Learning On and Near Low-Dimensional Subsets of the Wasserstein Manifold

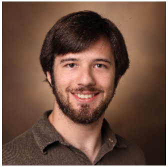
Alex Cloninger

University of California, San Diego
Department of Mathematics and
Halicioğlu Data Science Institute

Collaborators

Caroline Moosmüller (UNC)

Varun Khurana (UCSD)



Keaton Hamm (UTA)

Jinjie Zhang (UCSD)

Harish Kannan (UCSD)

Funding:

- NSF DMS 1819222 (collaborative with Xiuyuan Cheng)
- NSF DMS 2012266 (collaborative with Hrushikesh Mhaskar)
- NSF DMS 2111322 (Caroline)
- Intel Research

Outline

1. Learning on Distributions
2. Optimal Transport Theory and Embedding
   - Linear Optimal Transport
   - Compatible Transformations
3. Theory for Full Distributions
4. Theory for Finite Samples
5. Experiments

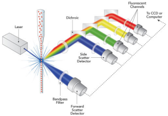
Learning on Distributions

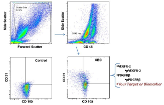
- Many applications where “data point” is a point cloud / distribution

$$X_k = \{x_i^{(k)}\}_{i=1}^{n_k} \sim \mu_k \in \mathbb{P}(\mathbb{R}^d), \qquad k \in \{1, ..., N\}$$

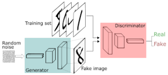
- Learning goals remain:
  - Construct classifier for “classes” of different μ_k
  - Clustering / dimension reduction to embed into low dimensional space

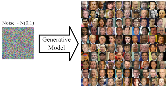
Flow Cytometry



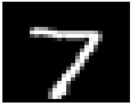


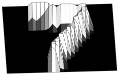
Generative Models





Images as Densities





Goals of Distribution Statistics

- What statistic to map to vector/Hilbert space $\phi : \mathbb{P} \to \mathcal{H}$
  - Deploy off the shelf classifiers
  - Ideally linear / low-complexity classifiers, regression, SVM
- Don't want ϕ to collapse information too much
  - Mean or first few moments far from injective
- Want to understand geometric structure of latent space
  - Minimal separation in $\mathbb{P}$ guarantees minimal separation in $\mathcal{H}$
- Want to be robust to perturbations and simple transformations
  - shifts / scalings / shearings / deformations
  - ϕ should be Lipschitz under these push forwards
- Would like ability to do distributed computation
  - features of data can be distributed across computers with low communication bandwidth

Comparison of Families of Distributions

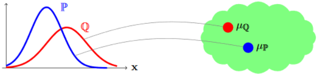
- **Question 1:** Given $\mu_k \in \mathbb{P}(\mathbb{R}^d)$ can we find embedding s.t.

$$d(\mu_k, \mu_\ell) \approx \|\phi_{\mu_k}(\cdot) - \phi_{\mu_\ell}(\cdot)\|?$$

  - Statistical distances expensive so want to avoid pairwise comparisons
  - How low dimensional can embedding ϕ be?

- **Question 2:** Given training data (μ_k, y_k) for $y_k \in \mathbf{C}$, can we find classifier

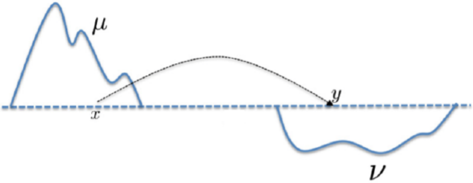
$$f : \mathbb{P} \to \mathbf{C}?$$

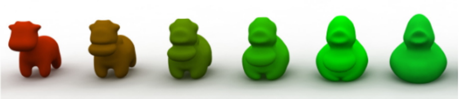


- **Question 3:** Can we control the embedding / classification error from sampling $X_k = \{x_i^{(k)}\}_{i=1}^{n_k} \sim \mu_k$ as a function of n_k?

Optimal mass transport (OMT)

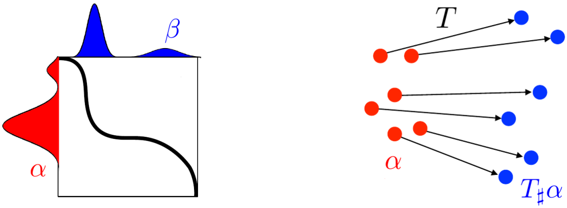
- Natural geometry for probability measures defined on a geometric space
- Move mass from pile into hole in the cheapest way possible respecting the underlying metric (Monge, 1781)





[Schmitz et al, SIAM J. Imaging Sci, 2018], [Solomon et al, SIGGRAPH, 2015]

Optimal mass transport (OMT)



- The **Wasserstein-2** distance between distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ is

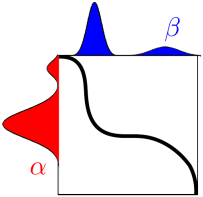
$$W_2(\mu,\nu)^2 = \min_{T \in \Pi_\mu^\nu} \int \| T(x) - x \|^2 \, d\mu(x).$$

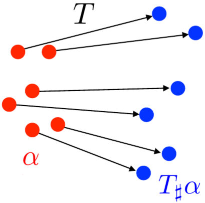
  with $T \in \Pi_\mu^\nu$ if $T_\sharp \mu = \nu$, i.e. $\mu(T^{-1}(A)) = \nu(A)$.
- The argmin is the **optimal transport map**, denote it by T_μ^ν. Exists and unique subject to regularity assumptions on μ.

[Brenier, Commun. Pure Appl. Math., 1991], [Kantorovich, Manag. Sci., 1958],

[Peyré et al, Found. Trends Mach. Learn., 2019]

Optimal mass transport (OMT)



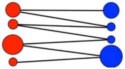


- The **Wasserstein-2** distance between distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ is

$$W_2(\mu,\nu)^2 = \min_{T \in \Pi_\mu^\nu} \int \|T(x) - x\|^2 \, d\mu(x).$$

with $T \in \Pi_\mu^\nu$ if $T_\sharp\mu = \nu$, i.e. $\mu(T^{-1}(A)) = \nu(A)$.

- The argmin is the **optimal transport map**, denote it by T_μ^ν. Exists and unique subject to regularity assumptions on μ.
- **Kantorovich relaxation**: Allow "mass splitting", coupling instead of function.



[Brenier, Commun. Pure Appl. Math., 1991], [Kantorovich, Manag. Sci., 1958],

[Peyré et al, Found. Trends Mach. Learn., 2019]

Outline

1. Learning on Distributions
2. Optimal Transport Theory and Embedding
   - Linear Optimal Transport
   - Compatible Transformations
3. Theory for Full Distributions
4. Theory for Finite Samples
5. Experiments

Outline

1. Learning on Distributions

2. Optimal Transport Theory and Embedding
   - Linear Optimal Transport
   - Compatible Transformations

3. Theory for Full Distributions

4. Theory for Finite Samples

5. Experiments

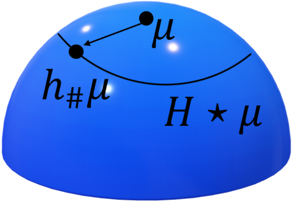
Dimensionality of Wasserstein Space

- $(\mathbb{P}, W_2)$ is infinite dimensional Riemannian manifold
  - Without assumptions on the data model, all hope is lost!
- **Assumption:** Data comes from push-forwards of one (or several) base distributions

$$\mathcal{H} \star \mu = \{h_\sharp \mu : h \in \mathcal{H}\}$$

$$(h_\sharp \mu)(A) = \mu(h^{-1}(A))$$

- **Assumption:** $\mathcal{H}$ made up of "simple" transformations (more on this later)
  - Push-forwards arc out low dimensional subsets of the Wasserstein manifold



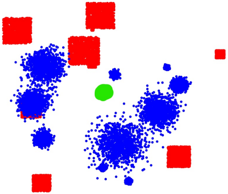
Linearized Optimal Transformation Features

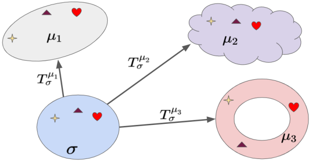
Instead of distances, think of transport plan as a new set of features

- LOT Embedding: fix reference distribution σ

$$\begin{aligned} F_\sigma :& \mathbb{P} \to L^2(\mathbb{R}^d, \sigma) \\ & \mu \mapsto T_\sigma^\mu \end{aligned}$$

- **Idea:** Define a registration between μ_i and μ_j according to how they are optimally aligned to σ
  - Hope this isn't too different than optimal registration of μ_i and μ_j





[Rohde et al. 2013, 2016, 2018 (algorithm, theory for CDF, algorithm involving Radon transform)]

Linear optimal transport (LOT)

- **Distance**:

$$W_2^{LOT}(\mu,\nu) = \|T_\sigma^\mu - T_\sigma^\nu\|_\sigma$$

- **Learning**:

$$\begin{aligned} f_\mu :\quad & \mathcal{P}(\mathbb{R}^d) \to \mathcal{C} \\ & \mu \mapsto f(T_\sigma^\mu) \qquad \text{for } f : L^2(\mathbb{R}^d,\sigma) \to \mathcal{C} \end{aligned}$$

Questions:

- What are natural actions $h \in \mathcal{H}$ for which $f(h_\# \mu) = f(\mu)$ is efficiently learnable?
- What does $W_2^{LOT}(\mu_i,\mu_j)$ tell us about $W_2(\mu_i,\mu_j)$?
- How stable is LOT w.r.t. perturbation push forwards h and w.r.t. finite sampling of μ_i?

[Rohde et al. 2013, 2016, 2018 (algorithm, theory for CDF, algorithm involving Radon transform)], [Aldroubi, et al. 2020 (Concurrent ArXiv pub., different theory)]

Outline

1. Learning on Distributions

2. Optimal Transport Theory and Embedding
   - Linear Optimal Transport
   - Compatible Transformations

3. Theory for Full Distributions

4. Theory for Finite Samples

5. Experiments

Compatible Transformations

- Need to find families of group actions that "interact nicely" with optimal transport
    - Easy to show that $(S \circ T_\sigma^\mu)_\# \sigma = S_\# \mu$
    - **Problem:** is this the optimal map from σ to $S_\# \mu$?

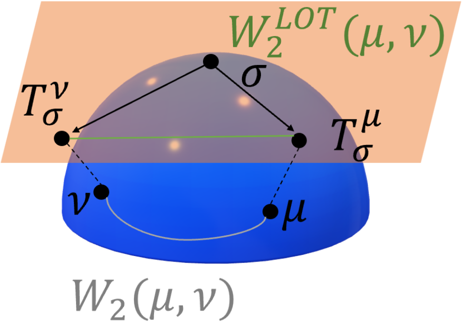
$$\begin{tikzcd}
\mu \arrow[r, "S_\#\mu"] & S_\#\mu \\
\sigma \arrow[u, "T_\sigma^\mu"] \arrow[ur, "T_\sigma^{S_\#\mu}"'] &
\end{tikzcd}$$

$$S \circ T_\sigma^\mu =?= T_\sigma^{S_\#\mu}$$

- Using optimal map basically regularizes the problem so that choice of a map from σ to μ_k is well-defined for each μ_k

LOT Distance on Wasserstein Manifold

- Distance between transport maps is distance on tangent plane at σ
- Looking for characterization of when tangent plane distance similar to Wasserstein distance
- **NB:** Possible to travel "far" from σ and still be near tangent plane



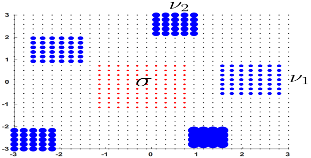
Compatible Action Examples Examples

With no assumptions on τ, σ

- **Shifts:** $\mu = (S_a)_{\#}\tau$ for $S_a(x) = x - a$
- **Scalings:** $\mu = (R_c)_{\#}\tau$ for $R_c(x) = c \cdot x$, $c > 0$
- **Affine:** Combination of shifts and scalings.
- **Why?** They satisfy $S = T_{\tau}^{S_{\#}\tau}$ (S is already optimal!)

With assumptions on τ, σ

- **Barycenter:** Any measure along OT curve
- **Principal axis shearing:** More on this later



Create tube around “simple” transformations

- **Perturbations:**
  $\mathcal{G}_{\varepsilon,R} = \{g \in L^2(\mathbb{R}^d) : \exists h \in A_{a,c} \text{ s.t. } \|g - h\| < \varepsilon \text{ and } \|h\| < R\}$

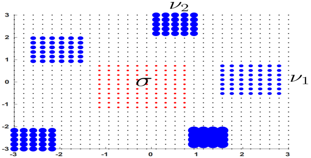
Compatible Action Examples Examples

With no assumptions on τ, σ

- **Shifts:** $\mu = (S_a)_{\#}\tau$ for $S_a(x) = x - a$
- **Scalings:** $\mu = (R_c)_{\#}\tau$ for $R_c(x) = c \cdot x$, $c > 0$
- **Affine:** Combination of shifts and scalings.
- **Why?** They satisfy $S = T_{\tau}^{S_{\#}\tau}$ (S is already optimal!)

With assumptions on τ, σ

- **Barycenter:** Any measure along OT curve
- **Principal axis shearing:** More on this later



Create tube around “simple” transformations

- **Perturbations:**
  $\mathcal{G}_{\varepsilon,R} = \{g \in L^2(\mathbb{R}^d) : \exists h \in A_{a,c}$ s.t. $\|g - h\| < \varepsilon$ and $\|h\| < R\}$
- **Open Q:** What are interesting deterministic functions in $\mathcal{G}_{\varepsilon,R}$?

Convexity For Group Actions

F_σ is *compatible* with action $\mathcal{H}$ if $\forall h \in \mathcal{H}$, $F_\sigma(h \star \mu) = h \star F_\sigma(\mu)$

- Ex: Shifts, scalings, combinations

Convexity (C., Moosmüller 2020)

If $\mathcal{H}$ convex and compatible w.r.t. F_σ, then $F_\sigma(\mathcal{H} \star \mu)$ is convex.

F_σ is $\delta-$*compatible* with action $\mathcal{H}$ if $\forall h \in \mathcal{H}$,

$$\|F_\sigma(h \star \mu) - h \star F_\sigma(\mu)\| < \delta$$

- Ex: Perturbations of shifts and scalings

Almost Convexity (C., Moosmüller 2020)

If $\mathcal{H}$ convex and $\delta-$compatible w.r.t. F_σ, then $F_\sigma(\mathcal{H} \star \mu)$ is $2\delta-$convex.

Means convex sum forms 2δ tube around set

Outline

1. Learning on Distributions
2. Optimal Transport Theory and Embedding
   - Linear Optimal Transport
   - Compatible Transformations
3. Theory for Full Distributions
4. Theory for Finite Samples
5. Experiments

Distances in LOT embedding space

Theorem (Almost Isometry (C., Moosmüller 2020))

Let σ, μ absolutely continuous and satisfy Caffarelli's regularity assumptions (convex supports). Let g, h be ε-perturbations of elementary transformations. Then we have

$$0 \leq \underbrace{W_2^{\mathrm{LOT}}(g_{\#}\mu, h_{\#}\mu)}_{\text{LOT Euclidean Dist.}} - \underbrace{W_2(g_{\#}\mu, h_{\#}\mu)}_{\text{Wasserstein-2 Dist.}} \leq C_{\sigma,\mu} \cdot \varepsilon + \overline{C_{\sigma,\mu}} \cdot \varepsilon^{1/2}$$

- **Corollary:** If g, h only shear+shift ($\varepsilon = 0$), then LOT is isometry.
- **Key proof ingredient:** $\frac{1}{2}$−Hölder type regularity:

$$W_2^{\mathrm{LOT}}(g_{\#}\mu, h_{\#}\mu) \leq c_1 \|g - h\|_\mu + c_2 \|g - h\|_\mu^{1/2}$$

  Basically follows from results by N. Gigli (2011)

- **Computational improvement:** To compute the $\binom{N}{2}$ distances between N distributions $g_{i\#}\mu$ need only N expensive OTs and $\binom{N}{2}$ cheap Euclidean distances.

Distances in LOT embedding space

Theorem (Almost Isometry (C., Moosmüller 2020))

Let σ, μ absolutely continuous and satisfy Caffarelli's regularity assumptions (convex supports). Let g, h be ε-perturbations of elementary transformations. Then we have

$$0 \leq \underbrace{W_2^{\text{LOT}}(g_\# \mu, h_\# \mu)}_{\text{LOT Euclidean Dist.}} - \underbrace{W_2(g_\# \mu, h_\# \mu)}_{\text{Wasserstein-2 Dist.}} \leq C_{\sigma,\mu} \cdot \varepsilon + \overline{C_{\sigma,\mu}} \cdot \varepsilon^{2/15}$$

- **Corollary:** If g, h only shear+shift ($\varepsilon = 0$), then LOT is isometry.
- **Key proof ingredient:** $\frac{2}{15}$−Hölder type regularity:

  $$W_2^{\text{LOT}}(g_\# \mu, h_\# \mu) \leq c_1 \|g - h\|_\mu + c_2 \|g - h\|_\mu^{2/15}$$

  Basically follows from results by Merigot et al. (2020)
- **Computational improvement:** To compute the $\binom{N}{2}$ distances between N distributions $g_{i\#}\mu$ need only N expensive OTs and $\binom{N}{2}$ cheap Euclidean distances.

Learning in LOT embedding space

Theorem (Linear Classifiers for Distributions (C., Moosmüller 2020))

Let σ absolutely continuous, $\mathcal{H}$ convex, and F_σ $\epsilon-$compatible with μ and ν orbits from $\mathcal{H}$. If

- *$\mathcal{H} \star \mu$, $\mathcal{H} \star \nu$ compact, and*
- *minimal separation $W_2(h_{1\#}\mu, h_{2\#}\nu) > \delta$ for $\delta = O(\epsilon^c)$,*

then $F_\sigma(\mathcal{H} \star \mu)$ and $F_\sigma(\mathcal{H} \star \nu)$ are linearly separable.

Theorem (Minimal Separation)

For g_1, g_2 perturbations of shifts and scalings, and μ, ν satisfying Caffarelli's regularity assumptions, $\delta = \max(\delta(\varepsilon, R, \sigma, \mu), \delta(\varepsilon, R, \sigma, \nu))$ where

$$\delta(\varepsilon, R, \sigma, \mu) := \left(\sqrt{\frac{4R}{K_\mu^\sigma}} + 2\right) \|f_\mu\|_\infty^{1/2} \varepsilon + \left(4R \|f_\mu\|_\infty^{1/2} \frac{W_2(\sigma, \mu) + R + \|Id\|_\mu}{K_\mu^\sigma}\right)^{1/2} \varepsilon^{1/2}$$

Proof Idea

- First version of this result by Rohde et. al. 2018 for $d = 1$ and $\varepsilon = 0$ ($\delta = 0$ in this case).
- Uses **Hahn-Banach theorem**. Key proof ingredient: Under compatibility condition: Convexity of $\mathcal{H}$ is preserved via LOT.
- Subresult: If $\mathcal{H}$ is convex and F_σ is (almost) compatible with action by $\mathcal{H}$, then $F_\sigma(\mathcal{H}_\sharp\mu)$ is (almost) convex.

Tailored References and Certain Deformations

Theorem (Conditions on transformations (Khurana, Kannan, C., Moosmüller 2022))

Same assumptions as above. If the Jacobian of T_σ^μ has a constant orthonormal basis given by an orthogonal matrix P (i.e. $J_{T_\sigma^\mu}(x) = P^\top D(x)P$), then

$$\mathcal{F}(P) = \left\{ x \mapsto P^\top \begin{bmatrix} f_1((Px)_1) \\ f_2((Px)_2) \\ \vdots \\ f_n((Px)_n)) \end{bmatrix} + b : \begin{matrix} f_j : \mathbb{R} \to \mathbb{R} \text{ is monotonically} \\ \text{increasing and differentiable} \\ \text{and } b \in \mathbb{R}^n \end{matrix} \right\}.$$

is the set of transformations for which the compatibility condition holds.

- Means we can shear as long as it respects the eigen-directions of the transport
- Comes down to maintaining convexity of $\nabla\varphi$

Tailored References and Certain Deformations

Theorem (Conditions on transformations (Khurana, Kannan, C., Moosmüller 2022))

Same assumptions as above. If the Jacobian of T_σ^μ has a constant orthonormal basis given by an orthogonal matrix P (i.e. $J_{T_\sigma^\mu}(x) = P^\top D(x) P$), then

$$\mathcal{F}(P) = \left\{ x \mapsto P^\top \begin{bmatrix} f_1((Px)_1) \\ f_2((Px)_2) \\ \vdots \\ f_n((Px)_n)) \end{bmatrix} + b : \substack{f_j:\mathbb{R}\to\mathbb{R} \text{ is monotonically} \\ \text{increasing and differentiable} \\ \text{and } b \in \mathbb{R}^n} \right\}.$$

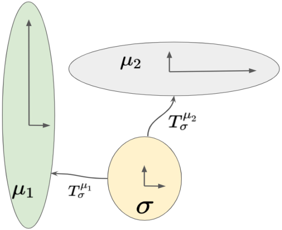
is the set of transformations for which the compatibility condition holds.

- Means we can shear as long as it respects the eigen-directions of the transport
- Comes down to maintaining convexity of $\nabla\varphi$

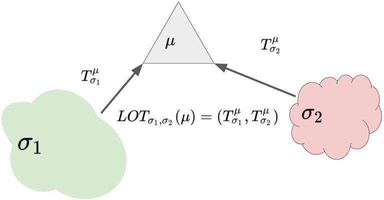
Motivated by tailored references, have equivalent theory of separability for multiple references $\{\sigma_i\}$

Tailored References and Certain Deformations

Certain Deformations:



Multiple References:


μ
$T_{\sigma_2}^{\mu}$
$T_{\sigma_1}^{\mu}$
$LOT_{\sigma_1,\sigma_2}(\mu) = (T_{\sigma_1}^{\mu}, T_{\sigma_2}^{\mu})$
σ_2
σ_1

Khurana, Kannan, C., Moosmüller 2022

Outline

1. Learning on Distributions

2. Optimal Transport Theory and Embedding
   - Linear Optimal Transport
   - Compatible Transformations

3. Theory for Full Distributions

4. Theory for Finite Samples

5. Experiments

Multiple Sources of Error

- Data sets only given through finite samples $X_k = \{x_i^{(k)}\}_{i=1}^{n_k} \sim \mu_k$
    - Not inherently an issue to finding transport map since σ still absolutely continuous
- Must sample $\{z_i\}_{i=1}^{m} \sim \sigma$ in order to have finite memory storage of each transport map $O(dm)$
- Solving transport on finite samples of σ leads to mass splitting and transport coupling $P_Z^{X_k} \in \mathbb{R}^{m \times n_k}$
    - Need to create approximate transport map $\widehat{T}_Z^{X_k} \in \mathbb{R}^{m \times d}$ through barycentric projection

$$\widehat{T}_Z^{X_k} = D^{\dagger} P_Z^{X_k} X_k, \text{ where } D_{ii} = \sum_j P_Z^{X_k}[i,j]$$

- **Goal:** Bound $\left| \frac{1}{\sqrt{m}} \| \widehat{T}_Z^{X_k} - \widehat{T}_Z^{X_\ell} \|_F - \| T_\sigma^{\mu_k} - T_\sigma^{\mu_\ell} \|_\sigma \right| < ?$

Known Rates

Easy Case: $\mathrm{supp}(\mu_k) \subset B(0, R)$ and connected

- Sampling of σ can be bounded by McDiarmid's inequality
- Transport $\mathbb{E}\| T_Z^{X_k} - T_\sigma^{\mu_k} \|_\sigma$ can be bounded for
  - finding transport through linear programming: Deb et al (2021)
  - finding transport through Sinkhorn: Niles-Weed et al (2021)

Difficult Case: μ_k have potentially unbounded support with tail decay faster than $\|x\|^{-d-2}$

- Requires need for $\mathrm{supp}(\sigma) \subset B(0, R)$
- Construct fictitious compactly supported $\widetilde{\mu_k}$ indistinguishable from μ_k given samples
- Bound distance between transports using Marigot et al (2021)

$$\| T_\sigma^\mu - T_\sigma^{\widetilde{\mu}} \|_\sigma \leq C \cdot W_1(\mu, \widetilde{\mu})^{\frac{1}{6}}$$

- Use “small set where r.v. unbounded” alternative to McDiarmid's inequality

Finite Sample Bounds

Theorem (Finite LOT Error (Khurana, Moosmüller, Hamm, C. 2022))

Let $\mu_k \in \mathcal{H} \star \mu$ where $\mathcal{H}$ are $\epsilon-$compatible transformations. Under technical assumptions on the smoothness of the densities of μ_k, given a finite sampling of size n and sampling of m points from σ, then with probability at least $1-\delta$,

$$\left|W_2(\mu_i,\mu_j)^2 - W_{2,LOT}(X_j,X_k)^2\right| \leq C_R\left(\epsilon^c + O(n^{-1/d}) + \sqrt{\frac{2\log(2/\delta)}{m}}\right).$$

Exponent of *n* can be ameliorated using kernel smoothing and smoothness of density

Finite Sample Bounds

Theorem (Finite LOT Error (Khurana, Moosmüller, Hamm, C. 2022))

Let $\mu_k \in \mathcal{H} \star \mu$ where $\mathcal{H}$ are $\epsilon-$compatible transformations. Under technical assumptions on the smoothness of the densities of μ_k, given a finite sampling of size n and sampling of m points from σ, then with probability at least $1-\delta$,

$$\left| W_2(\mu_i, \mu_j)^2 - W_{2,LOT}(X_j, X_k)^2 \right| \leq C_R \left(\epsilon^c + O(n^{-1/d}) + \sqrt{\frac{2\log(2/\delta)}{m}} \right).$$

Exponent of *n* can be ameliorated using kernel smoothing and smoothness of density

Given above bound:

- simply need slightly larger minimal separation between $\mathcal{H} \star \mu$ and $\mathcal{H} \star \nu$ to guarantee linear separability
- apply MDS perturbation results to guarantee stable low-dimensional embedding

Unsupervised Embedding

Theorem (Low Dimensional Embedding (Khurana, Moosmüller, Hamm, C. 2022))

Given $\{\mu_k\} \subset \mathcal{H} \star \mu$ of ϵ tube around compatible transformations and that has true low-dimensional embedding $Y \subset \mathbb{R}^{d'}$. Then the left singular vectors U of centered transport maps

$$\left[T_Z^{X_k}\right]_{k=1}^N - \frac{1}{N} 1_N^T \left[T_Z^{X_k}\right]_{k=1}^N$$

satisfies, for sufficiently large n and m and with probability $1-\delta$,

$$\min_{Q\in\mathcal{O}(d')} \|U - YQ\|_F \le C \cdot N \cdot \|Y^\dagger\| \left(\epsilon^c + O(n^{-1/d}) + \sqrt{\frac{2\log(2/\delta)}{m}}\right).$$

- Combines finite bound with Wassmap embedding of full distributions (Hamm et al 2021) and MDS Perturbation bound (Arias Castro et al 2020)
- Means we can reduce to $d' = C \cdot d$ embedding near isometrically

Outline

1. Learning on Distributions

2. Optimal Transport Theory and Embedding
   - Linear Optimal Transport
   - Compatible Transformations

3. Theory for Full Distributions

4. Theory for Finite Samples

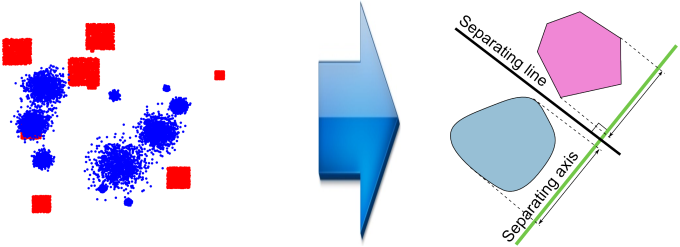
5. Experiments

Relevance in Experimental Setting

- Classification model

$$f(h_{\#}\mu) = f(\mu) \text{ for } h \in \mathcal{G}_{\varepsilon,R}$$

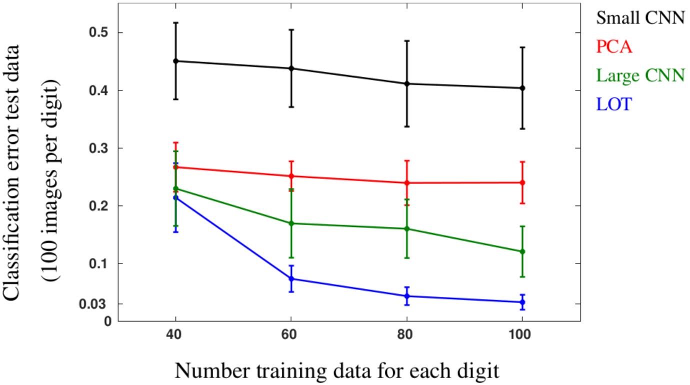
- Hyperplane classifier f_θ between $\mathcal{H} \star \mu$ and $\mathcal{H} \star \nu$
  - Sample σ to project T^{μ}_{σ} to T^{X}_{Z}
  - Reduce dimension through:
    - PCA if data on shared computer
    - JL-embedding if data on distributed computers



Experimental Validation (shifts and scalings)

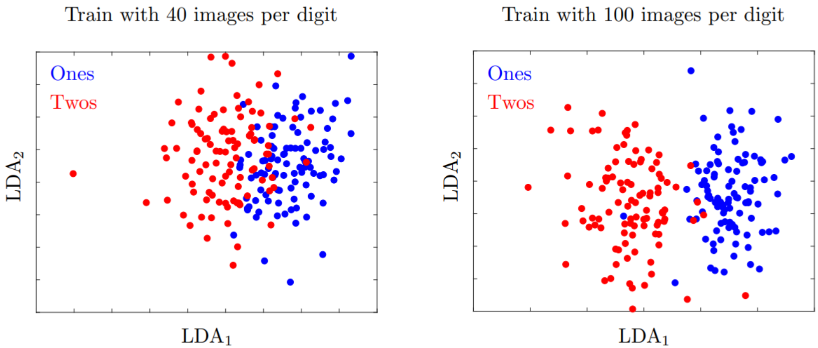
MNIST Classification Between 1's and 2's

- Data sampled from MNIST images
- Each image additionally augmented by random shift and scaling
- Sample *k* labeled examples of each class for training
- σ is centered normal distribution



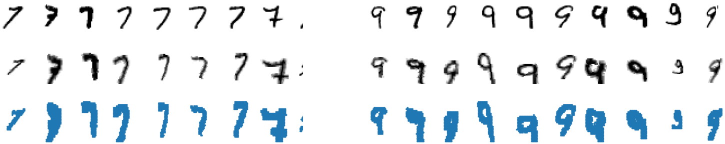
Experimental Validation (shifts and scalings)

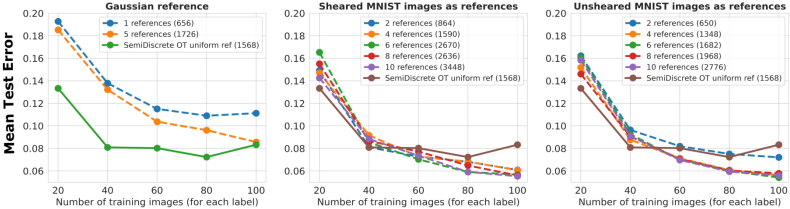
LDA embedding of test data



Experimental Validation (mild shearing)

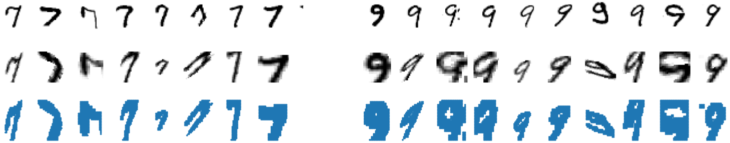
MNIST Classification Between 7's and 9's

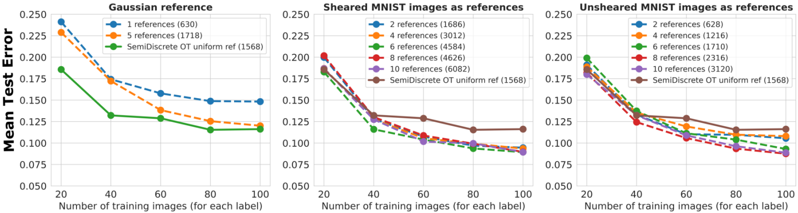




Experimental Validation (severe shearing)

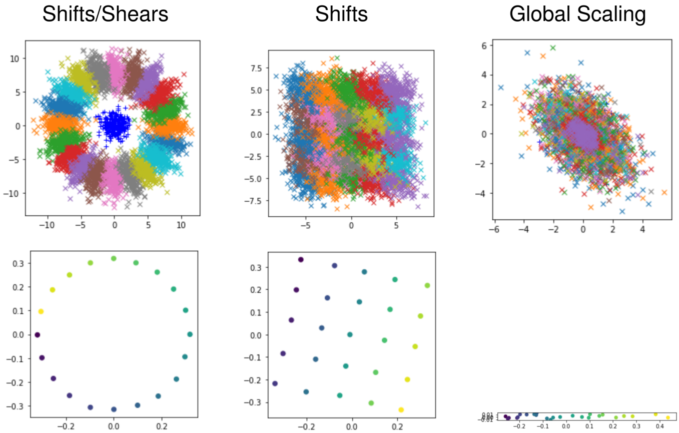
MNIST Classification Between 7’s and 9’s





PCA Embedding

- $\mathrm{supp}(\mu_k) \subset \mathbb{R}^2$
- $\mathcal{H}$ made up of almost compatible transformations



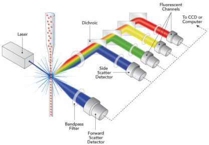
Shifts/Shears

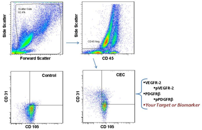
Shifts

Global Scaling

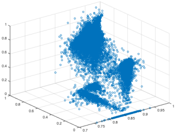
Flow Cytometry - Preliminary Results

- Flow cytometry: each patient is represented by 9D point cloud of cells

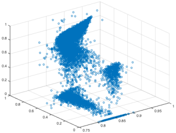




- Used to tell if people have blood disease
  - Medical test is to look at every 2D slice



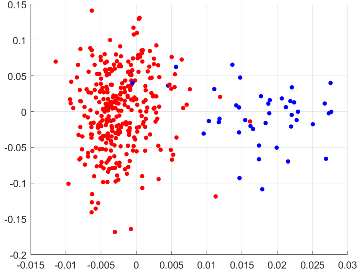
Healthy



AML

Flow Cytometry - Preliminary Results

- LOT using uniform reference distribution
- Sample transport maps and PCA reduce to 25 dimensions
- Fit linear SVM

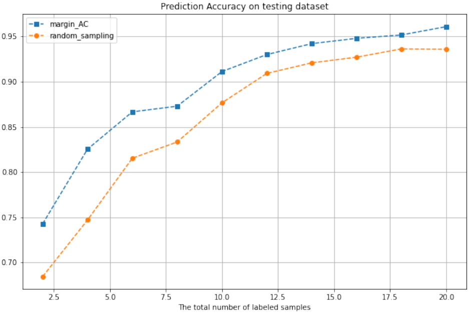


$n_{sick,tr} = 15$, $n_{healthy,tr} = 50$, $n_{sick,te} = 28$, $n_{healthy,te} = 265$

	Predicted Sick	Predicted Healthy
Sick	0.9286	0.0714
Healthy	0.0113	0.9887

Active Learning - Preliminary Results

Shifted 1s and 2s
Embed into LOT space
PCA Reduce space to lower dimension
Iteratively choose 2 labels per step
Refine sampling based of margin of remaining possible separators



LOT Summary

LOT feature space

- **Pro:** Requires only N OT computations, instead of $\binom{N}{2}$
- **Pro:** Well understood geometric structure in embedding space
- **Pro:** Linear separator is efficient to learn
  - Current research direction involving active learning
- **Con:** Assumes base distributions are absolutely continuous
  - Current research redefining compatibility to be a function of μ for measures with atoms
  - Current research direction involving entropic regularization on fixed grid

Thank you!

Questions?

References

1. C. Moosmüller, A. Cloninger. *Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations,* Information and Inference, 2022.
2. V. Khurana, H. Kannan, A. Cloninger, C. Moosmüller. *Learning sheared distributions using linearized optimal transport*, Sampling Theory, Signal Processing, and Data Analysis, 2022.
3. V. Khurana, C. Moosmüller, K. Hamm, A. Cloninger. *Stability of LOT Embeddings on Point Clouds*, To appear, 2022.
4. J. Zhang, C. Moosmüller, A. Cloninger. *Active learning of distributions with linearized optimal transport*, To appear, 2022.