Permutation invariant networks to learn Wasserstein metrics

Arijit Sehanobish Neal G. Ravindra David van Dijk

Internal Medicine (Cardiology) and Computer Science, Yale University

Introduction

- Understanding the space of probability measures on *Polish* space $\mathcal{X}$ under a Wasserstein metric W_p is an important problem in mathematical analysis.
- The Wasserstein metric has received lot of interest, particularly in Computer Vision, for its principled way of comparing distributions.
- Despite widespread use, the metric suffers from high computational cost, and robustness and non-differentiability issues.

Research Goals

- Can we propose a neural network that correctly computes the Wasserstein distance between 2 measures for out-of-training distributions?
- What properties of measures does such a network learn?
- What properties of Wasserstein space can be preserved in our encoded space?

Definitions and Theory

- $W_p(\mu,\nu) := \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \mathbb{E}(|X-Y|^p)^{1/p}, \quad p \geq 1$
- Alternate notation: $W_p(X,Y)$ if $X\sim\mu, Y\sim\nu$.
- $\mathbb{P}(\mathcal{X})$ under W_p complete and separable metric space.
- If X, Y degenerate at x, y, $W_p(X,Y) = |x-y|$.
- (Scaling) $W_p(aX, aY) = |a| W_p(X,Y), \forall\, a \in \mathbb{R}$.
- (Translation invariance) $W_p(X+x, Y+x) = W_p(X,Y),\ \forall x \in \mathcal{X}$
- $\mathbb{P}(\mathcal{X})$ flat space under W_1 but (sectional) curvature is non-negative under W_2.
- (Topology generated by W_p) **(i)** If $\mathcal{X} \subset \mathbb{R}^n$ is compact and $p \in [1,\infty)$, in the space $\mathbb{P}(\mathcal{X})$, we have $\mu_k \to \mu$ iff $W_p(\mu_k,\mu) \to 0$.
  (ii) If $\mathcal{X} = \mathbb{R}^n$, then $W_p(\mu_k,\mu) \to 0$ iff $\mu_k \to \mu$ and $\int |x|^p d\mu_k \to \int |x|^p d\mu$

Model

- Draw samples of size N from various distributions.
- Use DeepSets architecture to encode this set to get an permutation invariant encoding of the samples.
- Train encoder H_θ such that,

$$||H_\theta(X) - H_\theta(Y)|| = SD_p^\lambda(X,Y)$$

  SD_p^λ= Sinkhorn Distance, an entropy regularized Wasserstein distance.
- H_θ regarded as a Siamese Network, allowing us to compare samples from distributions.
- Our Wasserstein Loss function reads,

$$L_{wass} = \frac{1}{\binom{m}{2}} \sum (||H_\theta(S_X) - H_\theta(S_Y)|| - SD_p^\lambda(\mu,\nu))^2$$

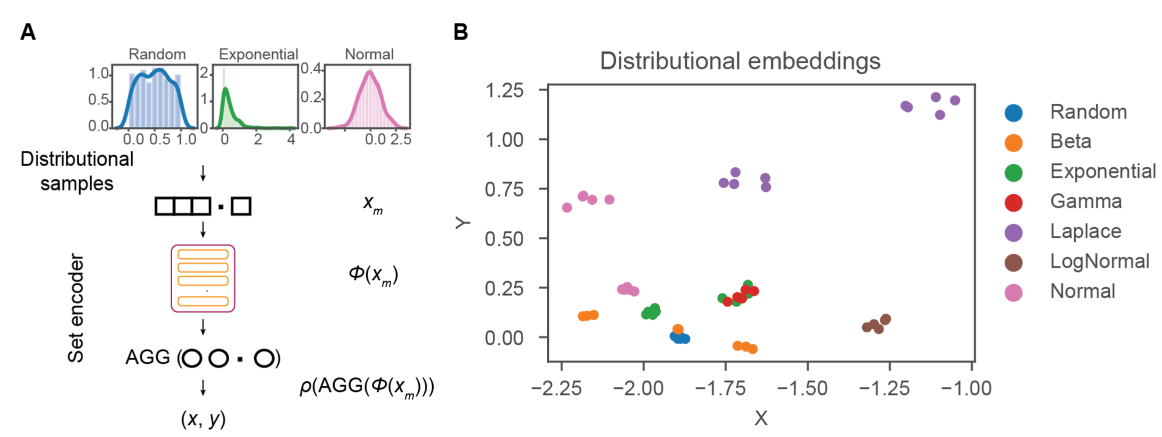

Figure (1) (**A**) Our encoder. (**B**) Low-dimensional embedding of encoded distributions.

Regularizers for preserving properties of Wasserstein space

$$\mathcal{L} := L_{Wass} + \frac{1}{\binom{m}{2}} \sum ((||H_\theta(X+x) - H_\theta(Y+x)|| - ||H_\theta(X) - H_\theta(Y)||)^2 + (||H_\theta(aX) - H_\theta(aY)|| - |a|||H_\theta(X) - H_\theta(Y)||))^2$$

- Regularizers to enforce Translation & Scaling laws.
- Translation law $\Longrightarrow$ $H_\theta(X)$, $H_\theta(X+x)$, $H_\theta(Y)$, $H_\theta(Y+x)$ form vertices of a parallelogram.
- SD_p^λ discretizes the space and changes the metric, thus lose some properties of Wasserstein metric.

Datasets Used

- Samples of size 500 drawn independently 50 times from uniform, Normal, Beta, Gamma, Exponential, Laplace, Log Normal and mixtures of Gaussian distributions with varying parameters.
- Samples of size 300 drawn independently 100 times from $2D$ Gaussians with various μ, Σ.

Experiments with W_1 metric

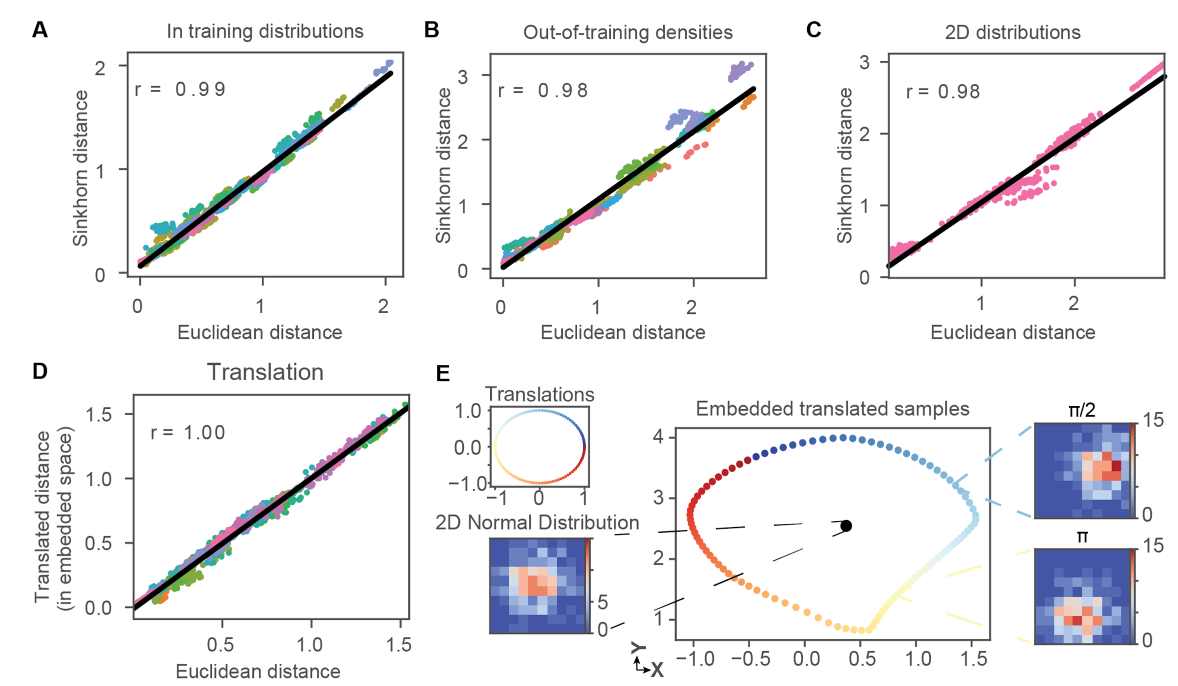

Figure (2) (**A–C**) Pearson's r between embedded and Sinkhorn distances. (**D**) Correlation after translations. (**E**) Samples from a multivariate Gaussian translated around a circular path.

- (Generalizing to out-of-sample-densities) Create new densities from training densities by changing parameters (Fig 2 B,C). Our model generalize to these densities and also calculate distances between 2 Dirac and between 2 Binomial distributions.
- (Translation) (**a**) Given samples X, Y, translate them by a random vector a, (**b**) Samples from a $2D$ Gaussian rotated around a circle (Fig 2 D,E).

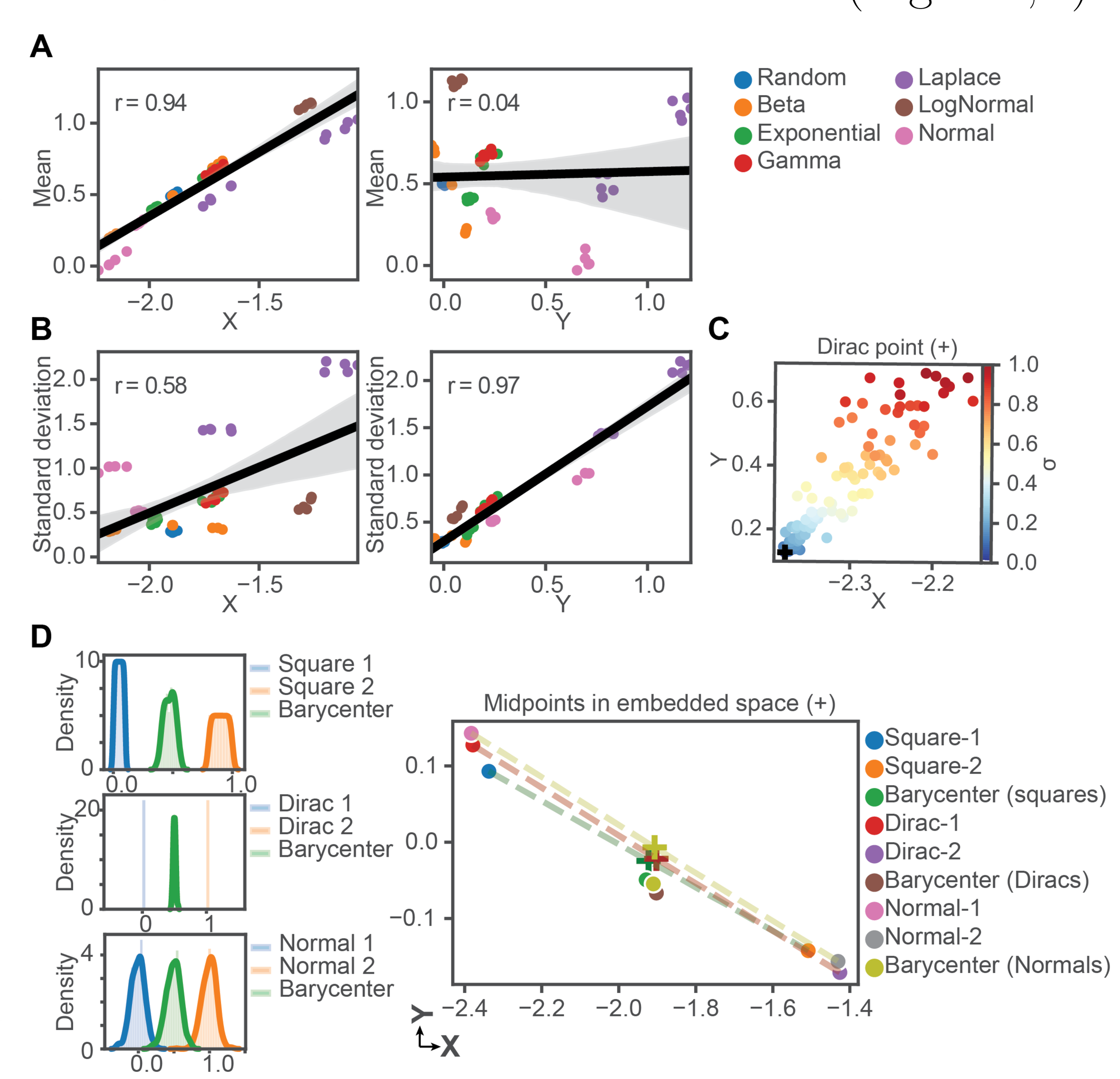

Figure (3) Person's r comparing axes to means (**A**) and standard deviations (**B**). (**C**) Convergence of samples from Gaussians with various standard deviations to the Dirac. (**D**) Barycenters of distributions (left) and midpoints of lines connecting the encoded samples (right).

More Results with W_1 metric

- (Statistical properties of measures) For encoded $1D$-distributions, found strong correlation between means (and variances) and x-coordinate (and y-coordinate) of encoded point(Fig 3A,B).
- (Respecting topology of the space) Choosing samples drawn from $N(0, 1/n)$ see that our encoded points converge to the point encoded by the Dirac measure (Fig 3C).
- (Wasserstein barycenters) Given two densities μ_1, μ_2, and $\hat{\mu}$ their Wasserstein barycenter, show that $H_\theta(\hat{\mu})$ can be approximated by the midpoint of the line joining $H_\theta(\mu_1)$ and $H_\theta(\mu_2)$. (Fig 3D). Distributions used in this experiment are
  1. $N(0, .1)$ and $N(1, .1)$.
  2. Dirac at 0 and 1.
  3. Uniform distribution in $[0, .1]$ and in $[.8, .1]$.
- More experiments found in our paper.

Discussion

- Developed permutation invariant Siamese Network to learn distances between probability measures.
- Showed our model can learn first and second moments.
- Experiments with translations and scaling showed our model respects various properties of Wasserstein space.
- Our model respects the topology by showing the convergence of samples from $N(0, 1/n)$ to Dirac measure at 0.
- Showed Wasserstein geodesics can be approximated by straight lines.
- Results with W_2 metric were weaker than with W_1 metric. We conjecture the reason behind this are :
  1. $\mathbb{P}(\mathcal{X})$ under W_1 is flat but not under W_2. Sinkhorn distance changes the W_2 metric differently than it changes the space under W_1
  2. Our target is a flat Euclidean space thus losing more structural information when mapping from the 2-Wasserstein space.
- In future work, we want to learn continuity properties of these neural networks and investigate if they can learn higher moments.