# Autumn 2021 Optimization & Machine Learning Talk I

## Computing the Distance Between Probability Measures: Wasserstein vs. Fisher-Rao

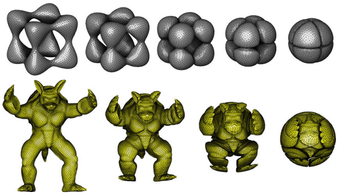
Axel G. R. Turnquist

NJIT Department of Mathematical Sciences

September 1, 2021

# Applications

- Comparing images
- Finding closeness of data fit
- Optimization algorithms
- Shape analysis
- Wasserstein GAN



# Why $L^2$ is Inappropriate

Distance between $\mu = \text{Unif}[0,1]$ and $\nu = \text{Unif}[2,3]$ vs.
$\mu = \text{Unif}[0,1]$ and $\text{Unif}[a, a+1]$.

$L^2$-distance is the same for both cases. $L^2$ only measures vertical distance, does not take in any *horizontal distance* into account.

# First Real Analysis Idea: Total Variation

Maybe we can compute

$$\mathrm{dist}(\mu, \nu) = \|\mu - \nu\|_{\mathrm{TV}} \tag{1}$$

In 1D, the total variation of a real-valued function $f$ on an interval $[a, b]$ is:

$$\|f\|_{\mathrm{TV}[a,b]} = \sup_{\mathcal{P}} \sum_{i=0}^{n_P} |f(x_{i+1}) - f(x_i)| \tag{2}$$

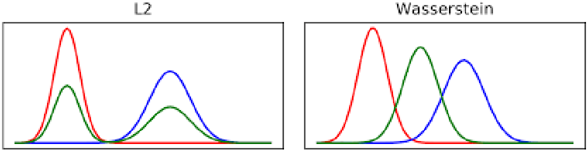
and one can see how the horizontal distance is taken into account.

# Total Variation Continued

For probability measures $\mu$ and $\nu$, the total variation between them is defined as:

$$\|\mu - \nu\|_{\mathsf{TV}} = 2 \sup \{|\mu(A) - \nu(A)| : A \in \Sigma\} \tag{3}$$

**Goal**: want a notion of **interpolation**. Need a manifold (metric) structure.



# Ideal Interpolation

# Metric vs. Distance

On a manifold, what is a **metric**? For finite-dimensional manifolds, a metric is like a positive definite matrix $A > 0$. It tells us what a dot product and magnitude are:

$$u \cdot v = u^T A v \tag{4}$$

$$\|u\|_A = \sqrt{u \cdot u} = \sqrt{u^T A u} \tag{5}$$

which defines the speed of a path $\gamma(t)$

$$\left\| \gamma(\dot{t}) \right\|_A = \sqrt{\dot{\gamma}^T A \dot{\gamma}} \tag{6}$$

This then gives us a notion of **distance** between $\gamma(0) = a$ and $\gamma(1) = b$:

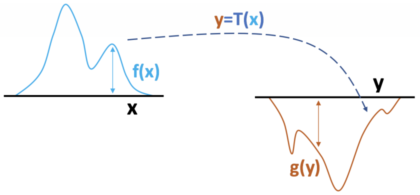
$$\operatorname{dist}(\gamma(0), \gamma(1)) := \inf_{\gamma : \gamma(0) = a, \gamma(1) = b} \int_0^1 |\dot{\gamma}(s)| \, ds \tag{7}$$

# Wasserstein Distance

The optimal transport problem uses the **change of variables** formula from Calculus. That is, given two probability measures $\mu$ and $\nu$ and a diffeomorphic mapping $T$, such that $T_{\#}\mu = \nu$:

$$\int_A \mu(x) = \int_A \nu(T(x))J_T(x) \tag{8}$$

for all measurable $A \subset \Omega$ where $J$ designates the Jacobian of the mapping $T$.



## Wasserstein Distance

Now we compute the *horizontal distance* by finding the map $T$ that takes the least amount of work to move from a point $x$ to $y$, like "shoveling dirt":

$$\mathrm{dist}(\mu, \nu) = \inf_T \int_{\mathbb{R}^n} c(x, T) d\mu(x) \tag{9}$$

such that $T$ satisfies the change of variables formula:

$$\int_A \mu(x) = \int_A \nu(T(x)) J_T(x) \tag{10}$$

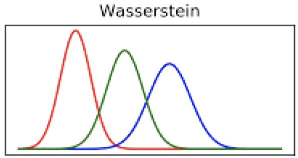
This represents a calculus of variations problem with an equality constraint. We get a distance by choosing $c(x, y) = d_M(x, y)^2$.

# Wasserstein Interpolation

This Wasserstein distance is actually way better than it initially seems. Defines a manifold structure on the space of probability measures on $M$! This gives us a **metric**, notion of Calculus, and **interpolation**, etc. Let $T_t = tT + (1-t)\mathrm{Id}$. Then,

$$\mu_t = T_{t\#}\mu \tag{11}$$

defines the Wasserstein interpolation and $\mu_1 = \nu$. The interpolation of two Gaussians is a Gaussian.



# First Statistical Idea: KL Divergence

The Kullback-Liebler **divergence** is also known as the **relative entropy**. The KL divergence $D_{\mathrm{KL}}(\mu, \nu)$ represents the "surprise" one receives in observing samples from $\nu$ when one actually expects samples from $\mu$:

$$D_{\mathrm{KL}}(\mu, \nu) = \int_{\Omega} \log\left(\frac{d\mu}{d\nu}\right) d\mu \tag{12}$$

**Note**: *not a distance*! If we can parametrize $\mu, \nu$ by a parameter $\theta$, then the Hessian of the KL divergence is known as the **Fisher information metric**:

$$g_{jk}(\theta) = \int_X \frac{\partial \log p(x,\theta)}{\partial \theta_j} \frac{\partial \log p(x,\theta)}{\partial \theta_k} p(x,\theta) dx \tag{13}$$

# Fisher-Rao

This Fisher information metric defines a **Riemannian structure** (means you can do Calculus) on the infinite-dimensional manifold of probability distributions $\mathcal{P}(M)$. Using an *explicit formula for the geodesics* on the Fisher-Rao manifold, one gets the **Fisher-Rao distance**:

$$\mathrm{dist}(\mu, \nu) = \sqrt{\mathrm{vol}(M)} \arccos\left(\frac{1}{\mathrm{vol}(M)} \int_M \sqrt{\frac{\mu}{\mathrm{vol}} \frac{\nu}{\mathrm{vol}}} \mathrm{vol}\right) \tag{14}$$

In general, deriving a formula for the Fisher-Rao distance is difficult (e.g. $M = \mathbb{R}^d$).

# We Can Get Bounds!

Happily, we can relate these with inequalities for $\mu, \nu \in \text{Dens}(M)$, the space of smooth densities strictly bounded away from zero on a compact manifold $M$:

$$\frac{\text{dist}_W(\mu,\nu)}{\text{diam}(M)} \leq \text{dist}_{\text{FR}}(\mu,\nu) \tag{15}$$

$$\text{dist}_{\text{TV}}(\mu,\nu) \leq \text{dist}_{\text{FR}}(\mu,\nu) \tag{16}$$

$$\text{dist}_{\text{FR}}(\mu,\nu) \leq \sqrt{\frac{\pi}{2}\text{dist}_{\text{KL}}(\mu,\nu)} \tag{17}$$

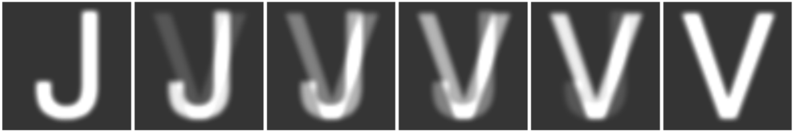
# Discussion

**Wasserstein**:

- In general result is not a mapping $T$ but a **joint probability distribution** $\pi$
- Hard to compute in higher-dimensions (**curse of dimensionality**)
- Regularity issues on some manifolds $M$, but behaves well for $\mathbb{R}^d$

**Fisher-Rao**:

- Naturally used in **information geometry**
- Not good for image interpolation



Source Target

# Questions?

# Highlighted Resources

- "Optimal Transport: Old and New" Cedric Villani
- "Topics in Optimal Transport" Cedric Villani
- "Diffeomorphic density matching by optimal information transport" Martin Bauer, Sarang Joshi & Klas Modin
- "On Choosing and Bounding Probability Metrics" Alison Gibbs & Francis Su
- "Computational Optimal Transport" , Gabriel Peyré & Marco Cuturi

# Future Talks

**Next Talk**:

September 9: Binan Gu
“Discrete Optimal Control on Graphs”