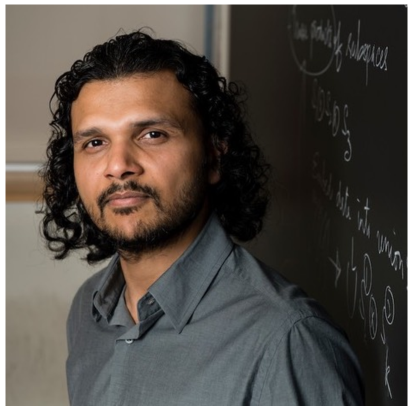
Intrinsic Models in Wasserstein Space with Applications to Molecular Dynamics

James M. Murphy

Brin MRC
February 22, 2024

Tufts UNIVERSITY

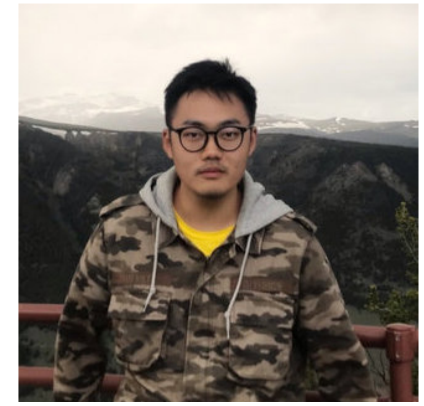
Collaborators



Aeron (Tufts ECE)

Ba (Harvard SEAS)

Damjanovic (Novo Nordisk)



Jiang (Tufts ECE)

Lin (Tufts Chemistry)

Masud (Tufts ECE)

Mueller (Tufts Math)

Tankala (Harvard SEAS)

Tasissa (Tufts Math)

Werenski (Tufts CS)

Learning in High Dimensions is Hard

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

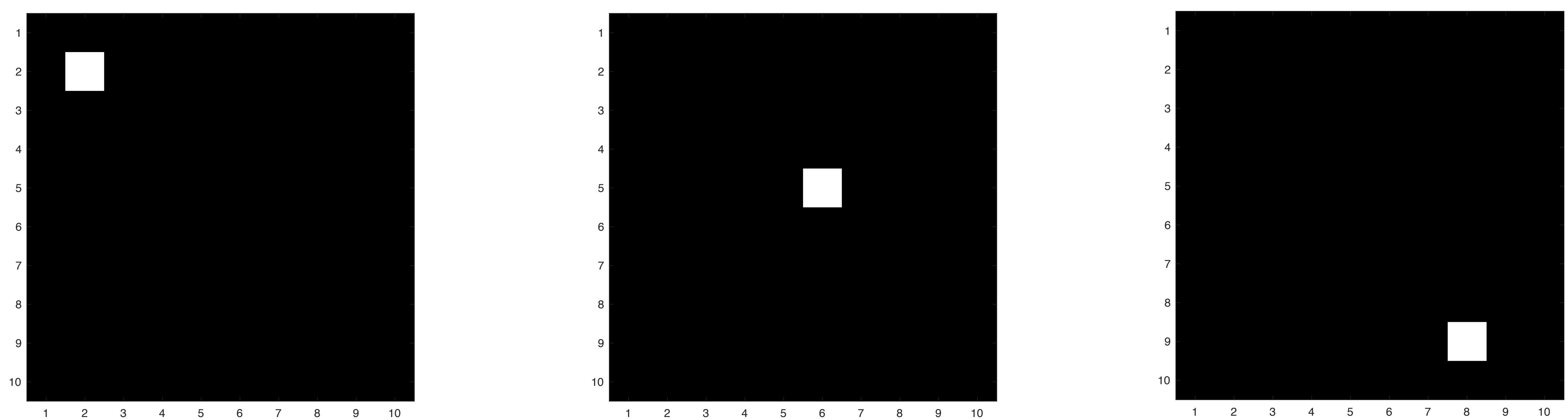
- The *curse of dimensionality* dooms inference in the absence of structural assumptions on the data.

- *Manifold Hypothesis*: data lies near low-dimensional subspace or manifold. Use local Euclidean distances to construct global distances (e.g., geodesics, Laplacian embeddings, diffusion distances,…)

Beyond the Euclidean

- Manifold learning 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.*

Data as Measures & 2-Wasserstein Space

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

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

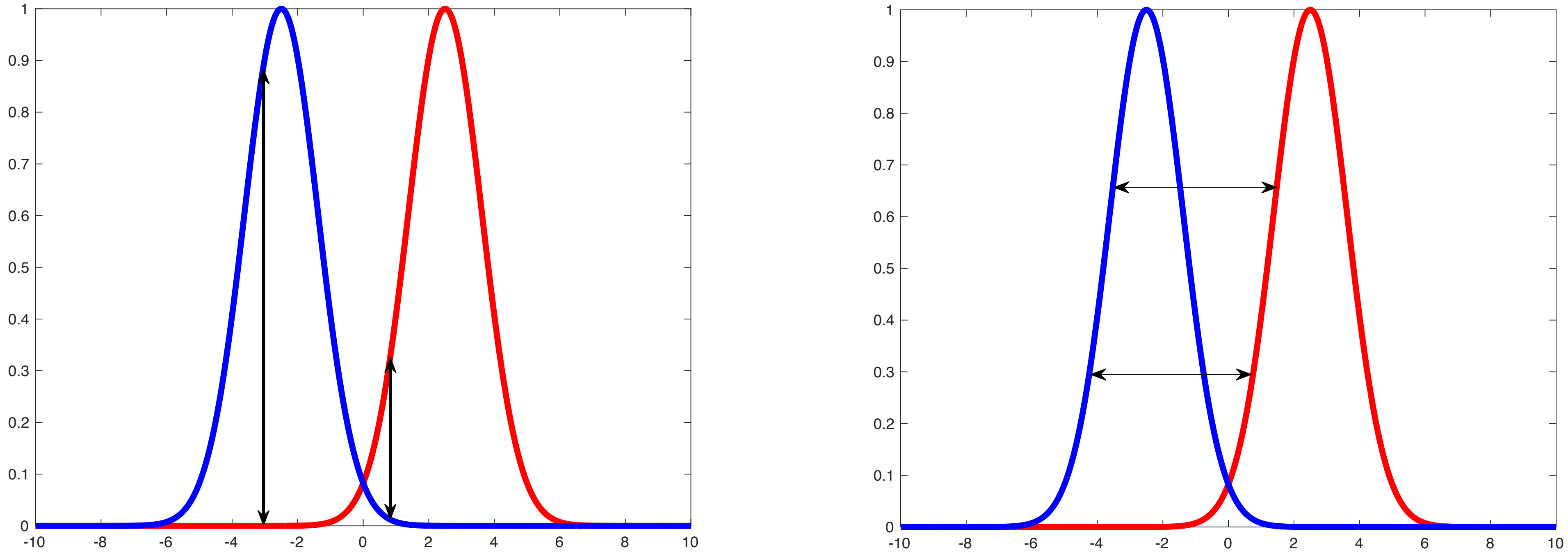
where the minimization is over all maps $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushforward μ onto ν:

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

Tufts UNIVERSITY

Optimal Transport Maps

- Pushforwards transfer mass from one distribution to another.



L^2 $\qquad$ W_2

- 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*.

Tufts UNIVERSITY

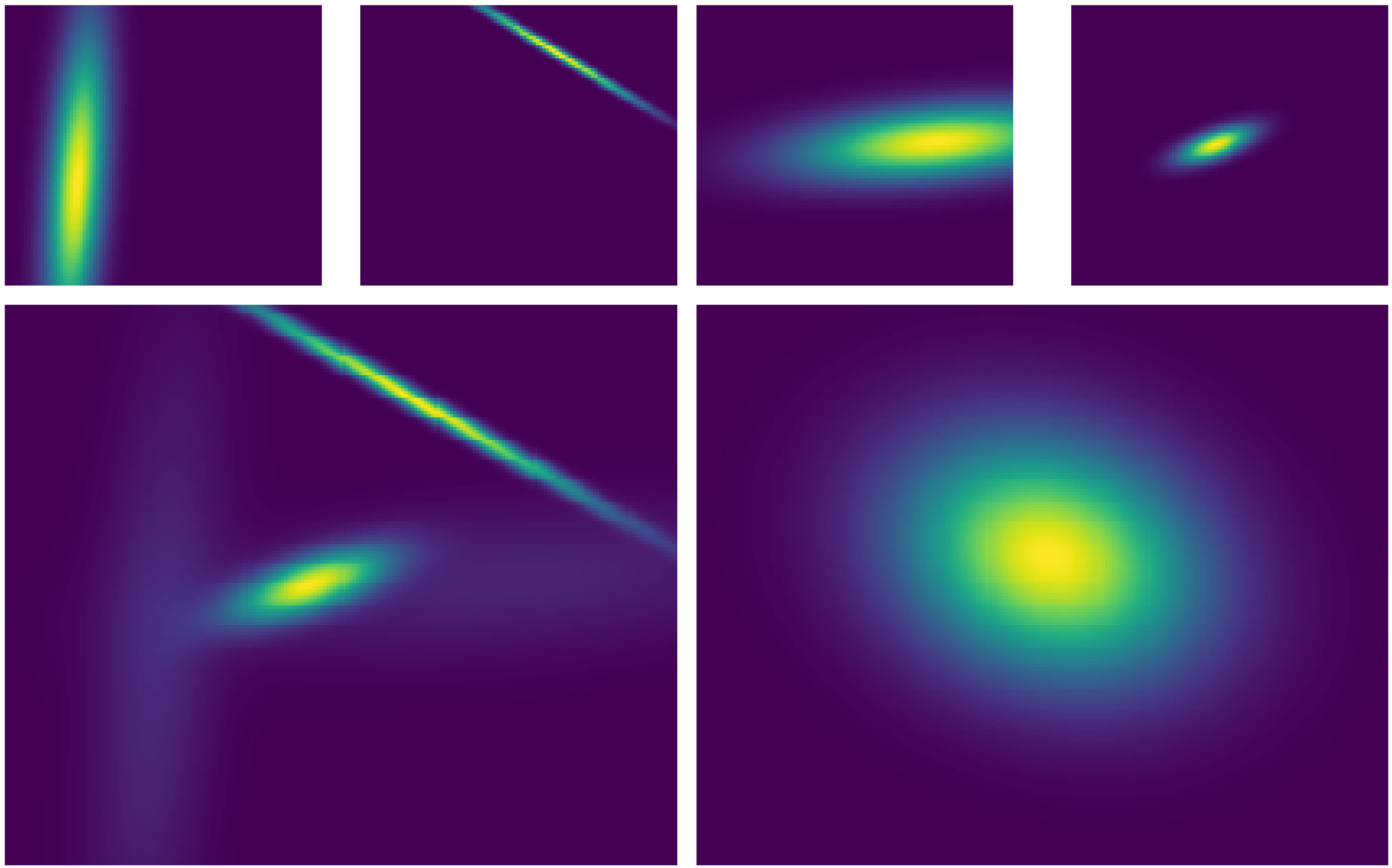
Averaging in $\mathcal{W}_2$: Barycenters

- 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

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

Barycenters Preserve Shape



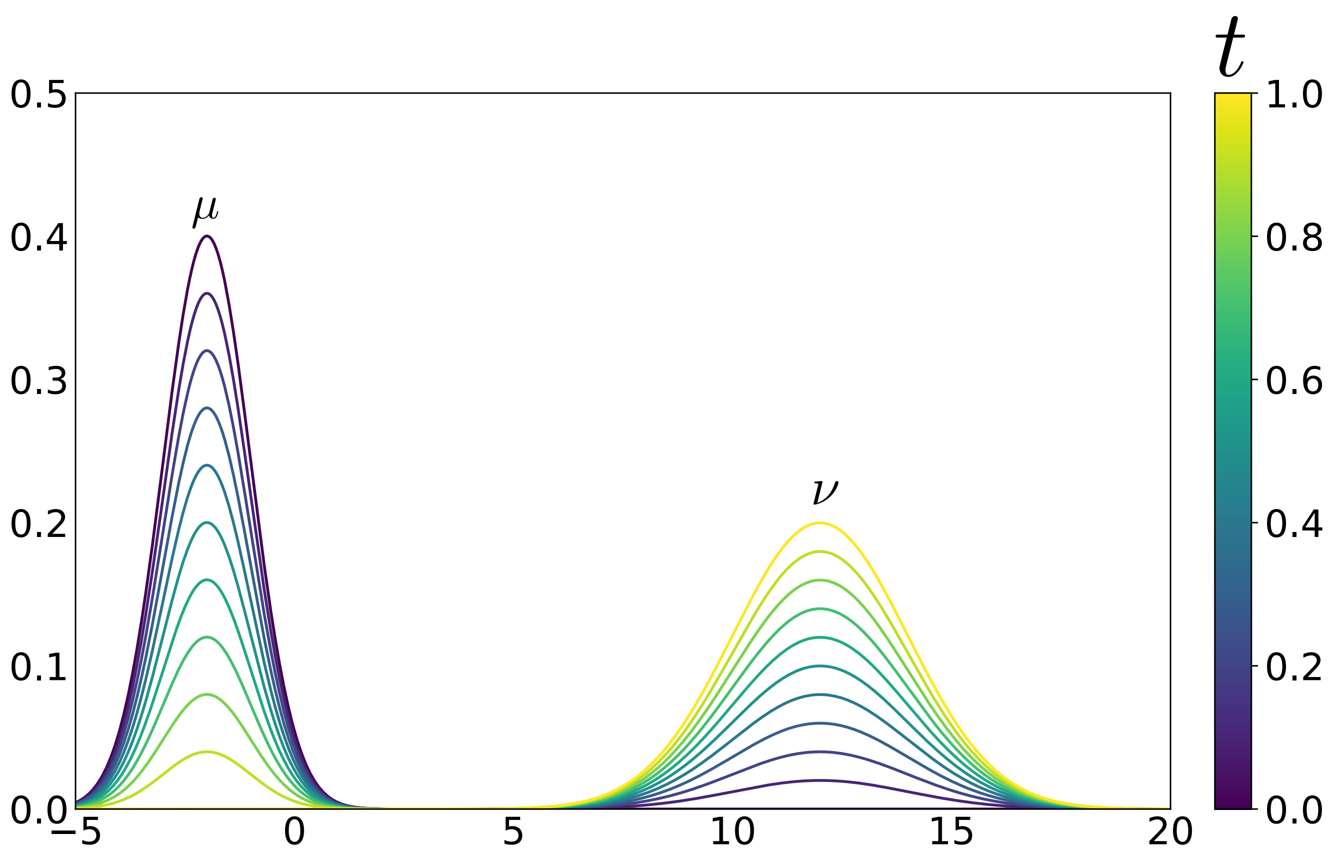
Euclidean Mixture

Wasserstein Barycenter

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

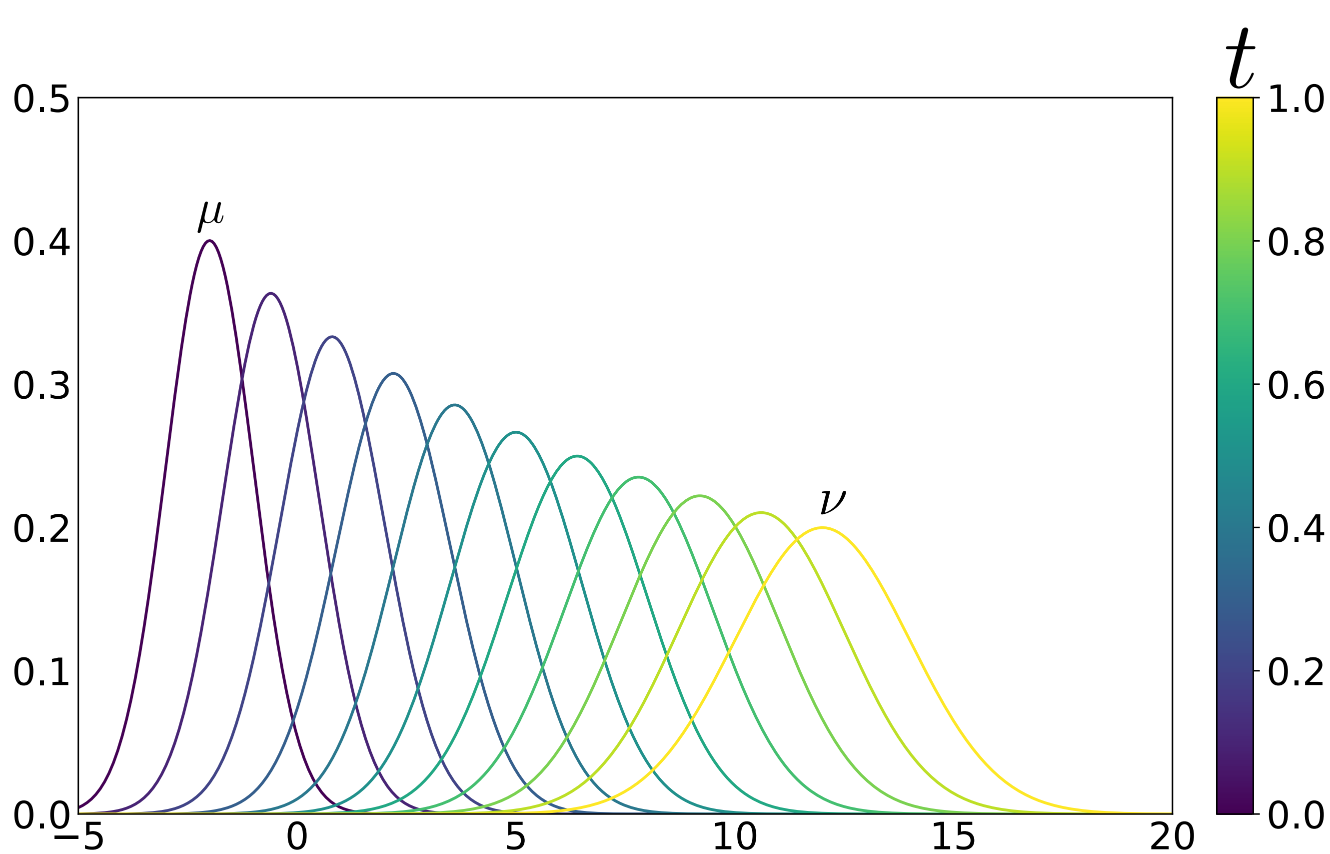
$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{\mathrm{d}})} \frac{1}{2} \sum_{i=1}^{p} \lambda_i W_2^2(\nu, \mu_i)$$

Wasserstein Geodesics are Barycenters



linear interpolation

$$t \mapsto (1-t)\mu + t\nu$$



Wasserstein geodesic
(McCann interpolation)

$$t \mapsto [(1-t)\mathrm{Id} + tT^*_{\mu\to\nu}]\#\mu$$

Tufts UNIVERSITY

The *Barycentric Coding Model*

- Let $\text{Bary}(\{\mu_i\}_{i=1}^p) = \{\nu_\lambda : \lambda \in \Delta^p\}$ 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

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

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

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

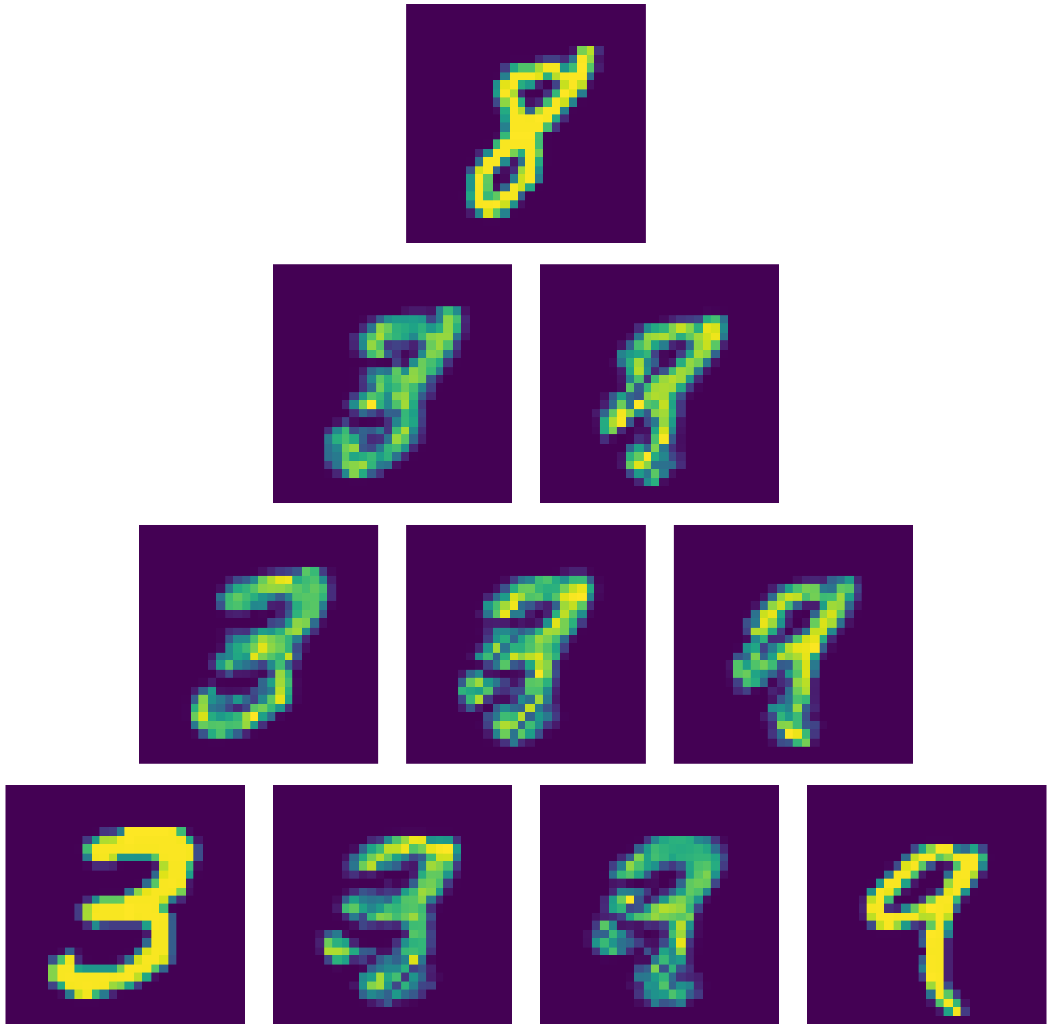
The Analysis Problem

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

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

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

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



- Can be thought of as histogram regression (Bonneel, Peyré, & Cuturi).

Exact Coefficients in BCM via Quadratic Program

Theorem. *(Werenski et al.) 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* $\mu_0 = \nu_{\lambda_*}$*.*

Remark: Holds in the exact case and can be generalized to *compatible measures* (when pairwise OT maps factor via composition).

Tufts UNIVERSITY

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 $M^i_{jk} = \frac{1}{2}||X_j - Y^i_k||_2^2$.

Solve for g^i as the optimal g in

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

entropy-regularized dual formulation (Kantorovich; Cuturi)

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

entropy-regularized OT map (Pooladian & Niles-Weed)

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 $\hat{\lambda} = \underset{\lambda\in\Delta^p}{\arg\min}\, \lambda^T \hat{A}\lambda$.

Tufts UNIVERSITY

Consistency of Entropic Estimation

Theorem. *(Werenski et al.) 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*

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

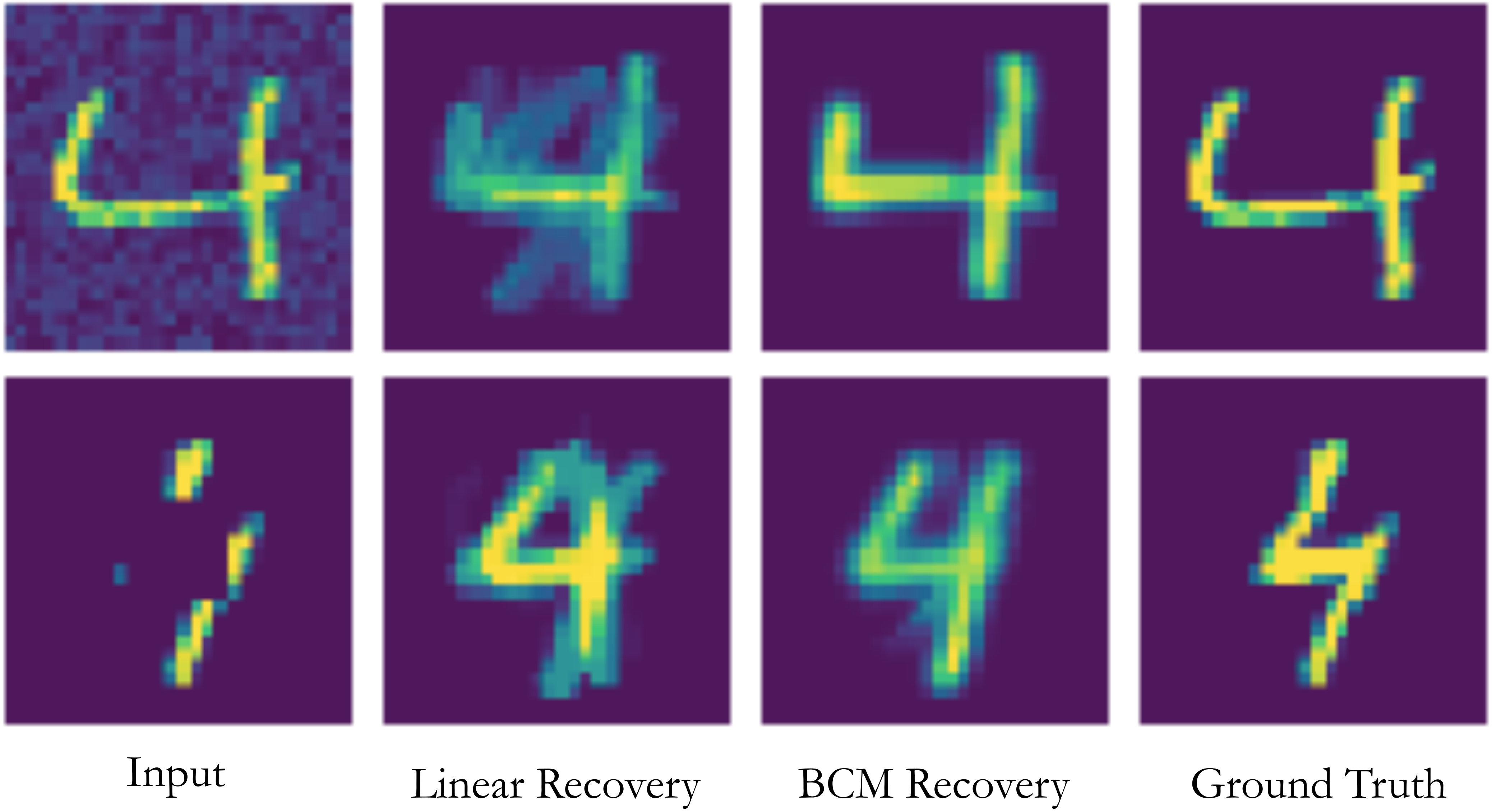
where $d' = 2\lceil d/2 \rceil$, and $\alpha \leq 3$ depends on the regularity of optimal maps.

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

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

- Convergence rate depends (poorly) on dimensionality and smoothness of OT maps.
- Regularity theory for OT maps relevant. Estimation of maps is provably hard!

Image Recovery



Remark: compared to running gradient descent, our closed-form program gives similar results with faster runtime.

Learning Reference Measures

- For a fixed dictionary $\mathscr{D} = \{\mathscr{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ and $\boldsymbol{\lambda} \in \Delta^m$, let

$$\mathrm{Bary}(\mathscr{D}, \boldsymbol{\lambda})$$

denote the barycenter generated from $\mathscr{D}$ with coordinates $\boldsymbol{\lambda}$.

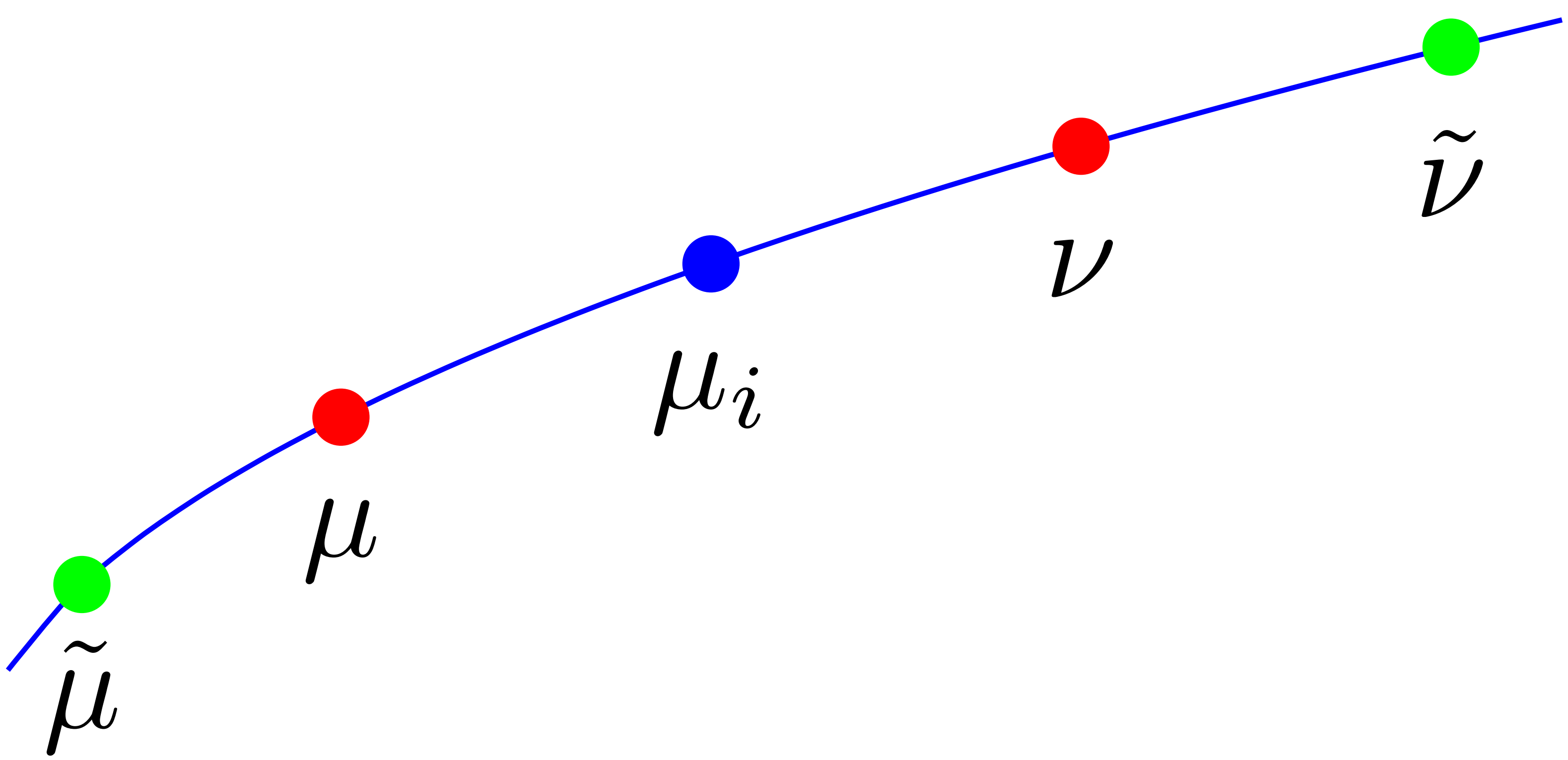
- Given observed data $\{\mu_i\}_{i=1}^n$, *Wasserstein dictionary learning (WDL)* (Schmitz et al.) solves:

$$(\mathscr{D}^*, \boldsymbol{\Lambda}^*) = \arg\min_{\mathscr{D}, \boldsymbol{\Lambda}} \sum_{i=1}^n W_2^2(\mathrm{Bary}(\mathscr{D}, \boldsymbol{\lambda}_i), \mu_i).$$

WDL Can Be Ill-Posed

- **Toy Example:** $\{\mu_i\}_{i=1}^n$ live on the Wasserstein geodesic between μ, ν.

- Then any measures $\tilde{\mu}, \tilde{\nu}$ that "extend" this geodesic can also generate $\{\mu_i\}_{i=1}^n$.



Geometric Wasserstein Dictionary Learning (GeoWDL)

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

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

- Learn a dictionary that *reconstructs well* using *nearby atoms;* $\rho > 0$.
- GeoWDL: $(\mathscr{D}^*, \mathbf{\Lambda}^*) = \arg\min_{\mathscr{D}, \mathbf{\Lambda}} \mathcal{G}(\mathscr{D}, \mathbf{\Lambda}, \{\mu_i\}_{i=1}^n)$.

Coding & Recovery Results

- Consider the coding problem for a *fixed* dictionary:

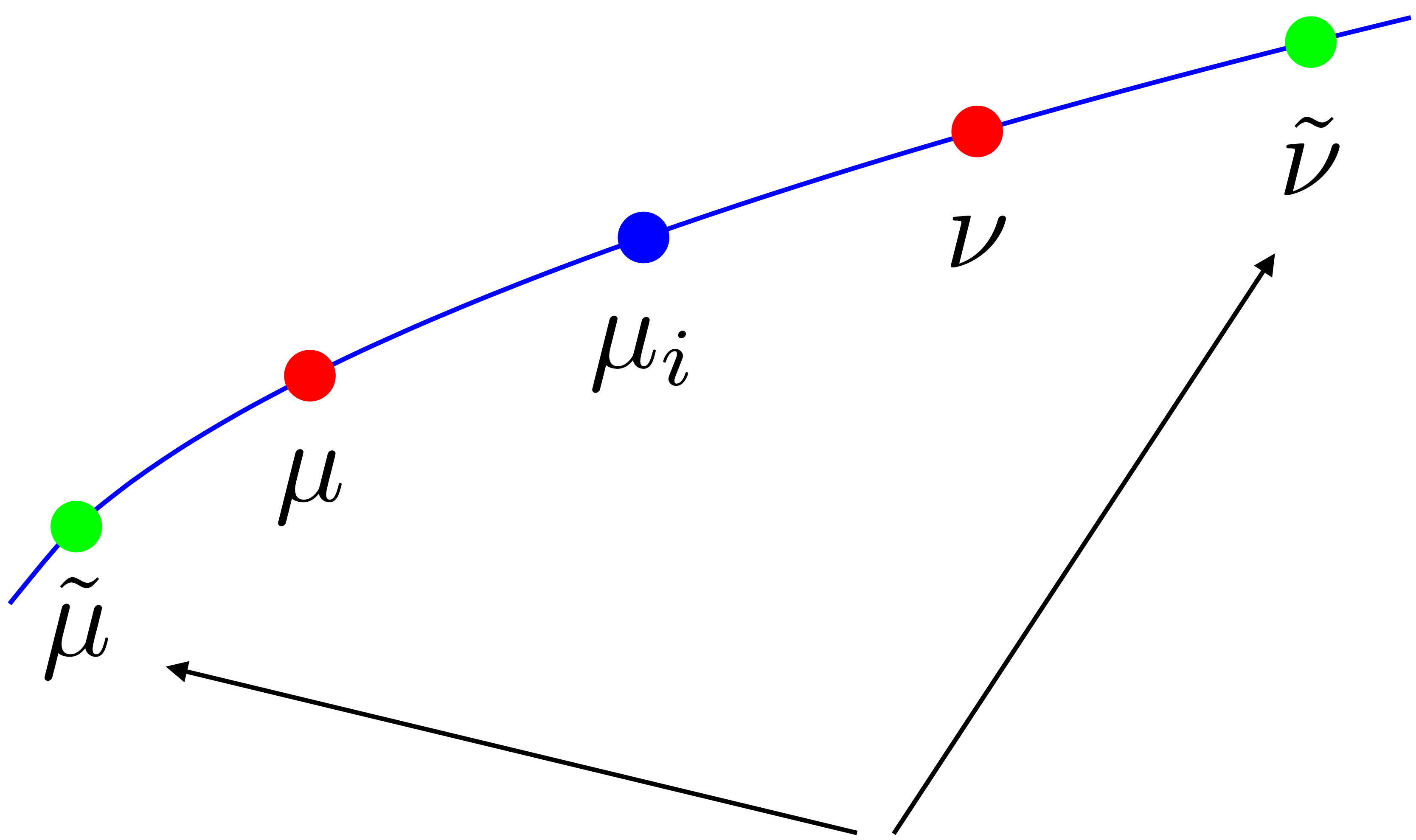
$$\begin{aligned}&\underset{\boldsymbol{\lambda}\in\Delta^m}{\arg\min}\sum_{j=1}^{m}\lambda_j W_2^2(\mathscr{D}_j,\mu)\\&\text{subject to } \mu=\mathrm{Bary}(\mathscr{D},\boldsymbol{\lambda}).\end{aligned}$$

- Generically, this has a unique solution and in particular resolves non-unique reconstruction issues.

- *Coefficient Properties, Informal:* Under particular generative models for μ, the optimal coefficients *concentrate and can exhibit sparsity*.

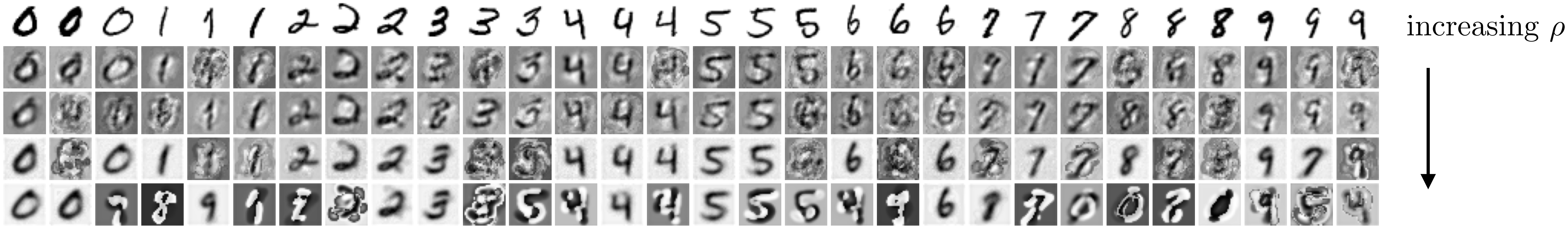
Provable Learning of Generators

- *Global Recovery, Informal:* If data lives on a Wasserstein geodesic (i.e., set of barycenters generated by μ, ν), minimizing the geometric regularizer subject to perfect reconstruction learns μ, ν .

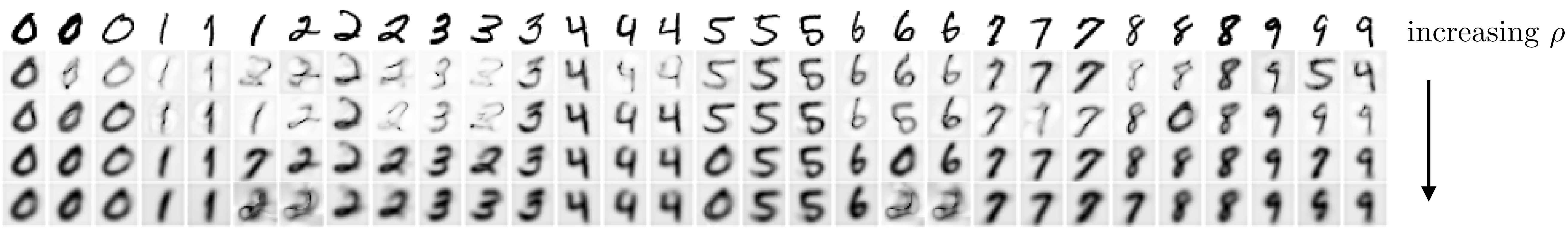


Geometric regularizer prevents learning these "extensions."

Learning MNIST: Linear v. Wasserstein

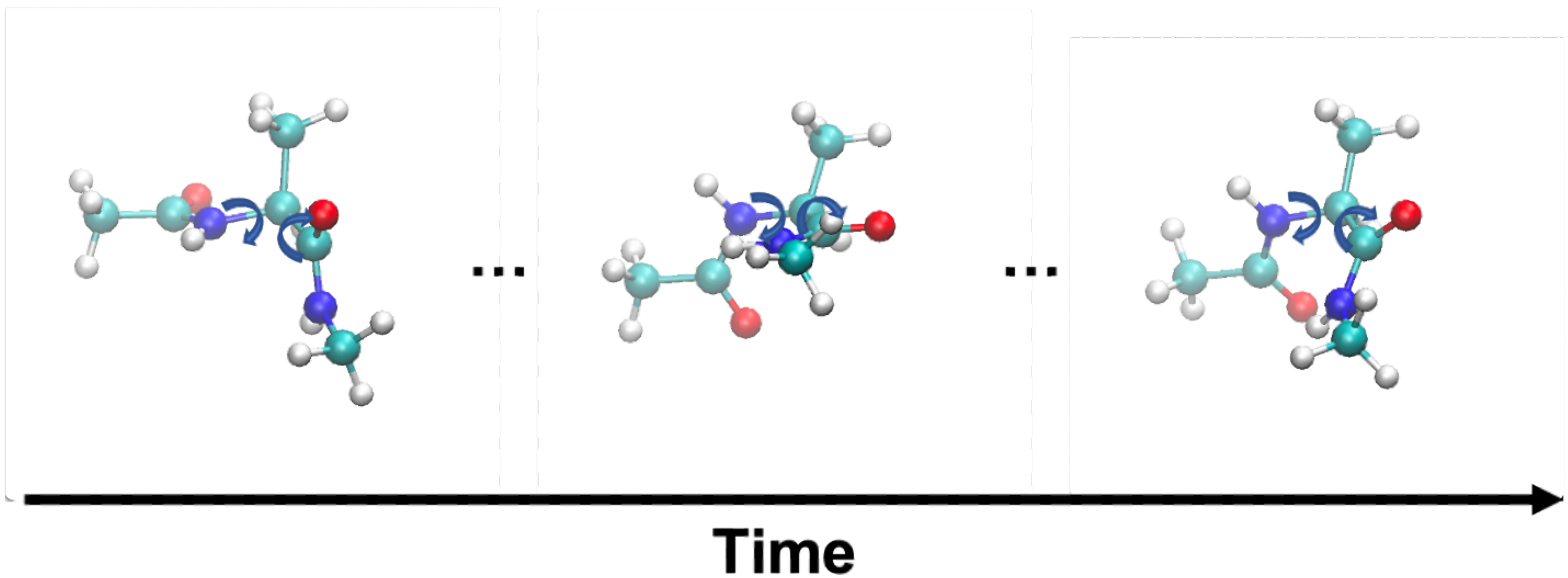


Linear Reconstruction Model with Geometric Regularization



GeoWDL

Reduced Order Modeling for MDS



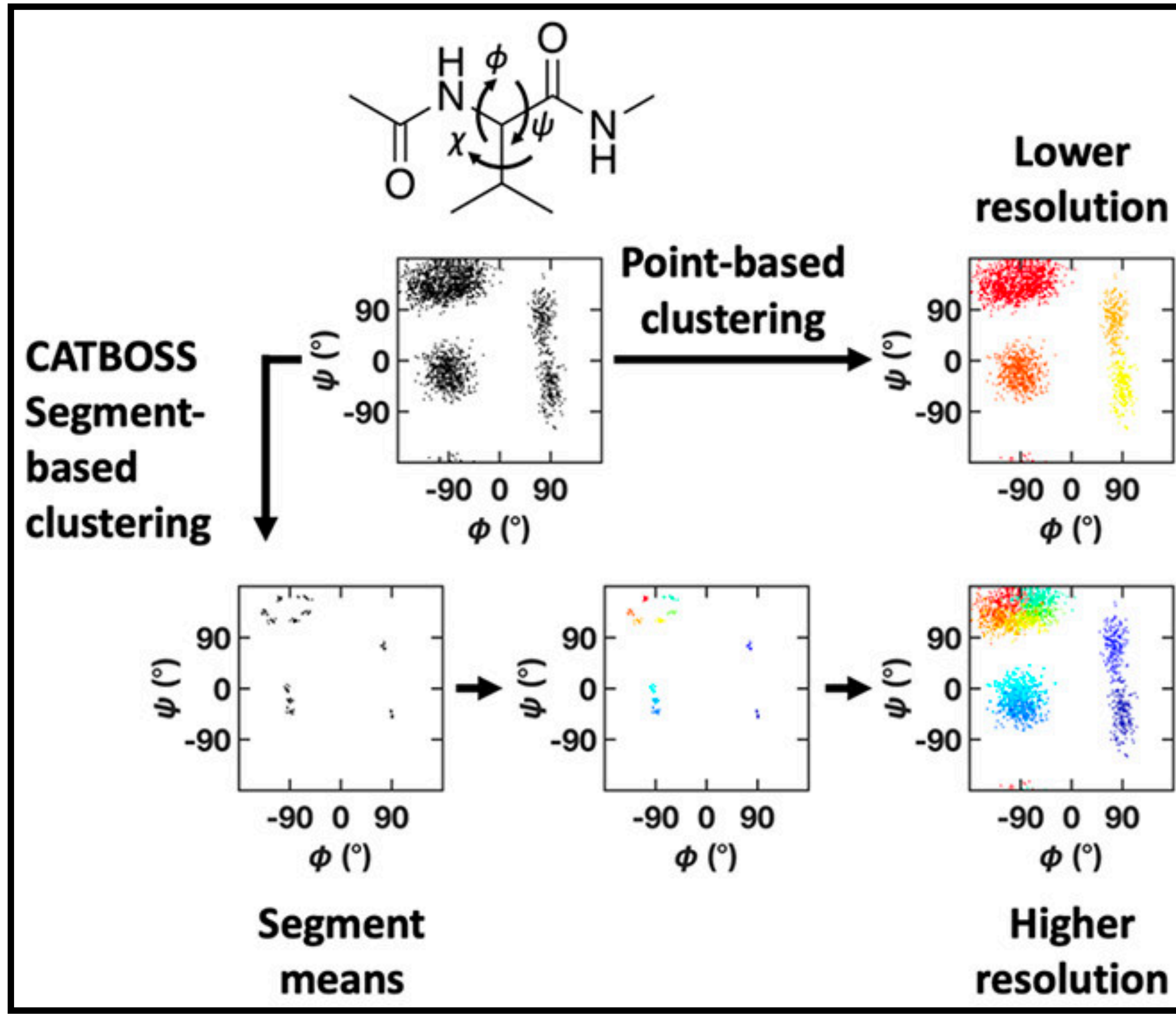
- Molecular dynamics simulations (MDS) are a crucial tool in computational chemistry.

- High-dimensional time series—want to identify canonical configurations (metastable states).

- How to characterize the global dynamics of the system?

“Cluster Analysis of Trajectories Based on Segment Splitting”

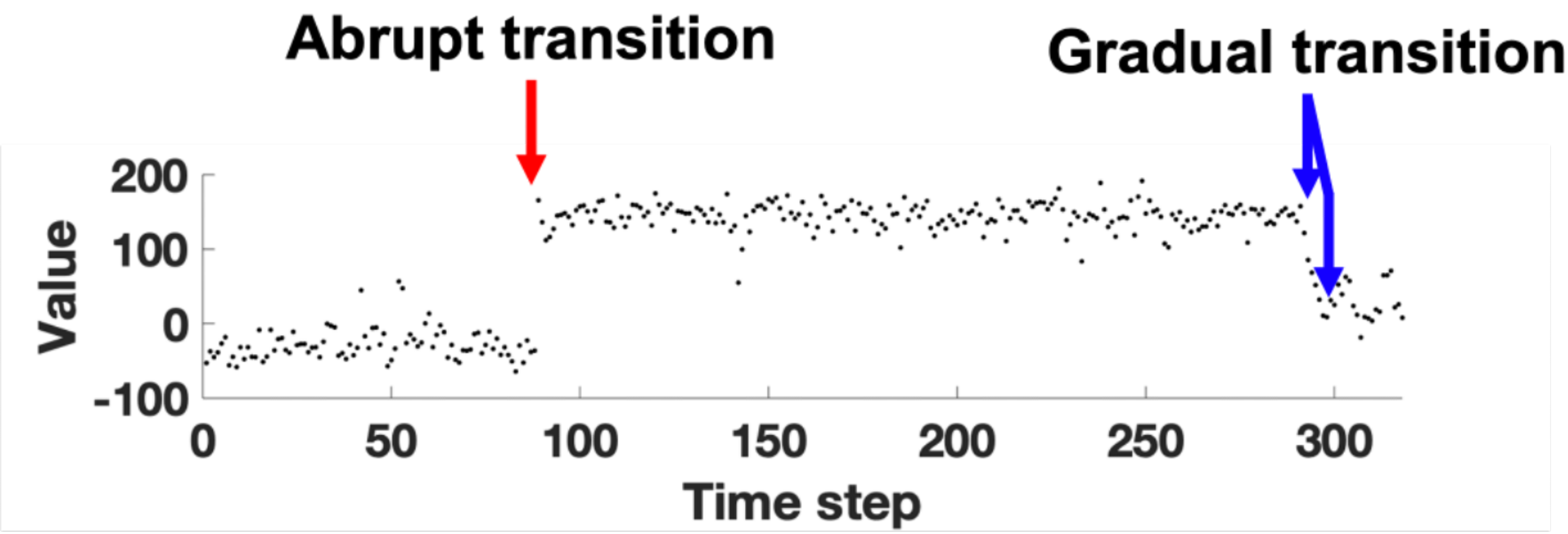
- CATBOSS: Cluster time *segments*, not points.
- Compare segments using *Wasserstein distances*.
- Improve speed, gain robustness to random fluctuations around a metastable state.



Tufts UNIVERSITY

Beyond Metastable States: Capturing Transitions

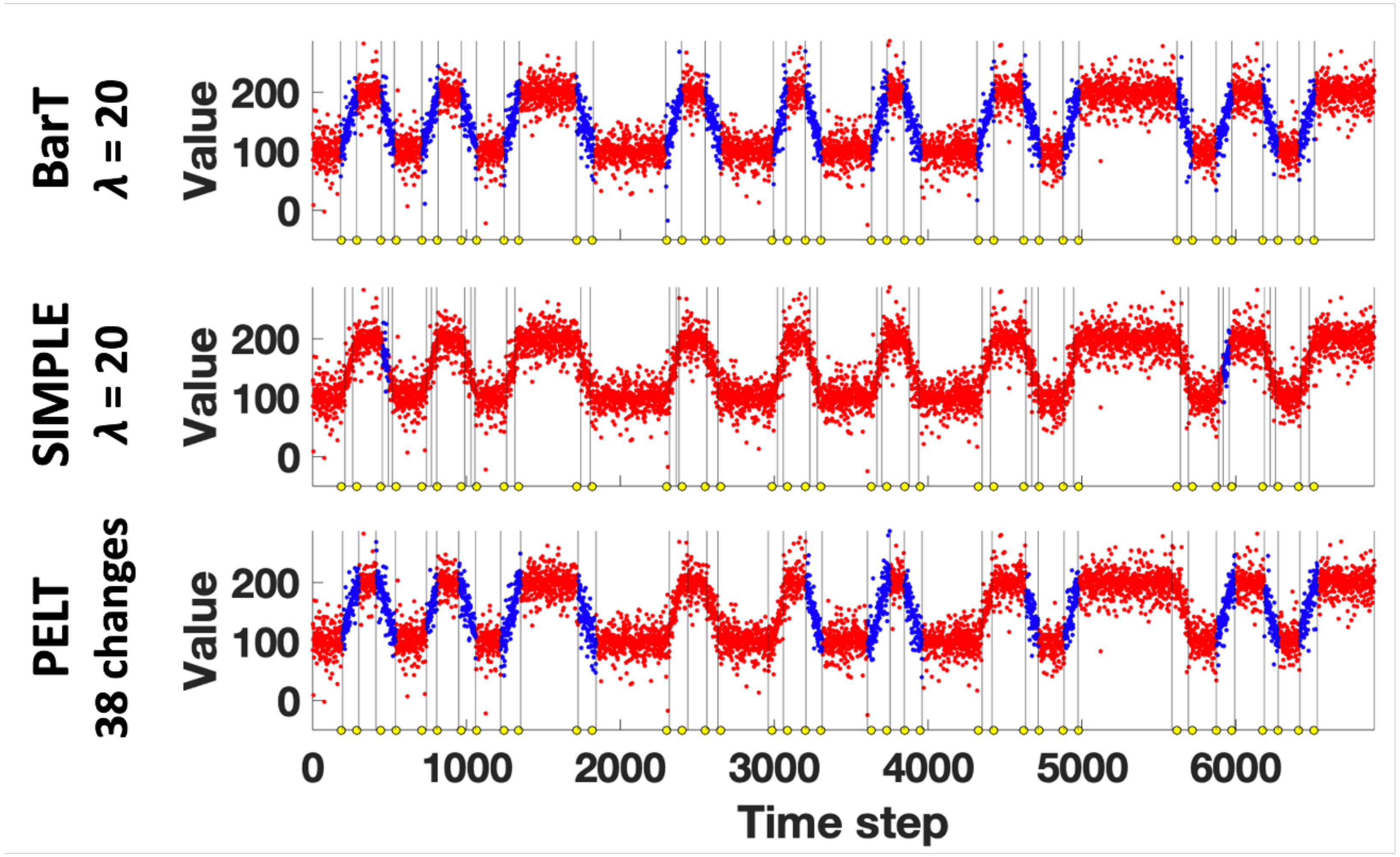
- Abrupt changes sometimes occur depending on sampling resolution, but gradual transitions also occur.
- How to characterize and learn them?



BarT: Barycentric Modeling of MDS Transitions

- Idea: transition regions are sampled from Wasserstein barycenters of two metastable states.

- **<u>BarT:</u>** incorporate this into change point detection and clustering.



- Limitation: initial efforts put parametric assumptions on metastable states (therefor on transition regions). WDL for non-parametric case?

Tufts UNIVERSITY

Papers & Support

Damjanovic, Lin, Murphy
"Modeling Changes in Molecular Dynamics Time Series as Wasserstein Barycentric Interpolations"
SAMPTA, 2023

Damjanovic, Murphy, and Lin.
"CATBOSS: Cluster analysis of trajectories based on segment splitting."
Journal of Chemical Information and Modeling, 2021

Masud, Werenski, Murphy, Aeron
"Multivariate Soft Rank Via Entropic Optimal Transport: Sample Efficiency and Generative Modeling"
Journal of Machine Learning Research, 2023

Mueller, Aeron, Murphy, Tasissa
"Geometric Sparse Coding in Wasserstein Space"
TAG in ML Workshop at ICML, 2023

Tasissa, Tankala, Murphy, Ba
"K-Deep Simplex: Manifold Learning via Local Dictionaries"
IEEE Transactions on Signal Processing, 2023

Werenski, Jiang, Tasissa, Aeron, Murphy
"Measure Estimation in the Barycentric Coding Model"
ICML, 2022

Werenski, Masud, Murphy, Aeron.
"On Rank Energy Statistics via Optimal Transport: Continuity, Convergence, and Change Point Detection"
IEEE Transactions on Information Theory, 2024

DISC

NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY
UNITED STATES OF AMERICA

NSF

DMS 1912737
DMS 1924513
DMS 2309519
DMS 2318894

THE CAMILLE & HENRY DREYFUS FOUNDATION

Tufts UNIVERSITY

Code and Contact Information

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

Contact: jm.murphy@tufts.edu

Thanks for Your Attention!

Tufts
UNIVERSITY