \nabla\phi_0(x,y)\rangle \, dy.$$

Similarly for the other integral we have for $y_x < y \leq 1 - x$

$$\begin{aligned}
\langle e_1, T(x,y)\rangle &\leq \lim_{y_0^- \to y_x} \langle e_1, T(x,y_0)\rangle \\
&\leq \lim_{y_0^- \to y_x} \langle e_1, \nabla\phi_0(x,y_0)\rangle = \langle e_1, \nabla\phi_0(x,y_x)\rangle
\end{aligned}$$

which implies

$$\int_{y_x}^{1-x} \langle e_1, \nabla\phi_0(x,y) - T(x,y)\rangle \, dy \geq \int_{y_x}^{1-x} \langle e_1, \nabla\phi_0(x,y) - \nabla\phi_0(x,y_x)\rangle \, dy.$$

The third inequality uses Lemma 3. The fourth uses that $y_x \in [0, 1-x]$ and minimizes the bound.

Using this inequality and the fact that $S$ is negligible we can compute

$$\begin{aligned}
&\int \|T - \nabla\phi_0\|_2 \, dU(C_0) \\
= &2\int_0^1 \int_0^{1-x} \|T - \nabla\phi_0(x,y)\|_2 \, dy dx \\
\geq &2\int_0^1 \frac{x(1-x)^2}{32} dx = \frac{1}{192}.
\end{aligned}$$

**Algorithm 5** MLE

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

This bound holds uniformly over $\mathcal{V}(C_0)$ and disproves Question 2. As noted above Question 1 is equivalent to Question 2. □

## E DETAILS FOR VERSIONS OF ALGORITHM 3

We consider three versions of Algorithm 3.

1. (**BCM**) `EstimateCoordinate` is done using the $W_2$BCM with $\eta = N(0, \hat{\Sigma}_{\text{emp}})$ and $\mu_i = \mathcal{N}(0, \Sigma_i)$. Next construct $A_{ij}$ as described in Section 4 and select $\hat{\lambda}_{\text{bcm}}$ as a minimizer in (8). `EstimateCovariance` is done using Algorithm 1 in (Chewi et al., 2020) with coordinate $\hat{\lambda}_{\text{bcm}}$ and $\Sigma_1, ..., \Sigma_m$.

2. (**LBCM**) Fix a $\mu_0 = \mathcal{N}(0, \Sigma_0)$, set $\eta = \mathcal{N}(0, \hat{\Sigma}_{\text{emp}})$, and set $\mu_i = \mathcal{N}(0, \Sigma_i)$. Compute the transport maps $T_{\mu_0}^{\mu_i}$ (which are of the form $T_{\mu_0}^{\mu_i}(x) = C_i(x)$) and $T_{\mu_0}^{\eta}$ using

$$C_i = S_\eta^{-1/2}\left(S_\eta^{1/2} S_i S_\eta^{1/2}\right)^{1/2} S_\eta^{-1/2}. \tag{26}$$

   Select $\hat{\lambda}_{\text{lbcm}}$ as a minimizer in (7). `EstimtateCovariance` is done using the estimate

$$\left(\sum_{i=1}^m (\hat{\lambda}_{lbcm})_i C_i\right)^T \Sigma_0 \left(\sum_{i=1}^m (\hat{\lambda}_{lbcm})_i C_i\right)$$

   which corresponds to the covariance of $\left(\sum_{i=1}^m (\hat{\lambda}_{lbcm})_i T_{\mu_0}^{\mu_i}\right) \#\mu_0$.

3. (**Maximum Likelihood Estimation (MLE)**) `EstimateCoordinate` is done by setting $\eta = N(0, \hat{\Sigma}_{emp})$ and returning $\hat{\lambda}_{\text{mle}}$ as the minimizer of $\min_{\lambda \in \Delta^m} D_{KL}(\eta \mid\mid \nu_\lambda)$ which can be numerically approximated. `EstimateCovariance` is done using using Algorithm 1 in (Chewi et al., 2020) with coordinate $\hat{\lambda}_{\text{mle}}$ and $\Sigma_1, ..., \Sigma_m$. We defer details of how the MLE strategy is implemented to Appendix F.

## F IMPLEMENTATION DETAILS FOR MLE

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 PyTorch. The procedure is summarized in Algorithm 5.

Here `SimplexProject`$(x)$ is defined as

$$\texttt{SimplexProject}(x) = \operatorname*{arg\,min}_{y \in \Delta^m} \|x - y\|_2 .$$

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 6. 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 6, and the gradient is obtained by back propagation through it.

**Algorithm 6** ComputeLoss

**Input:** $\{S_i\}_{i=0}^m$, $\lambda$, FPIters $> 0$, SQIters $> 0$
BC $\leftarrow S_0$
**for** $j = 1, ...,$ FPIters **do**
  BC_root $\leftarrow$ `SquareRoot`(BC, SQIters)
  BC_root_inv $\leftarrow$ `Invert`(BC_root)
  BC $\leftarrow$ BC_root_inv ($\sum_{i=1}^m \lambda_i$`SquareRoot`(BC_root, SQIters,$S_i$BC_root)) BC_root_inv
**end for**
**Return** $\mathrm{Tr}(\mathrm{BC}^{-1}S_0)$ + log det BC

The parameters that we chose in our experiments are given in Table 2. 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.

# G IMPLEMENTATION DETAILS FOR 2-WASSERSTEIN BARYCENTER SYNTHESIS

**Algorithm 7** Iterative 2-Wasserstein Barycenter Synthesis

**Input:** Initial probability measure $\rho_0$, step-size $\alpha > 0$, iterations $k \geq 1$.
**for** $1 \leq \ell \leq k$ **do**
  **for** $1 \leq j \leq m$ **do**
    Compute $\pi_{\rho_{\ell-1}}^{\mu_j} = \arg\min_{\pi \in \Pi(\rho_{\ell-1}, \mu_j)} \int \frac{1}{2}\|x-y\|^2 d\pi(x,y)$.
    Define map $\bar{T}_{\rho_{\ell-1}}^{\mu_j}(x) \triangleq \mathbb{E}_{(X,Y)\sim\pi_{\rho_{\ell-1}}^{\mu_j}}[Y|X=x]$
  **end for**
  $\rho_\ell = [(1-\alpha)Id + \alpha\sum_{j=1}^m \lambda_j \bar{T}_{\rho_{\ell-1}}^{\mu_j}]_{\#}[\rho_{\ell-1}]$
**end for**
**Return** $\rho_k$

We synthesize an approximate 2-Wasserstein barycenter using Algorithm 7. When an optimal transport map exists between iterates $\rho_\ell$ and references $\mu_j$, Algorithm 7 agrees with the algorithm implemented in (Zemel and Panaretos, 2019), where they establish convergence under regularity assumptions on the target measures. Algorithm 7 can also be viewed as a forward Euler discretization of the Wasserstein gradient flow for the functional $\sum_{j=1}^m \lambda_j W_2^2(-, \mu_j)$.

# H IMAGE ALGORITHMS

**Algorithm 8** Image to Measure

**Input:** Matrix representation of an image $\mathtt{I} \in \mathbb{R}_+^{d\times d}$
Set $s = \sum_{i,j=1}^d \mathtt{I}_{i,j}$
**Return** $\frac{1}{s}\sum_{i,j=1}^d \mathtt{I}_{i,j}\delta_{(i/d, j/d)}$

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

Table 2: Parameters used when performing the maximum likelihood estimation.

**Algorithm 9** Measure to Image

1: **Input:** Support locations $S \in \mathbb{R}^{n\times 2}$, Masses $\{b_k\} \in \mathbb{R}^n_+$, Output size $d$, Resolution $r$, Bandwidth $b$, Lower bound $\ell$.
2: Create $\mathtt{K} \in \mathbb{R}^{rd\times rd}_+$ with $\mathtt{K}_{ij} = \sum_{k=1}^n b_k \exp\left(-\frac{\|(i/rd, j/rd) - S_k\|_2^2}{b^2}\right)$
3: Set $\mathtt{K} = \left(\sum_{i,j=1}^{rd} \mathtt{K}_{ij}\right)^{-1} \cdot \mathtt{K}$
4: For each $i, j$ Set $\mathtt{K}_{ij} = \mathtt{K}_{ij} \cdot \mathbf{1}[\mathtt{K}_{ij} > \ell]$
5: Create $\mathtt{I} \in \mathbb{R}^{d\times d}$ with $\mathtt{I}_{ij} = \sum_{k,l=1}^{r} \mathtt{K}_{(i-1)*r+k,(l-1)*r+l}$
6: **Return:** $\left(\sum_{i,j=1}^{d} \mathtt{I}_{ij}\right)^{-1} \cdot \mathtt{I}$

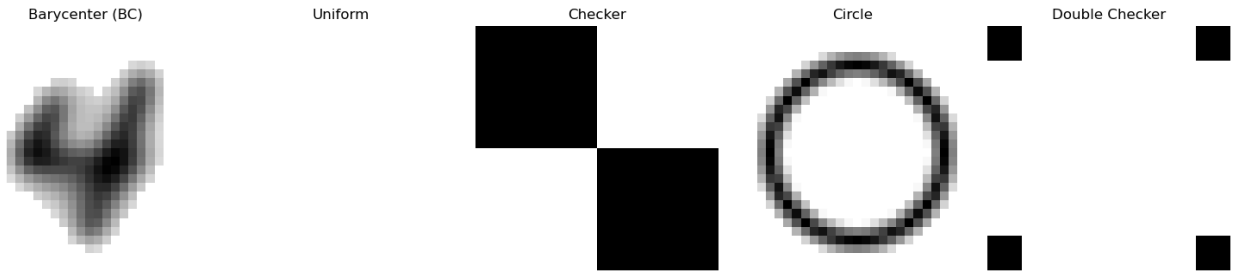


Figure 5: Base measures used in Figure 6

# I ADDITIONAL EXPERIMENTS

To conclude we briefly consider the impact that is had by the choice of the base measure in the LBCM in the problem above. We consider four choices for this measure which are presented in Figure 5. The first measure is the convolutional barycenter of the 10 reference digits using the method of (Solomon et al., 2015). The second is a uniform image, the third uses two squares in the corners, the fourth is supported on a circle, and the final is supported on squares supported in the four corners of the image. The reconstruction of the same digits using each of these as the base measure is shown in Figure 6.

In general the barycenter digit tends to give visually the best reconstruction while the uniform and checkered

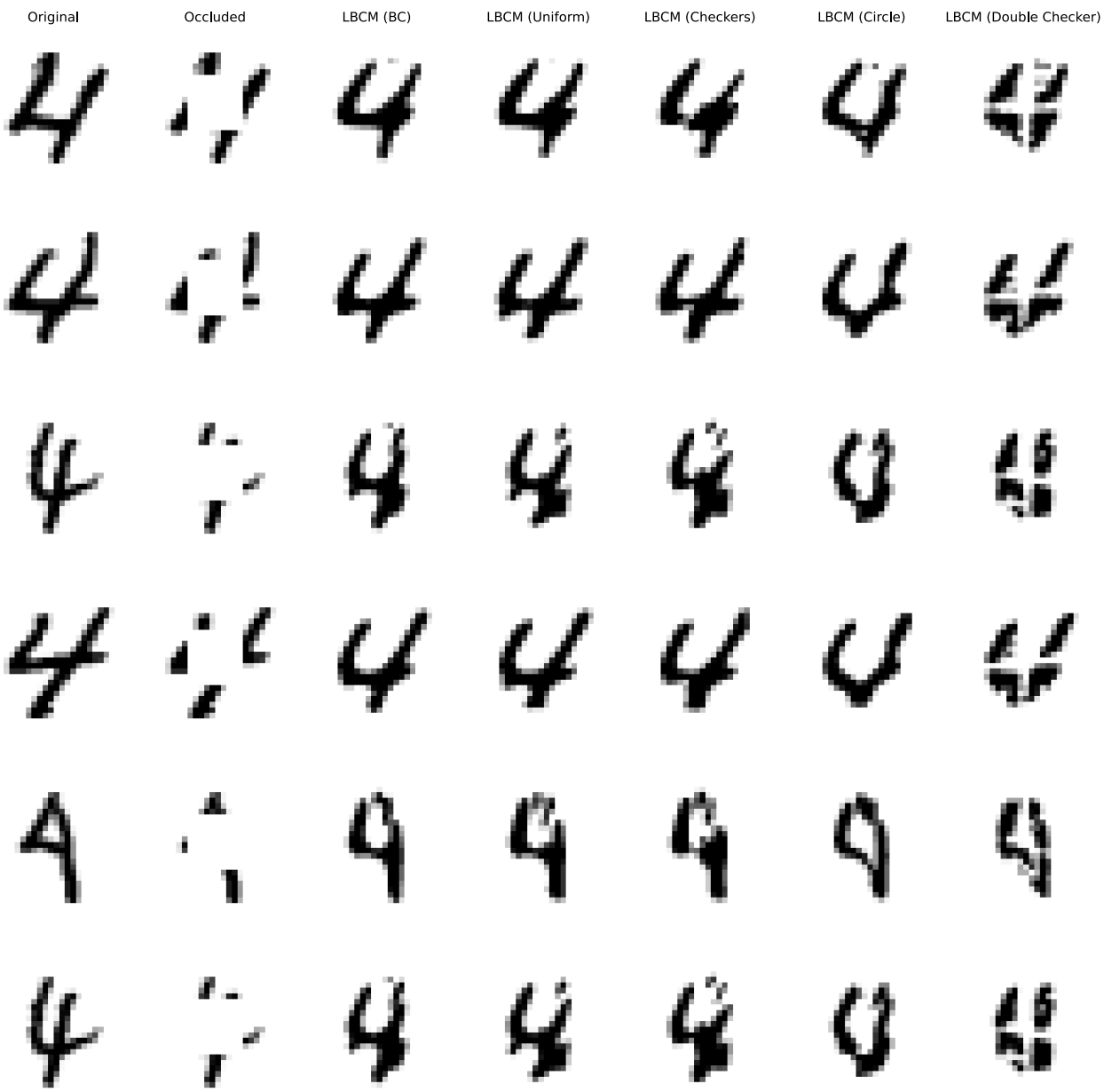


Figure 6: Reconstruction of occluded digits using the LBCM with four different base measures

base measures perform comparably to each other and the circle yields the worst reconstructions. This highlights the fact that the choice of the base measure is meaningful and can significantly impact the recovery. The question of which base measures lead to the best performance is left to future work.