Learning tasks in the Wasserstein space

Caroline Moosmüller

University of North Carolina at Chapel Hill

Joint with Alex Cloninger, Keaton Hamm, Harish Kannan, Varun Khurana, Jinjie Zhang

Kantorovich initiative seminar, Oct 27, 2022

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Outline

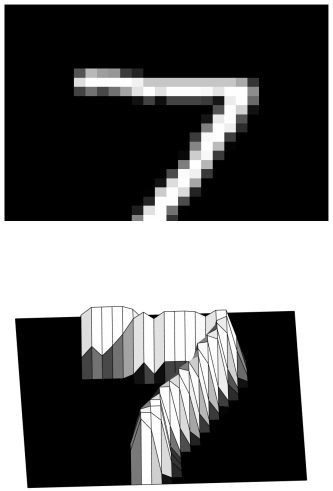
Comparison of Distributions

- **Question 1:** Given distributions $\mu_i, i = 1, \ldots, N$ (or data points sampled from μ_i), find all pairwise distances

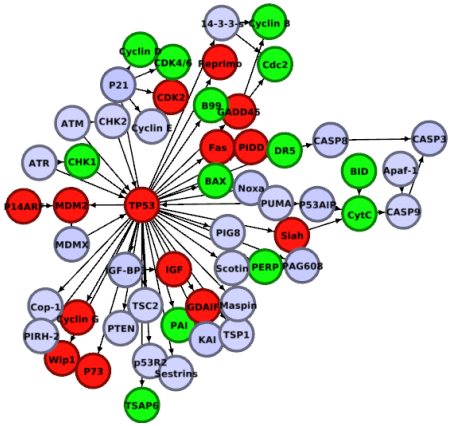
$$d(\mu_i, \mu_j) = \quad ?$$

$\rightarrow$ Unsupervised learning

Images as Distributions



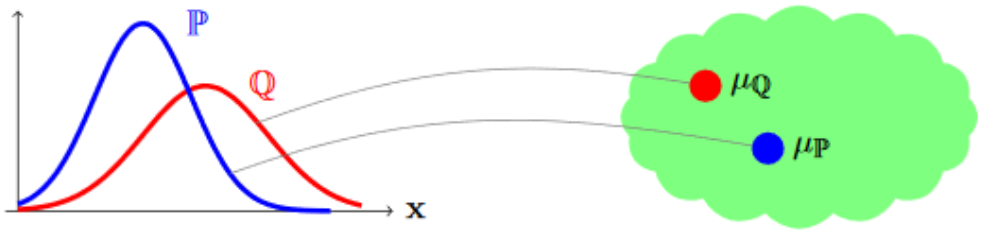
Gene expression data



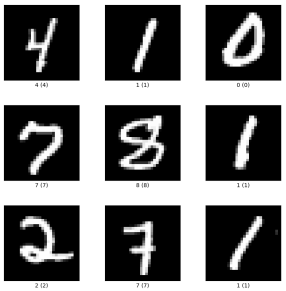
Learning distributions

- **Question 2:** Given training data (μ_k, y_k), with classes $y_k \in \mathbf{C}$

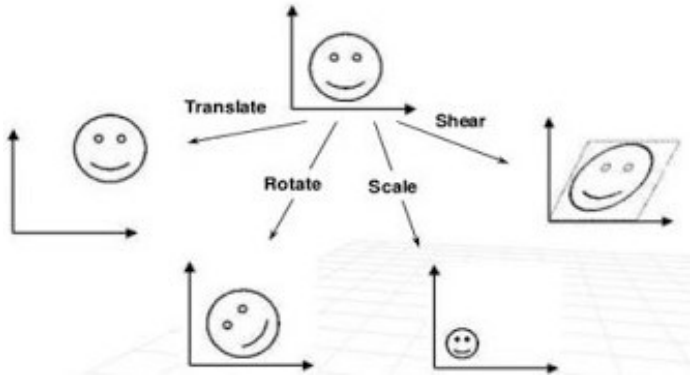
Learn a function: $f : \mathcal{P} \to \mathbf{C}$



Classify digits



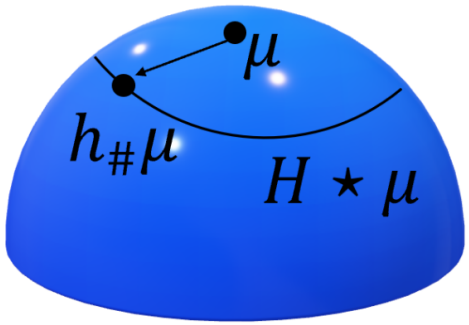
Transformation invariance



Assumptions and goals

We address question 1 and 2 by embedding $\mathcal{P}$ into an L^2-space using optimal transport. Then use linear distance and linear classifier.

- Assumptions
  - Distributions are generated by simple functions $\mathcal{H} \subset L^2(\mathbb{R}^d)$, i.e. $\mathcal{P} = \{h_\sharp \mu : h \in \mathcal{H}\}$



  - $\mathcal{H}$ consists of shifts, scalings, shearings, perturbations
  - Computation scales with complexity of $\mathcal{H}$

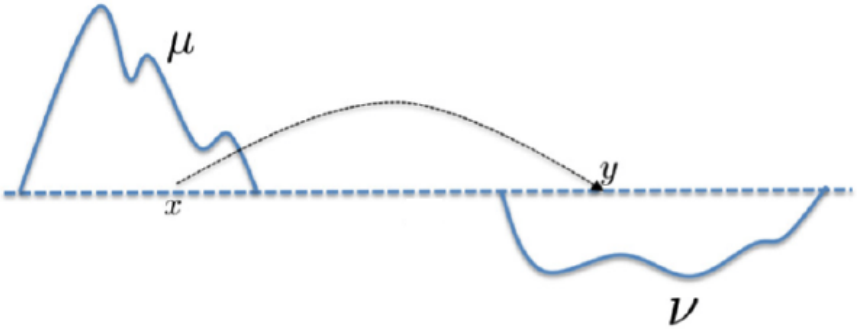
- Goal for feature space
  - **Unsupervised**: Euclidean distance in feature space approximates Wasserstein distance between distributions
  - **Supervised**: Separability of different families of distributions using linear methods

Outline

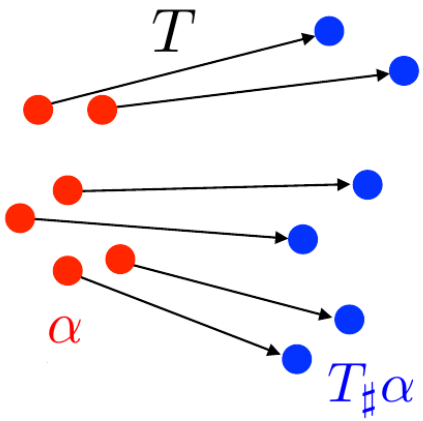
Outline

Optimal mass transport (OMT)

- Move mass from pile into hole in the cheapest way possible respecting the underlying metric (Monge, 1781)



- Find function T with $T_\sharp \mu = \nu$ that minimizes work



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

- The argmin is the **optimal transport**, T_μ^ν, the min is **Wasserstein distance**
- Exists and unique subject to regularity assumptions on μ, ν.

Pros and Cons of OMT

- **Pros**: W_2 is a metric on $\mathcal{P}(\mathbb{R}^d)$, well-developed theory, interpolation, incorporates geometry of space
- **Con**: Computation of T_μ^ν **usually slow** (linear program).
  - **Help**: add regularizer, Sinkhorn algorithm
  - **Still**: pairwise distances between $\mu_k, k = 1, \ldots, N$ needs $\binom{N}{2}$ OMT computations.
- **Con**: Complexity independent of the family of distributions
- **Learning?** unsupervised: Wasserstein distance!
  supervised: Embedding needed!

[Cuturi, NIPS 2013], Book: "Computational optimal transport" by Peyré, Cuturi, 2019

Outline

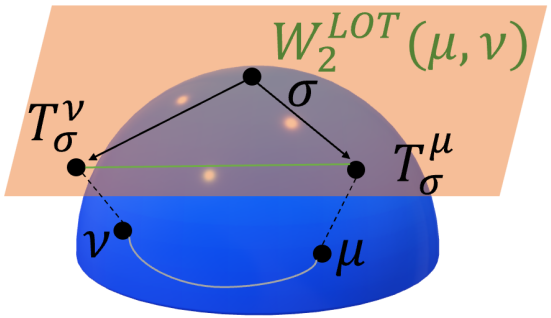
Linear optimal transport (LOT)

Think of transport plan as a new set of features.

- **LOT embedding**: Pick a reference distribution σ:

$$F_\sigma : \quad \mathcal{P}(\mathbb{R}^d) \to L^2(\mathbb{R}^d, \sigma)$$
$$\mu \mapsto T_\sigma^\mu$$

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



- **Learning**:

$W_2(\mu, \nu)$

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

Questions: For which distributions $\mathcal{P}(\mathbb{R}^d)$? $W_2^{LOT}(\mu, \nu) \approx W_2(\mu, \nu)$?

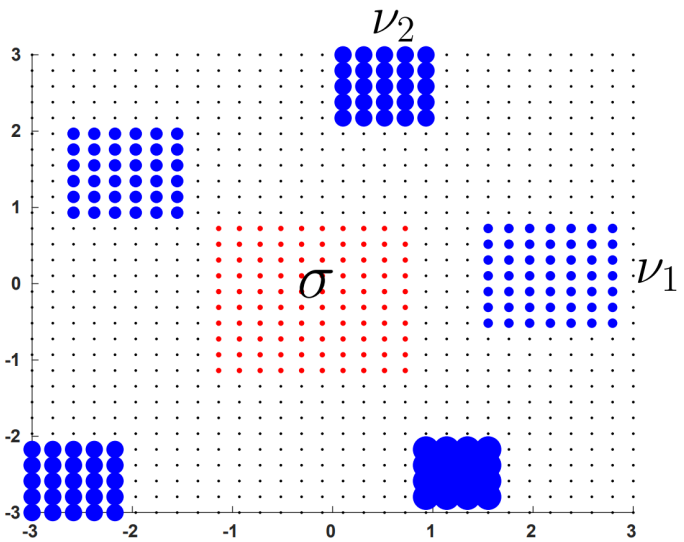
[Rohde et al. 2013, 2016, 2018], [Aldroubi, et al. 2021]

For which distributions?

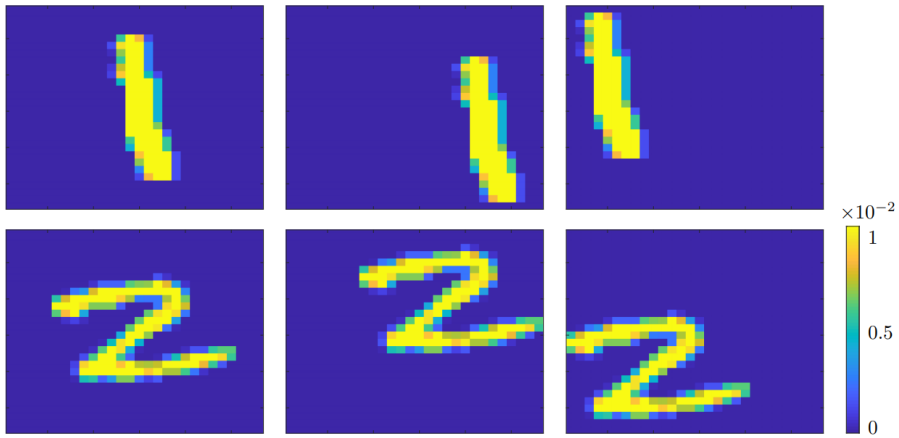
Consider distributions μ created by simple deformations of a template distribution τ:

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

Shifts and scalings of template σ

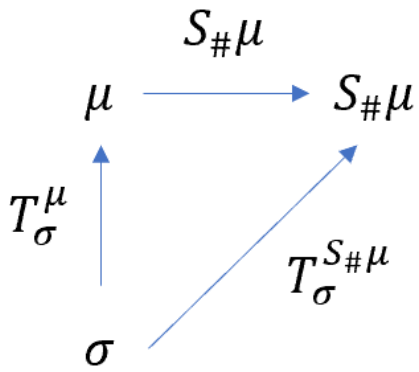


Shifts of template 1 and 2



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 $S \circ T_\sigma^\mu$ the optimal map from σ to $S_\# \mu$?



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

Key observations

- This composition is the identity

$$\begin{array}{ccccc} \mathcal{P}(\mathbb{R}^d) & \to & L^2(\mathbb{R}^d, \tau) & \to & \mathcal{P}(\mathbb{R}^d) \\ \mu & \mapsto & T_\tau^\mu & \mapsto & T_{\tau\,\#}^{\mu}\tau = \mu \end{array}$$

  i.e. the pushforward map is the **left-inverse** to LOT.
- Is it also **right-inverse**?

$$\begin{array}{ccccc} L^2(\mathbb{R}^d, \tau) & \to & \mathcal{P}(\mathbb{R}^d) & \to & L^2(\mathbb{R}^d, \tau) \\ h & \mapsto & h_\#\tau & \mapsto & T_\tau^{h_\#\tau} \stackrel{?}{=} h, \end{array}$$

- In general: no! Yes for shifts + scalings, and for shears with symmetric positive definite matrix. Others?
- Compatibility condition: $T_\tau^{h_\#\tau} = h$.
- **Almost compatibility condition**:

$$\| T_\tau^{h_\#\tau} - h\|_\tau < \varepsilon.$$

  Hence we allow perturbations of affine transformations.

Outline

Distances in LOT embedding space

Theorem (Almost Isometry (M., Cloninger 2022))

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

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

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

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

  Basically follows from results by N. Gigli (2011). No regularity, but weaker bound in [Merigot et al. 2020].
- **Computational improvement:** To compute the $\binom{N}{2}$ distances between N distributions $g_{i\#}\tau$ need only N expensive OTs and $\binom{N}{2}$ cheap Euclidean distances.

Learning in LOT embedding space

Theorem (Linear Classifiers for Distributions (M., Cloninger 2022))

Let σ, τ_1, τ_2 absolutely continuous in $\mathcal{P}(\mathbb{R}^d)$, $\mathcal{H}$ convex set of ε-perturbations of elementary transformations. If

- $\mathcal{H}_\sharp \tau_1$, $\mathcal{H}_\sharp \tau_2$ compact, and
- minimal distance $W_2(h_{1\#}\tau_1, h_{2\#}\tau_2) > \delta$,

then $F_\sigma(\mathcal{H}_\sharp \tau_1)$ and $F_\sigma(\mathcal{H}_\sharp \tau_2)$ are linearly separable in $L^2(\mathbb{R}^d, \sigma)$.

- δ can be given explicitly based on $\sigma, \tau_1, \tau_2, \varepsilon$.
- 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: 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 $F_\sigma(\mathcal{H}_\sharp \tau)$ is (almost) convex.

Theorem (Conditions on transformations (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

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

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

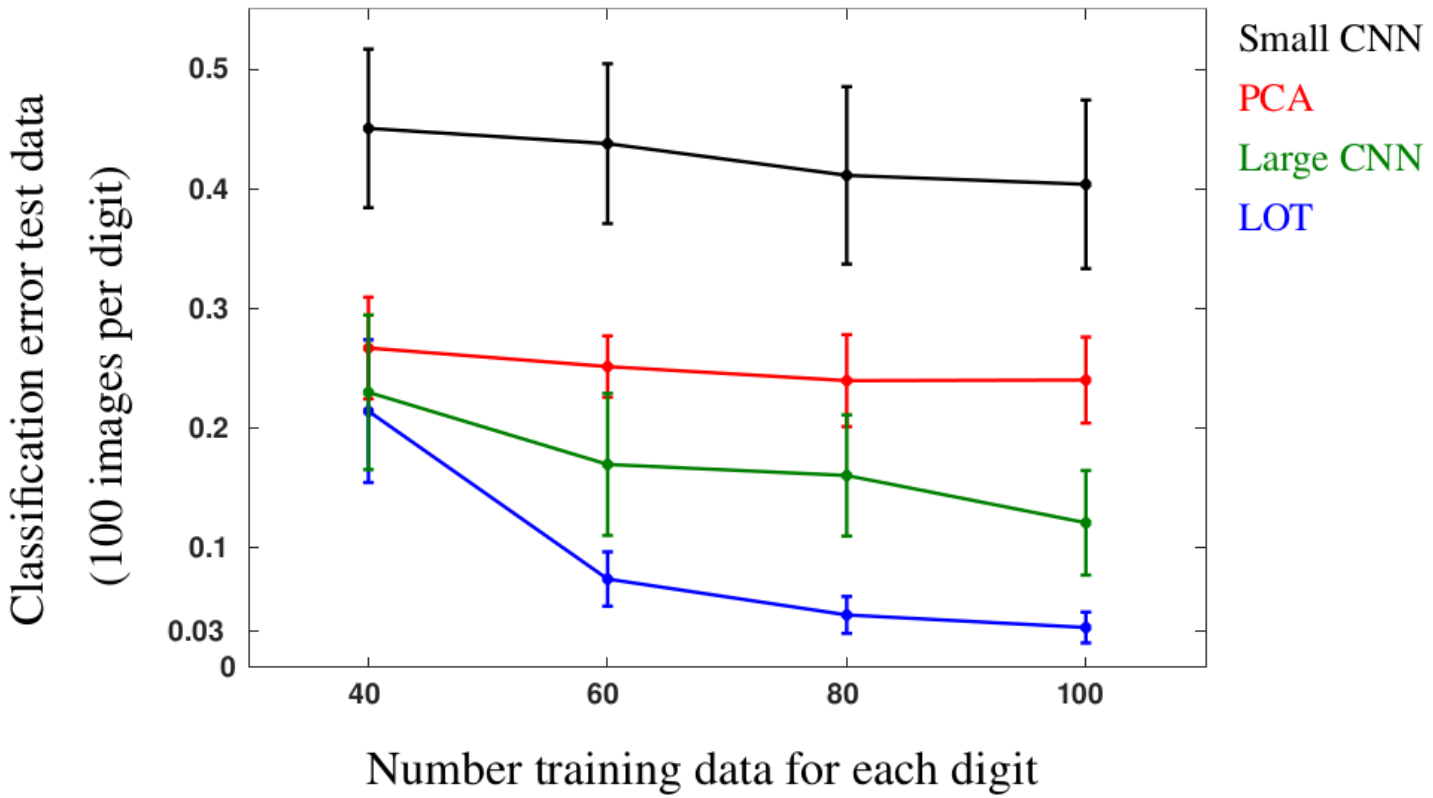
- In particular $S(x) = Ax + b$ with $A = P^T DP$ (diagonalizable by P).
- Comes down to maintaining convexity of $\nabla \varphi$

Outline

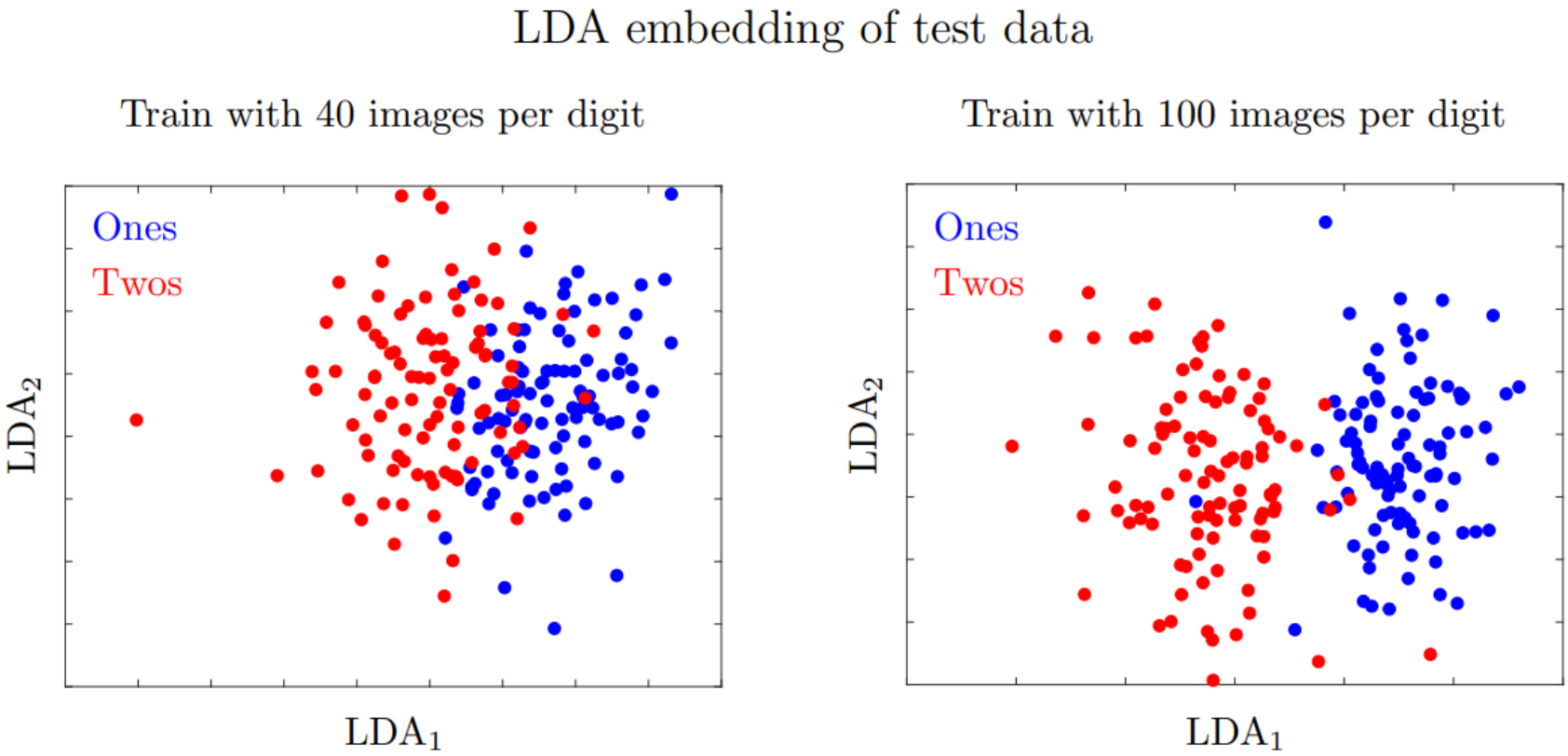
Experimental Validation

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

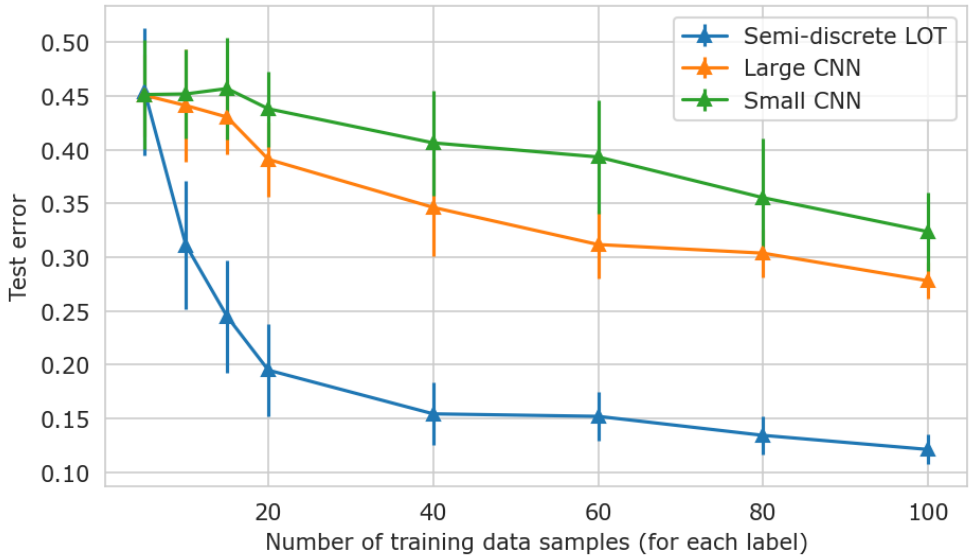


Experimental Validation (mild shearing)



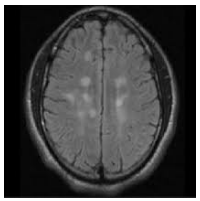
Experimental Validation (severe shearing)

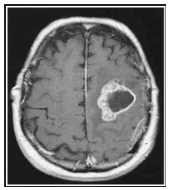
MNIST Classification Between 7's and 9's

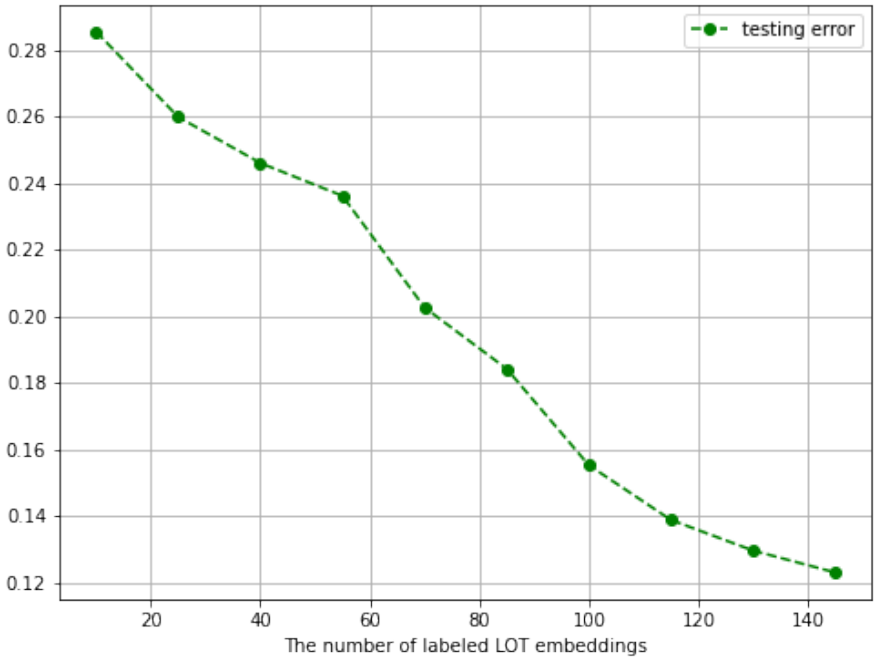


Active learning in LOT space

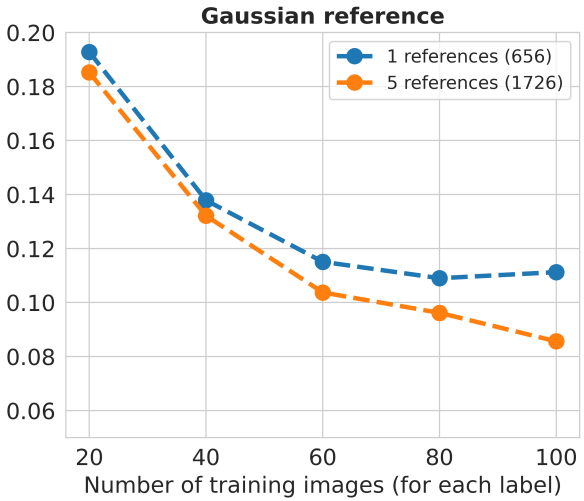
- Iteratively choose 5 labels per step
- Refine sampling based of margin of remaining possible separators

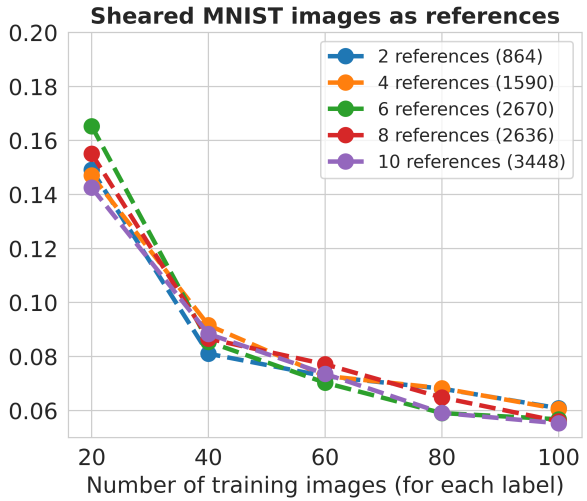


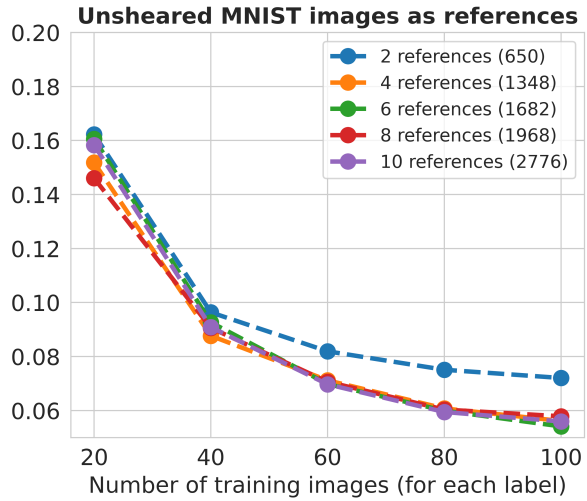




Multiple embeddings, different references







- Use **multiple embeddings** to improve separability
- Quantify **number of embeddings** needed to achieve given separation level $\delta > 0$ allowing functions bounded by L
- References **from the data set** achieve better separation

Outline

Summary

LOT feature space

- **Pro:** In shift/scalings/perturbation set-up, requires only N expensive OT computations, instead of $\binom{N}{2}$
- **Pro:** Apply linear methods in embedding space to separate classes of distributions, can be learned efficiently.
  - Current research: Active learning
- **Con:** What happens beyond an ε perturbation?
  - Current research: Redefine compatibility, deal with references which are not absolutely continuous
- **Con:** Still: Slow to compute each of N embeddings
  - Current research: Combine with entropic regularization on grid
- Current research: Deal with samples, find error bounds

Thank you!

Questions?

References

- V. Khurana, H. Kannan, A. Cloninger, C. Moosmüller. *Learning sheared distributions using linearized optimal transport*, Sampling Theory, Signal Processing, and Data Analysis, 2022.
- J. Zhang, C. Moosmüller, A. Cloninger. *Active learning of distributions with linearized optimal transport*, working paper 2022.
- C. Moosmüller, A. Cloninger. *Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations,* Information and Inference: A Journal of the IMA, 2022.