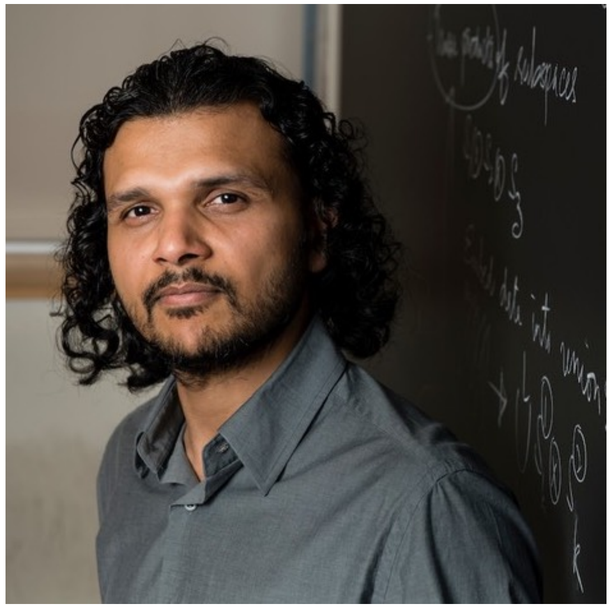
Measure Estimation in the Barycentric Coding Model: Geometry, Statistics, and Algorithms

James M. Murphy
Department of Mathematics

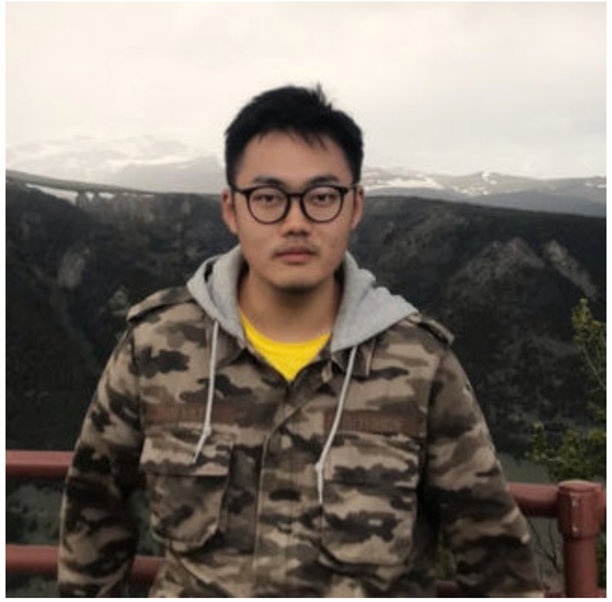
October 1, 2022

Tufts
UNIVERSITY

Collaborators at Tufts



Shuchin Aeron, ECE



Ruijie Jiang, ECE

Abiy Tasissa, Math

Matt Werenski, CS

Tufts UNIVERSITY

Learning in High Dimensions is Hard

- High-dimensional problems (e.g. many variables relative to number of observations) are hard for machine learning.

- The *curse of dimensionality* dooms inference in the absence of structural assumptions on the data:

  $$\textit{If } \{x_i\}_{i=1}^n \textit{ is a uniform, i.i.d. sample from } [0,1]^D, \textit{ then } x_i \textit{ is distance} \approx n^{-\frac{1}{D}} \textit{ from its nearest neighbor.}$$

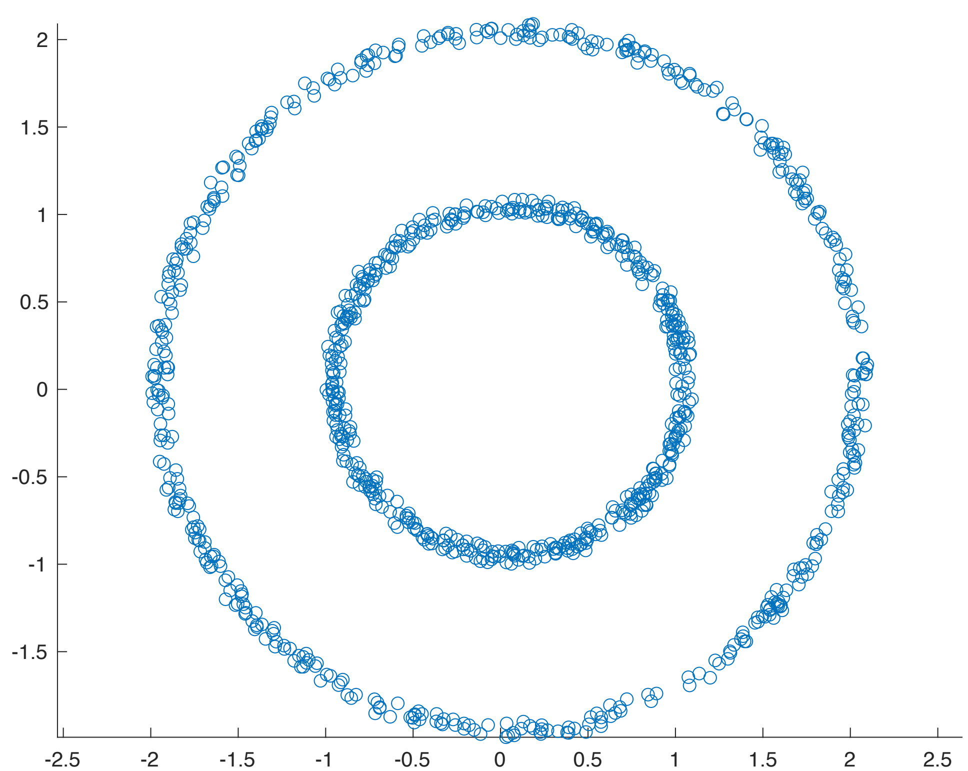
- Pairwise distances may not be informative—nothing is close to anything else.

Classical Approach to Breaking Curse

- Popular model: data are near a low-dimensional subspace or manifold.

- That is, the data actually live near $\mathcal{M}^d \subset \mathbb{R}^D$ where one aims to develop methods that depend exponentially on d.

- When $d \ll D$, one may hope to break the curse.



Ambient dimension is 2, but data is (approximately, locally) 1 dimensional.

Tufts UNIVERSITY

“Think Globally, Fit Locally”

- How to get methods that depend on manifold dimension rather than ambient dimension?

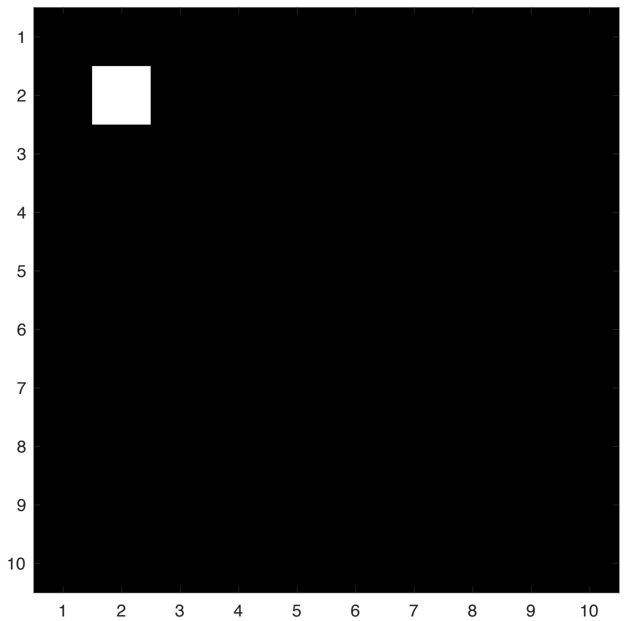
- **Main Idea of Manifold Learning:** local Euclidean information (e.g., nearest neighbor calculations) can be leveraged to make global inferences.

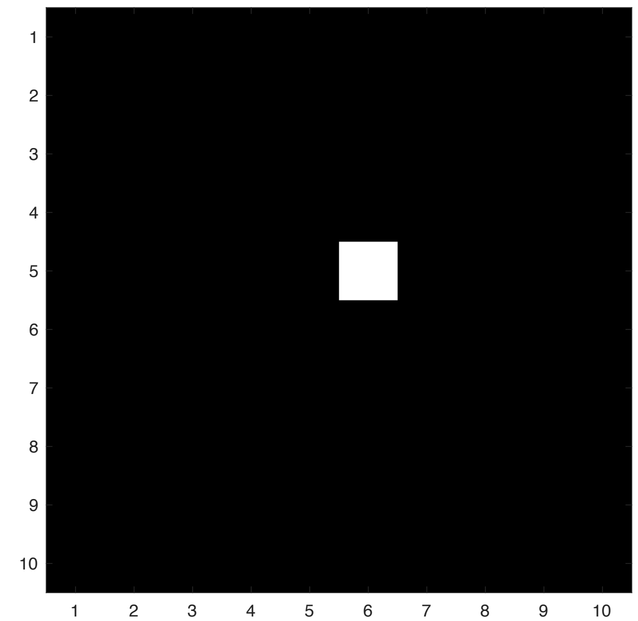
- Data is locally low dimensional, so “zoom in” finely enough for this to be the limiting factor.

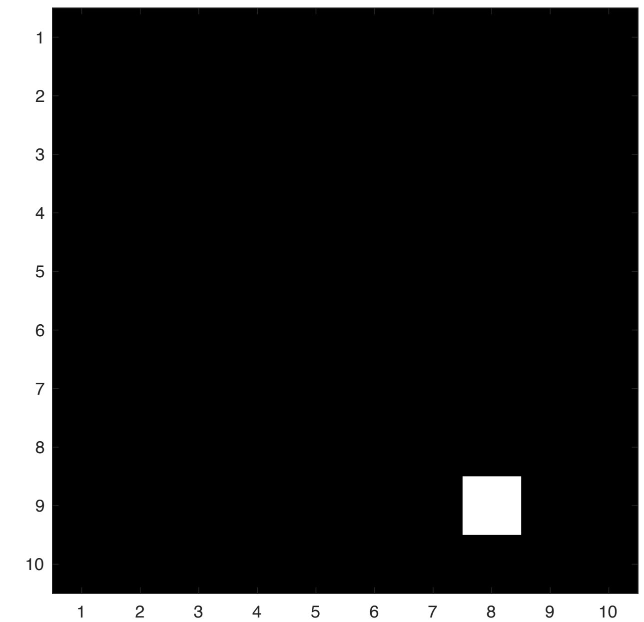
- This is typical done with Euclidean distances and a graph is constructed, from which global information can be gleaned: geodesics, PDE/diffusions on graphs, structure-preserving embeddings,…

Beyond Euclidean Distances

- Methods based on local Euclidean distances may be insufficient to capture the geometry of certain data.
- **Toy example**: black and white images with single white pixel:







- Everything is equally far in Euclidean distance, and therefore in any graph metric.
- Need to capture the distance *between the support of these images*.

Tufts UNIVERSITY

Data as Measures: Wasserstein-2 Metric

- Let $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}})$ denote the space of absolutely continuous measures (i.e., having density with respect to the Lebesgue measure) with finite second moment.

- For two measures $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}})$, the Wasserstein-2 metric is

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

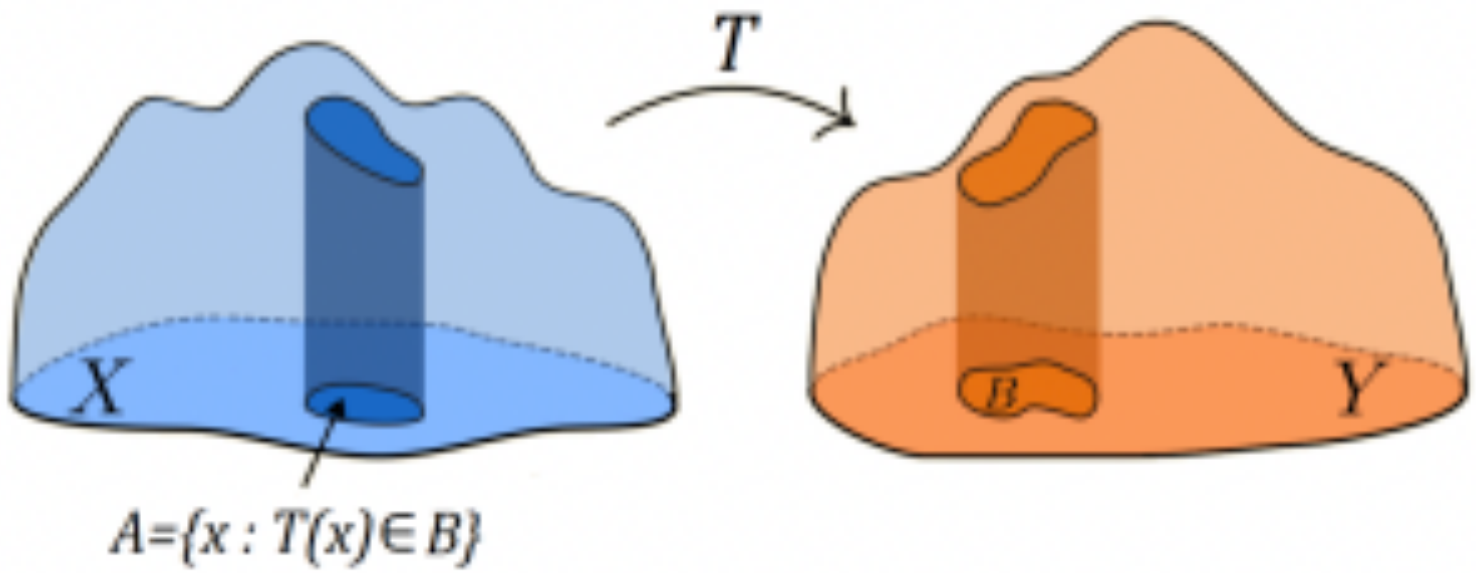
where the minimization is over all maps $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushforward μ onto ν:

$$T\#\mu = \nu \;\leftrightarrow\; \nu[B] = \mu[T^{-1}(B)] \quad \text{for all Borel sets } B\,.$$

Tufts UNIVERSITY

Optimal Transport Maps

- Pushforwards transfer mass from one distribution to another.



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

  is the optimal transport map. It pushes forward in the "most efficient" way.

Image Source: M. Thorpe (https://www.math.cmu.edu/~mthorpe/OTNotes)

Tufts UNIVERSITY

Existence and Generalization

- This is not well-defined for general measures, and this *Monge* formulation is less tractable than the *Kantorovich* formulation:

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

$$\Pi(\mu,\nu) = \left\{ \gamma : \mathbb{R}^{2d} \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}^d} \gamma(x,y)dx = \nu(y),\; \int_{\mathbb{R}^d} \gamma(x,y)dy = \mu(x) \right\}.$$

- Under our assumptions, these formulations are equivalent and a unique T exists. We'll return to the Kantorovich form when computing.

Averaging in $\mathcal{W}_2$: Barycenters

- Let $\Delta^p = \left\{ \lambda = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p : \lambda_i \geq 0, \sum_{i=1}^{p} \lambda_i = 1 \right\}.$

- For measures $\{\mu_i\}_{i=1}^p \subset \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}})$ and coordinates $\lambda \in \Delta^p$, define the *Wasserstein-2 barycenter* as

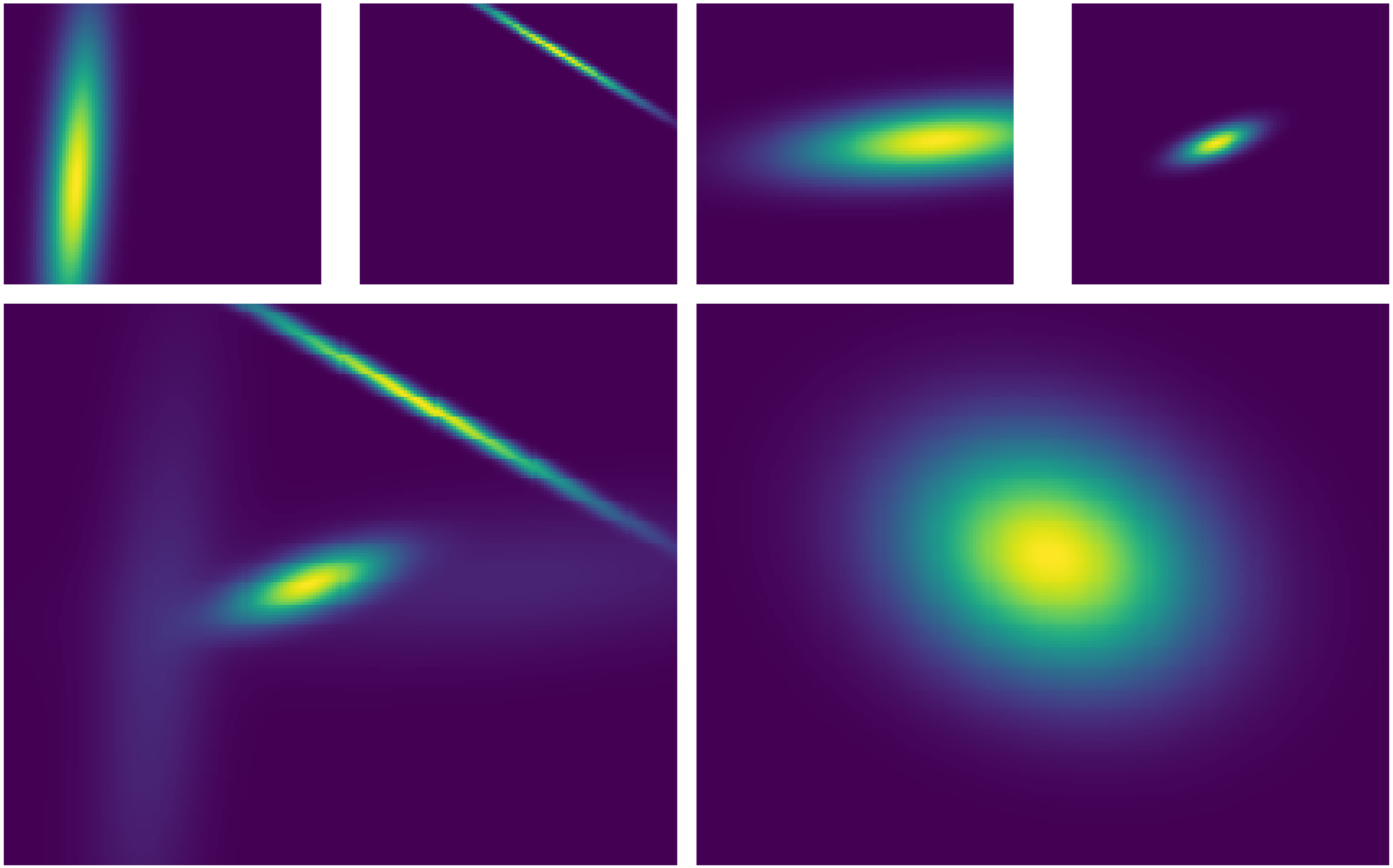
$$\nu_\lambda = \underset{\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}})}{\arg\min} \frac{1}{2} \sum_{i=1}^{p} \lambda_i W_2^2(\nu, \mu_i)$$

- This is well-defined and unique under our assumptions.

- ν_λ is the "right" way of averaging in the space of measures.

Tufts UNIVERSITY

Barycenters Preserve Structure



Euclidean Mixture

Wasserstein Barycenter

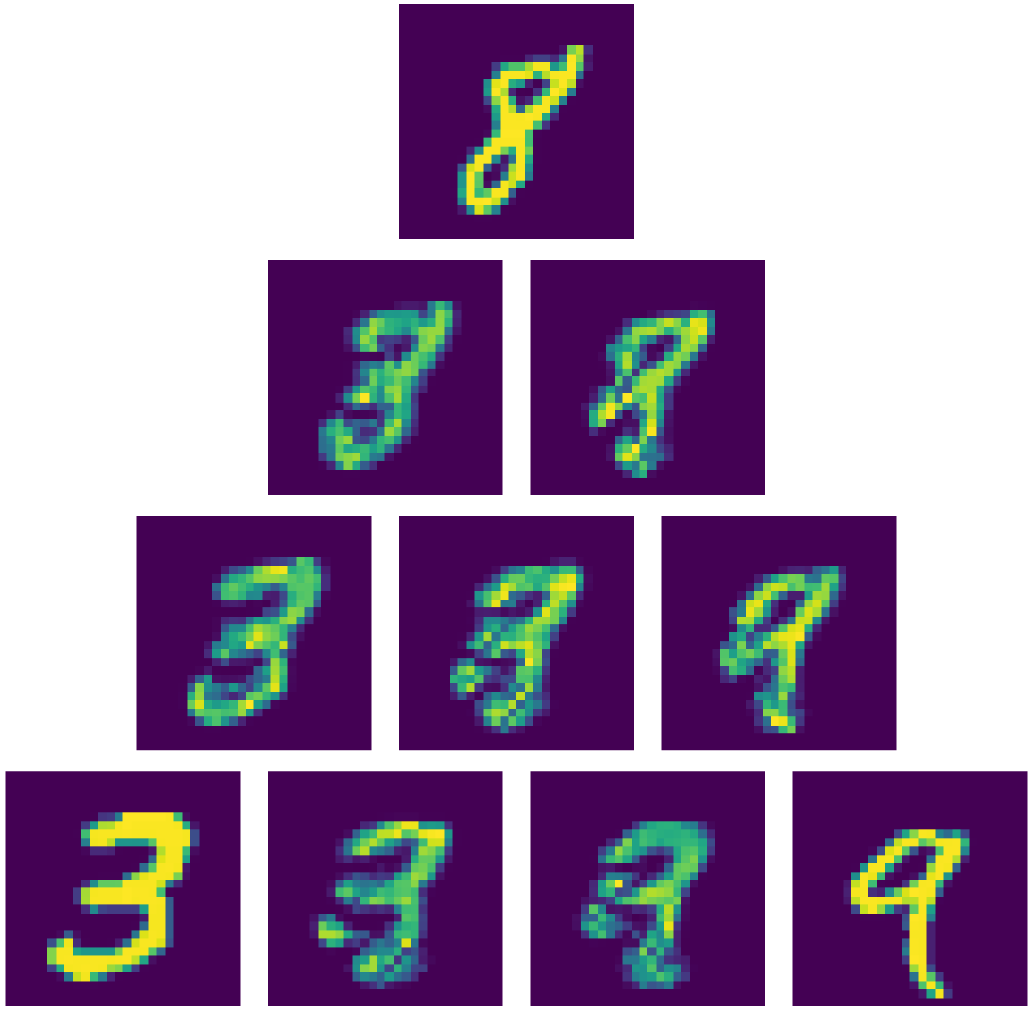
$$\sum_{i=1}^{p} \lambda_i \mu_i$$

$$\underset{\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}})}{\arg\min} \frac{1}{2} \sum_{i=1}^{p} \lambda_i W_2^2(\nu, \mu_i)$$

The Synthesis Problem

- The *synthesis problem* is solving

$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}})} \frac{1}{2} \sum_{i=1}^{p} \lambda_i W_2^2(\nu, \mu_i).$$



- Existence and uniqueness theory, consistent estimation procedures, and fast numerical schemes have been developed in the past decade (McCann; Agueh and Carlier; Alvarez-Esteban et al.; Bigot and Klein; Claici, Chien, and Solomon; Schmitz et al.; Yang et al. …)

The *Barycentric Coding Model*

- Let $\mathrm{Bary}(\{\mu_i\}_{i=1}^p) = \{\nu_\lambda : \lambda \in \Delta^p\}$ be the set of all barycenters that can be generated from $\{\mu_i\}_{i=1}^p$.

- We denote by the *barycentric coding model (BCM)* the identification of a measure

$$\mu_0 \in \mathrm{Bary}(\{\mu_i\}_{i=1}^p)$$

with its coordinates $\lambda \in \Delta^p$.

- $\mathrm{Bary}(\{\mu_i\}_{i=1}^p)$ can be thought of as the "span" of the reference measures, but with respect to the geometry of Wasserstein space.

Tufts UNIVERSITY

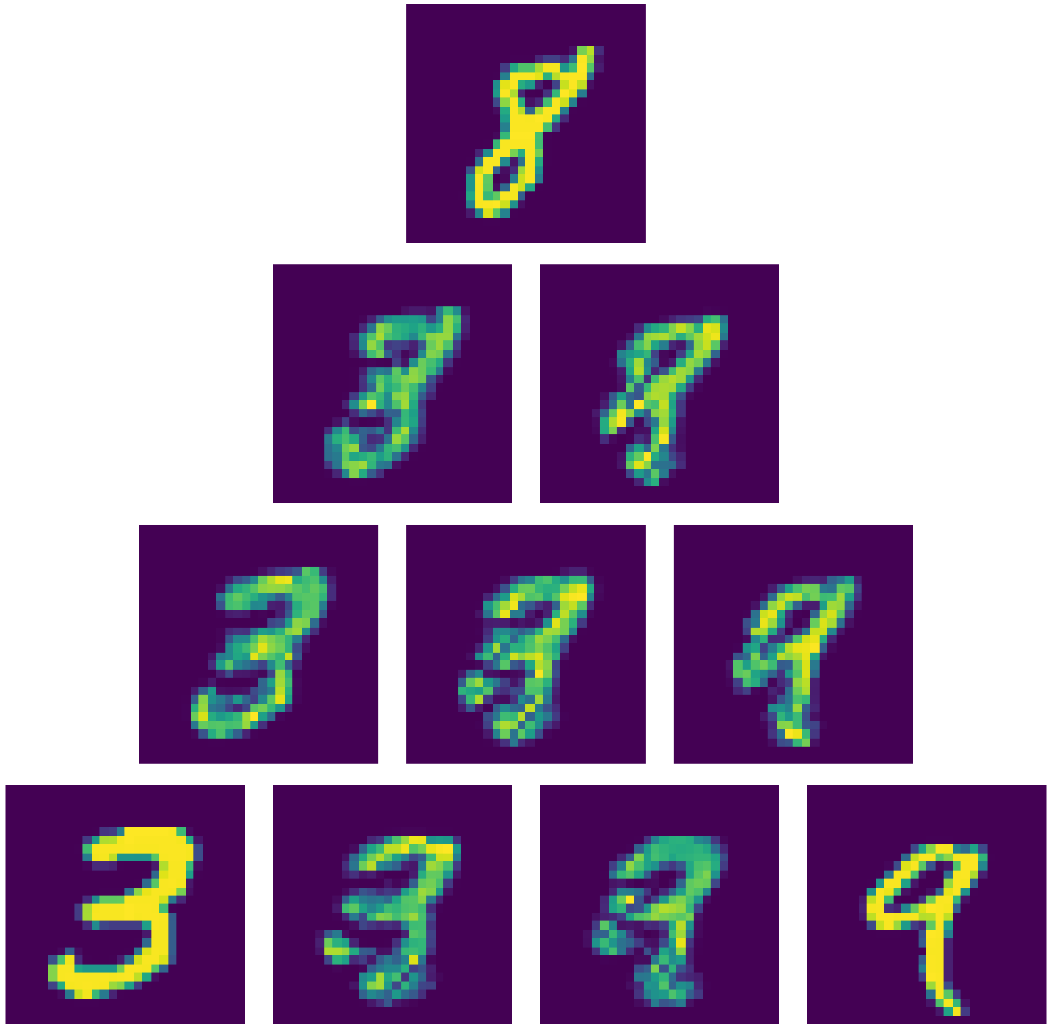
The Analysis Problem

- Given a measure μ_0 and reference measures $\{\mu_i\}_{i=1}^{p}$, the *analysis problem* is solving

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

- If $\mu_0 \in \mathrm{Bary}(\{\mu_i\}_{i=1}^{p})$, then:

$$\min_{\lambda \in \Delta^p} W_2^2\left(\mu_0, \nu_\lambda\right) = 0.$$



- Some computational methods known (Bonneel, Peyré, and Cuturi) but no existence and uniqueness results nor rigorous estimation procedures.

BCM as Low-Parameter Model

- The set $\text{Bary}(\{\mu_i\}_{i=1}^p) = \{\nu_\lambda : \lambda \in \Delta^p\}$ can be interpreted as a p-parameter subspace in the space of measures.

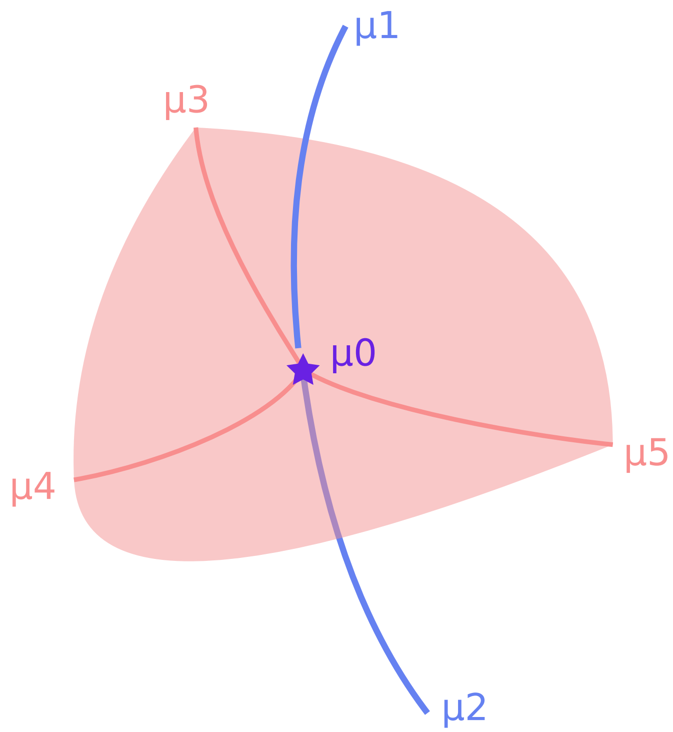
- This can be contrasted with a linear subspace, in terms of number of parameters needed to uniquely specify an element.

- Unlike linear subspaces, however, there is not a notion of orthogonal projection to quickly compute coordinates.

Tufts UNIVERSITY

Basic Questions

- Unique representations in $\mathrm{Bary}(\{\mu_i\}_{i=1}^p) = \{\nu_\lambda : \lambda \in \Delta^p\}$?



Not always, but perhaps generically.

- How to check if $\mu_0 \in \mathrm{Bary}(\{\mu_i\}_{i=1}^p)$?

- More generally, how to solve

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

Tufts UNIVERSITY

BCM as Quadratic Program

Theorem. *(Aeron, Jiang,* ***M.****, Tasissa, Werenski) Suppose* $\{\mu_i\}_{i=0}^p$ *are sufficiently regular. Then* $\mu_0 \in \mathrm{Bary}(\{\mu_i\}_{i=1}^p)$ *if and only if*

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

where $A \in \mathbb{R}^{p\times p}$ *is given by* $A_{ij} = \int_{\mathbb{R}^d} \langle T_i(x) - \mathrm{Id}(x), T_j(x) - \mathrm{Id}(x)\rangle d\mu_0(x)$ *for* T_i *the optimal transport map between* μ_0 *and* μ_i*. Furthermore, if the minimum value is 0 and* λ_* *is an optimal argument, then* $\mu_0 = \nu_{\lambda_*}$*.*

- $T_i(x) - \mathrm{Id}(x)$ is the displacement of the vector $x \in \mathbb{R}^d$ when transported by the map T_i which optimally transports μ_0 to μ_i .

- $\langle T_i(x) - \mathrm{Id}(x), T_j(x) - \mathrm{Id}(x)\rangle$ can be thought of as the angle between the displacement associated to the optimal transport map between μ_0 to μ_i with that of μ_0 to μ_j.

- Integrating with respect to μ_0 quantifies the average angle between displacements.

Proof Sketch and Interpretation

- The main idea is to understand the minimizers of the *variance functional* $G_\lambda : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}}) \to \mathbb{R}$ given by

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

- This has Fréchet derivative $\nabla G_\lambda(\nu) = -\sum_{i=1}^{p} \lambda_i \, (T_i - \mathrm{Id}).$

- Under regularity conditions on the optimal transport maps, solutions to the synthesis problem occur at the critical points of the Fréchet derivative.

- Then the result follows by computing

$$||\nabla G_\lambda(\mu_0)||_{\mu_0}^2 = \lambda^T A \lambda.$$

Tufts UNIVERSITY

Projection onto Barycentric Span?

- If $\mu_0 \notin \mathrm{Bary}(\{\mu_i\}_{i=1}^p)$, we can still find the minimizer of the quadratic form $\lambda \mapsto \lambda^T A \lambda$.

- A natural question then is, does $\lambda_* = \arg\min_{\lambda \in \Delta^p} \lambda^T A \lambda$ approximate well

$$\arg\min_{\lambda \in \Delta^p} W_2^2\left(\mu_0, \nu_\lambda\right)?$$

- In certain cases ($d = 1$; all measures are Gaussian, …), solving the quadratic program gives the exact projection!

Tufts UNIVERSITY

OT in Practice: Entropic Regularization

- Given i.i.d samples $\{X_i\}_{i=1}^n \sim \mu,\ \{Y_i\}_{i=1}^n \sim \nu$, the discrete (Kantorovich) W_2 problem solves

$$\arg\min_{\substack{\pi \in \mathbb{R}_+^{n\times n} \\ \pi \mathbf{1} = \mathbf{1} \\ \pi^T \mathbf{1} = \mathbf{1}}} \sum_{j=1}^n \sum_{k=1}^n \|X_j - Y_k\|_2^2 \cdot \pi_{jk}$$

- This has complexity in n at least $O(n^3)$—too slow.

- To improve complexity, one can consider *entropic regularization*: for $\epsilon > 0$, solve:

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

Entropic Estimation of BCM Coordinates

Algorithm 1 Estimate λ

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

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

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

Solve for g^i as the optimal g in

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

Define $\hat{T}_i(x) = \dfrac{\sum_{i=1}^n Y_i \exp\left(\frac{1}{\epsilon}(g^i(Y_i) - \frac{1}{2}||x - Y_i||_2^2)\right)}{\sum_{i=1}^n \exp\left(\frac{1}{\epsilon}(g^i(Y_i) - \frac{1}{2}||x - Y_i||_2^2)\right)}$.

end for

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

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

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

Tufts UNIVERSITY

Consistency of Entropic Estimation

Theorem. *(Aeron, Jiang, **M.**, Tasissa, Werenski) Let $i, j \in \{1, ..., p\}$ and suppose that μ_i, μ_j, μ_0 are supported on bounded domains and that the maps T_i and T_j are sufficiently regular. Let $X_1, ..., X_{2n} \sim \mu_0, Y_1, ..., Y_n \sim \mu_i, Z_1, ..., Z_n \sim \mu_j$. For an appropriately chosen ϵ, let $\hat{T}_i$ and $\hat{T}_j$ be the entropic maps computed using $\{X_i\}_{i=1}^n, \{Y_i\}_{i=1}^n, \{Z_i\}_{i=1}^n$. Then we have*

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

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

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

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

- Solution to the sample-driven, entropic problem converges to the true one.
- Rate depends on dimensionality and smoothness.

Tufts UNIVERSITY

Application: Covariance Estimation

- Consider sampling from a measure $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$ for known zero mean Gaussians $\{\mu_i\}_{i=1}^p$.

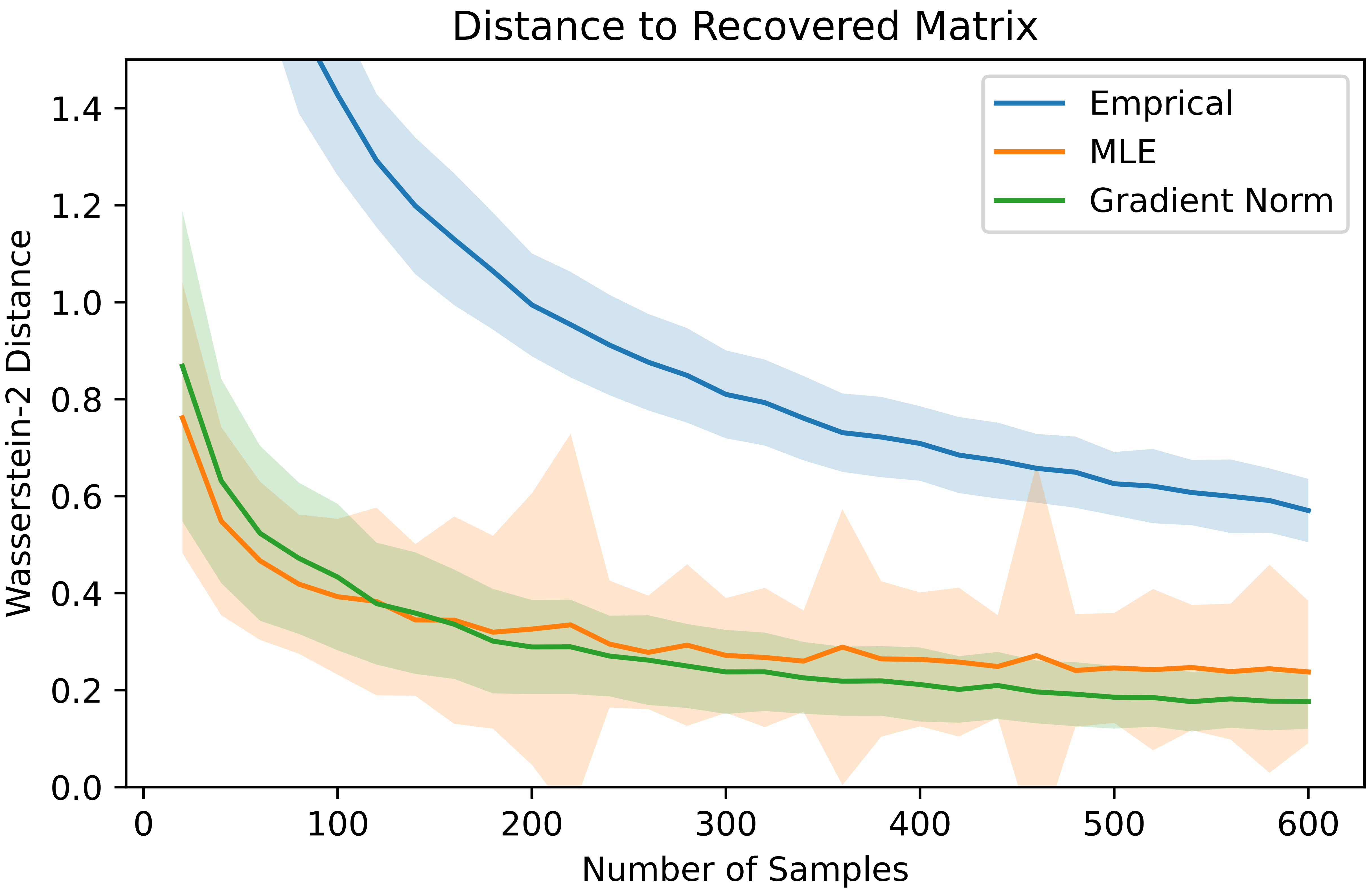
- Then μ_0 is necessarily Gaussian. In fact, its covariance matrix is structured.

Corollary. *For $i = 1, \ldots, p$, let $\mu_i = \mathcal{N}(0, S_i)$ with $S_i \in \mathbb{S}^d_{++}$. Then μ_0 is a barycenter if and only if $\mu_0 = \mathcal{N}(0, S_0)$ for some $S_0 \in \mathbb{S}^d_{++}$, and $\min_{\lambda \in \Delta^p} \lambda^T A \lambda = 0$, where the matrix A is given by $A_{ij} = \text{Tr}\left((C_i - I)(C_j - I)S_0\right)$ for $C_i = S_0^{-1/2}\left(S_0^{1/2} S_i S_0^{1/2}\right)^{1/2} S_0^{-1/2}$. Furthermore, if the minimum value is zero and λ_* is a optimal argument, then $\mu_0 = \nu_{\lambda_*}$*

- We can plug in the empirical covariance matrix for S_0, solve the QP above, and use the learned coefficients to estimate the covariance matrix of the measure observed only through samples.

Tufts UNIVERSITY

Numerical Results



Tufts UNIVERSITY

Supervised NLP: Documents as Distributions

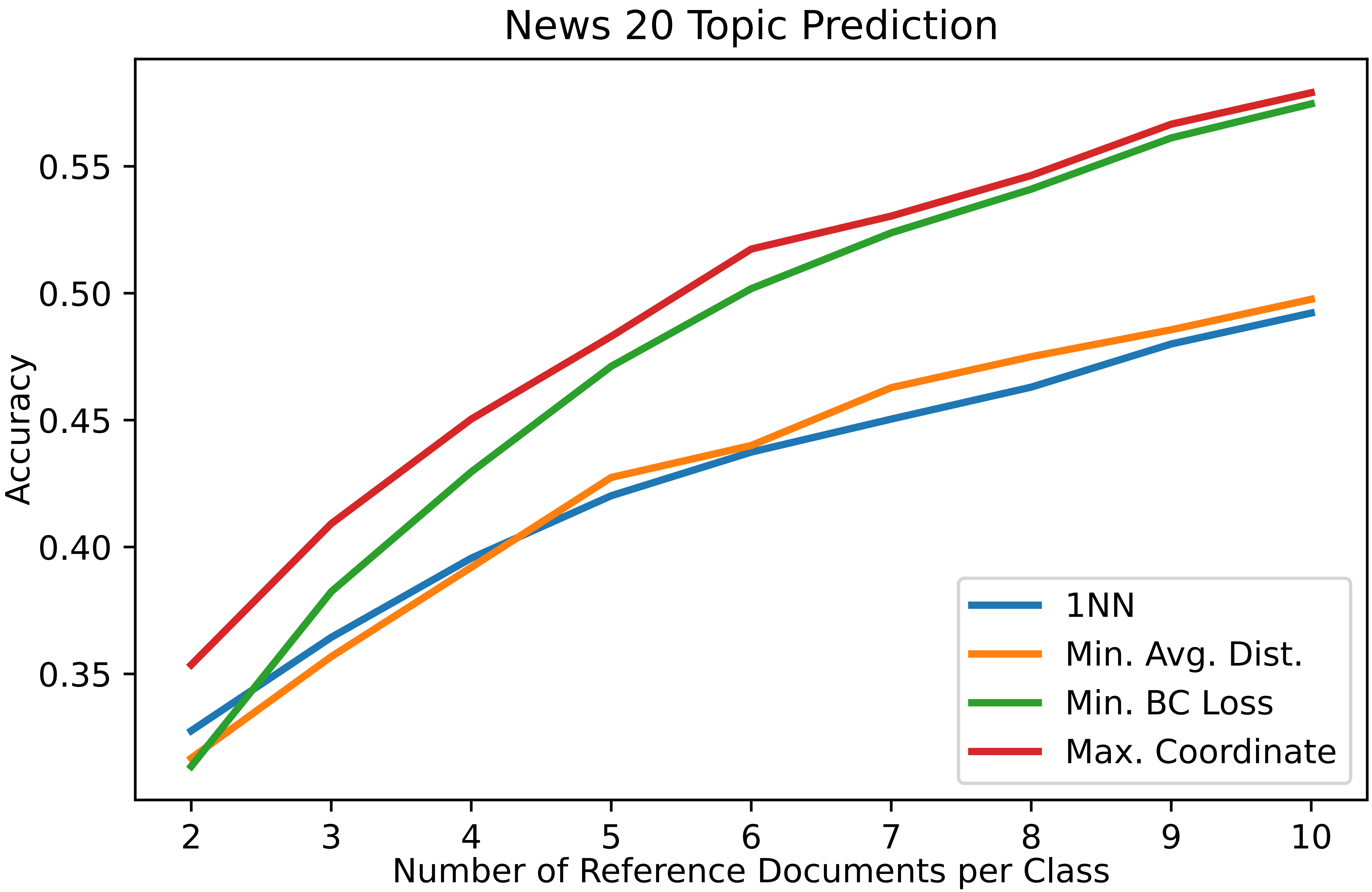
- We can consider a written document as a probability distribution in the space of words.

- Under this model, we can consider documents of different classes (e.g., science documents, sports documents,…) and use the BCM to decompose a new document using representatives of these classes.

- The corresponding coefficients can be used to determine a label for the new document.

- Note that standard word embeddings can be used to reduce the dimensionality and improve the learning rate of the BCM coordinates.

Classification with Few Labels



Tufts UNIVERSITY

Ongoing Research and Open Problems

- Representational capacity of $\{\mu_i\}_{i=1}^{p}$ as $p \to \infty$? Note that if the measures are Gaussian, the BCM is not a universal approximator!
- Regularized representation and dictionary learning.
- Can we efficiently estimate λ without estimating the OT maps? All we need are angles, not the maps themselves.
- Connections to linear OT framework.

Paper & Support

Werenski, Jiang, Tasissa, Aeron, **Murphy**
"Measure Estimation in the Barycentric Coding Model"
International Conference on Machine Learning
2022

The Camille & Henry Dreyfus Foundation

NSF

Tufts University

DISC

DMS 1912737
DMS 1924513
CCF 1934553

Tufts University

Code and Contact Information

Code: https://jmurphy.math.tufts.edu/Code/

Contact: jm.murphy@tufts.edu

Thanks for Your Attention!