Numerical Optimal Transport

http://optimaltransport.github.io

Density Fitting

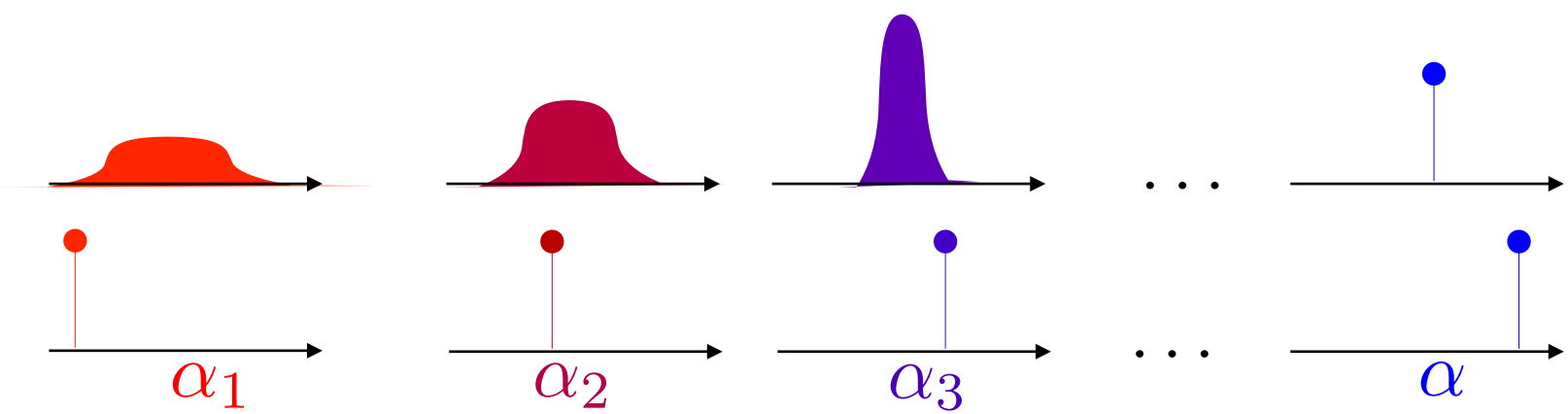
Gabriel Peyré

www.numerical-tours.com

Weak vs Strong Topology

Random vectors		Radon measures
$\mathbb{P}(X \in A)$	$=$	$\int_A \mathrm{d}\alpha(x)$
Convergence in law:		Weak* convergence:
$\forall$ set A $\quad \mathbb{P}(X_n \in A) \overset{n\to+\infty}{\longrightarrow} \mathbb{P}(X \in A)$		$\forall$ continuous function f $\quad \int f \mathrm{d}\alpha_n \overset{n\to+\infty}{\longrightarrow} \int f \mathrm{d}\alpha$

Weak convergence:

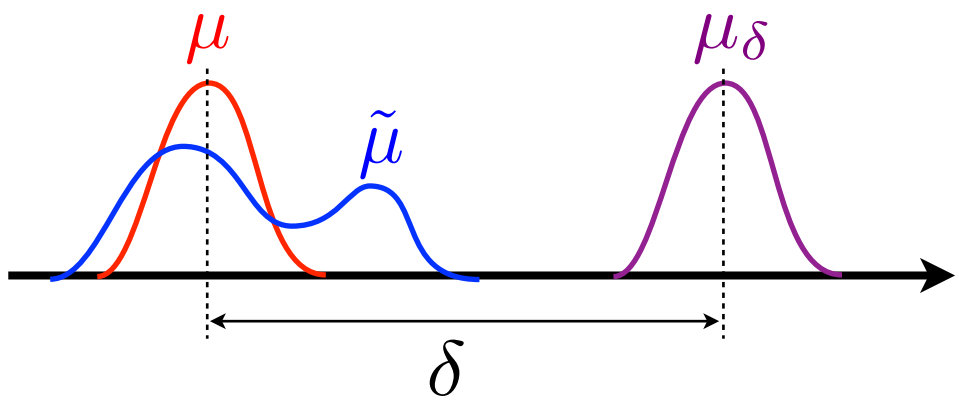


Overview

- **Csiszar Divergences**
- Dual Norms and MMD
- Minimum Kantorovitch Estimators
- Deep Generative Models Fitting

Metrics on the Space of Measures

$$\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$$
$$\mathrm{d}\tilde{\mu}(x) = \tilde{\rho}(x)\mathrm{d}x$$
$$\mathrm{d}\mu_\delta(x) = \rho(x-\delta)\mathrm{d}x$$



Bins-to-bins metrics:

Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(\mu,\tilde{\mu}) = \int \rho(x)\log\frac{\rho(x)}{\tilde{\rho}(x)}\mathrm{d}x$$

Hellinger distance:

$$D_{\mathrm{H}}(\mu,\tilde{\mu})^2 = \int \left(\sqrt{\rho(x)} - \sqrt{\tilde{\rho}(x)}\right)^2 \mathrm{d}x$$

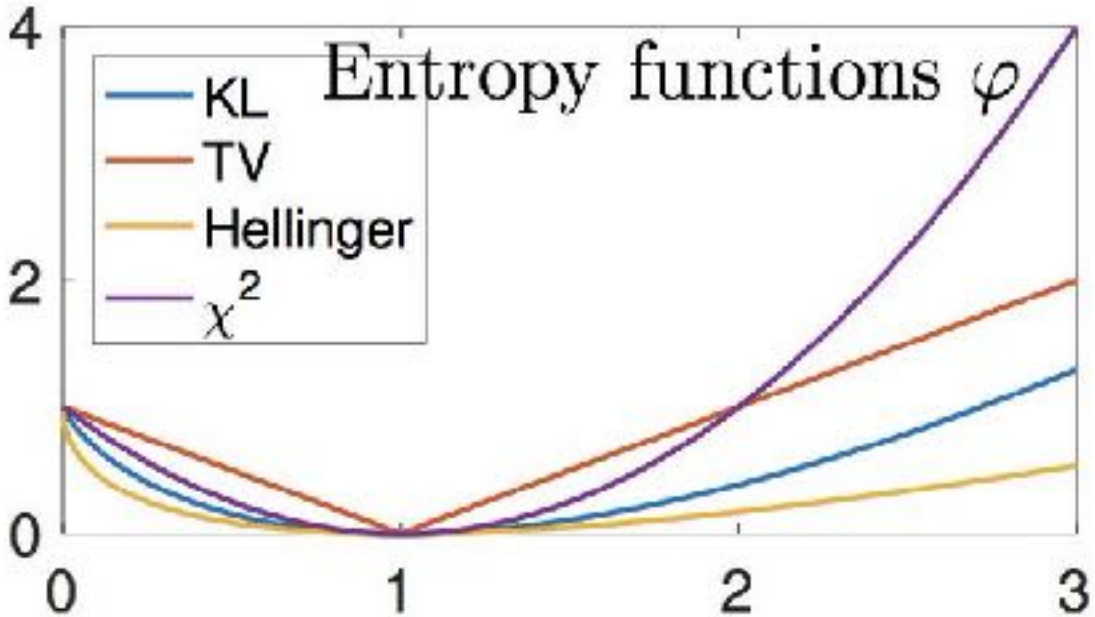
Effect of translation:

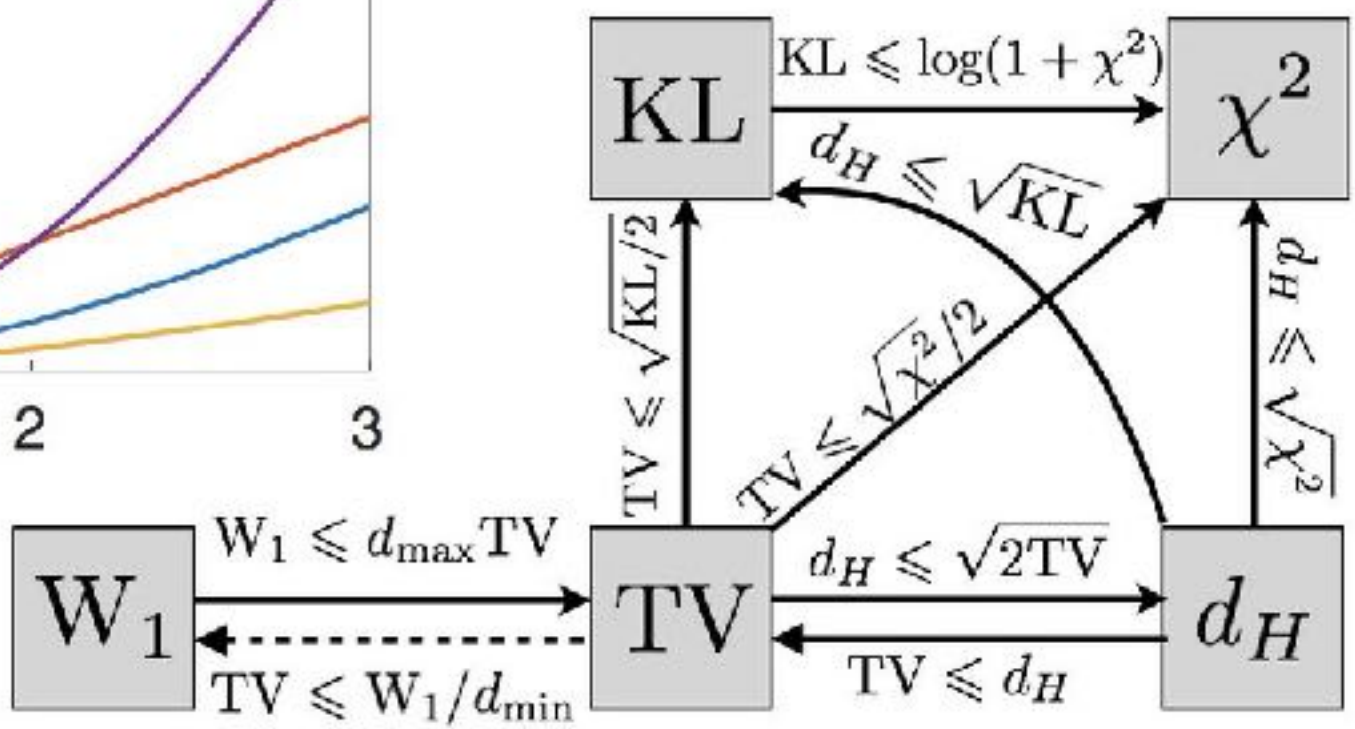
$$D(\mu,\mu_\delta) \approx \mathrm{cst} \qquad W_2(\mu,\mu_\delta) = \delta$$

Csiszar Divergence

$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta + \varphi'_\infty \alpha^\perp(\mathcal{X})$$

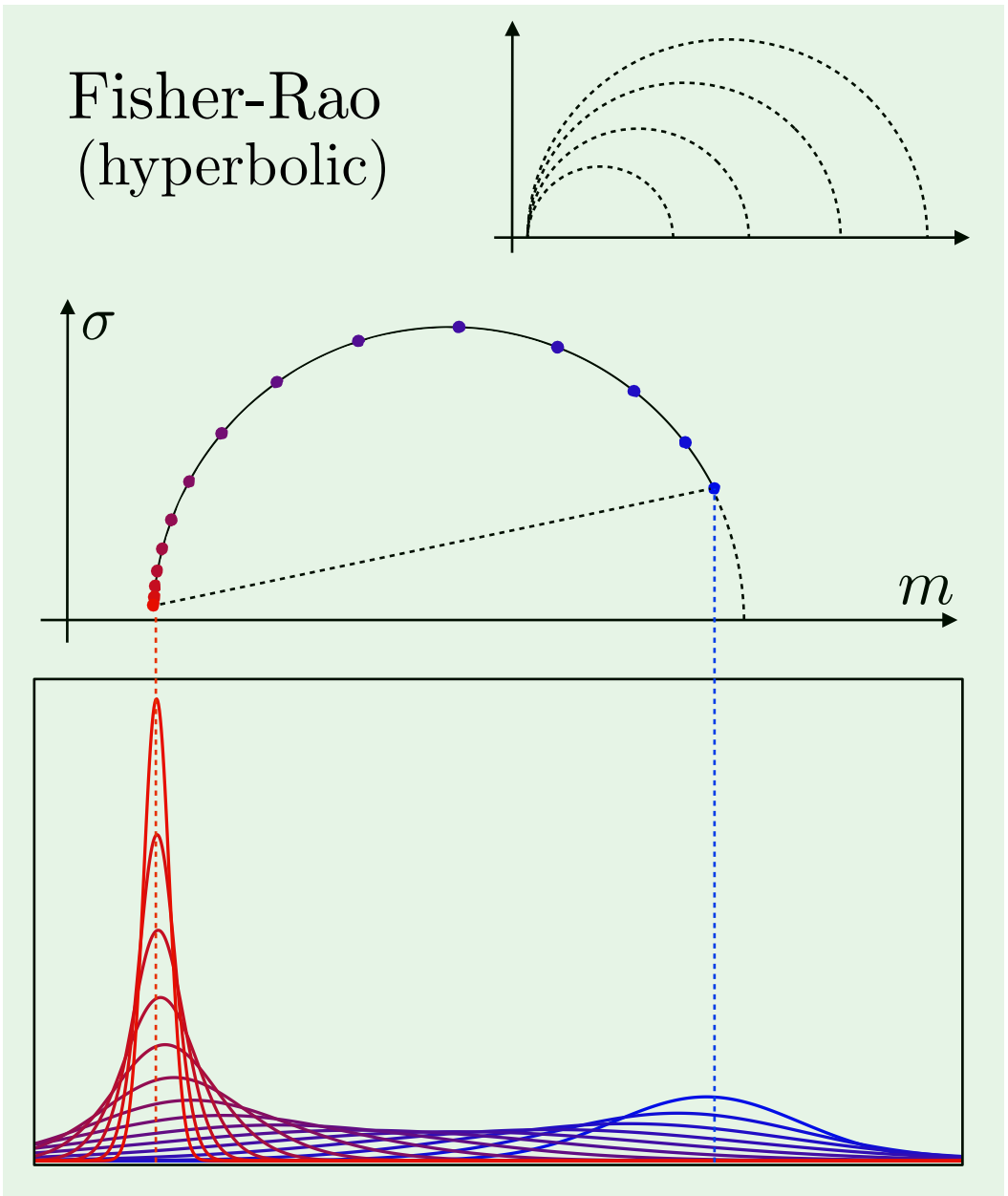
$$\varphi'_\infty = \lim_{x \uparrow +\infty} \varphi(x)/x \in \mathbb{R} \cup \{\infty\}$$

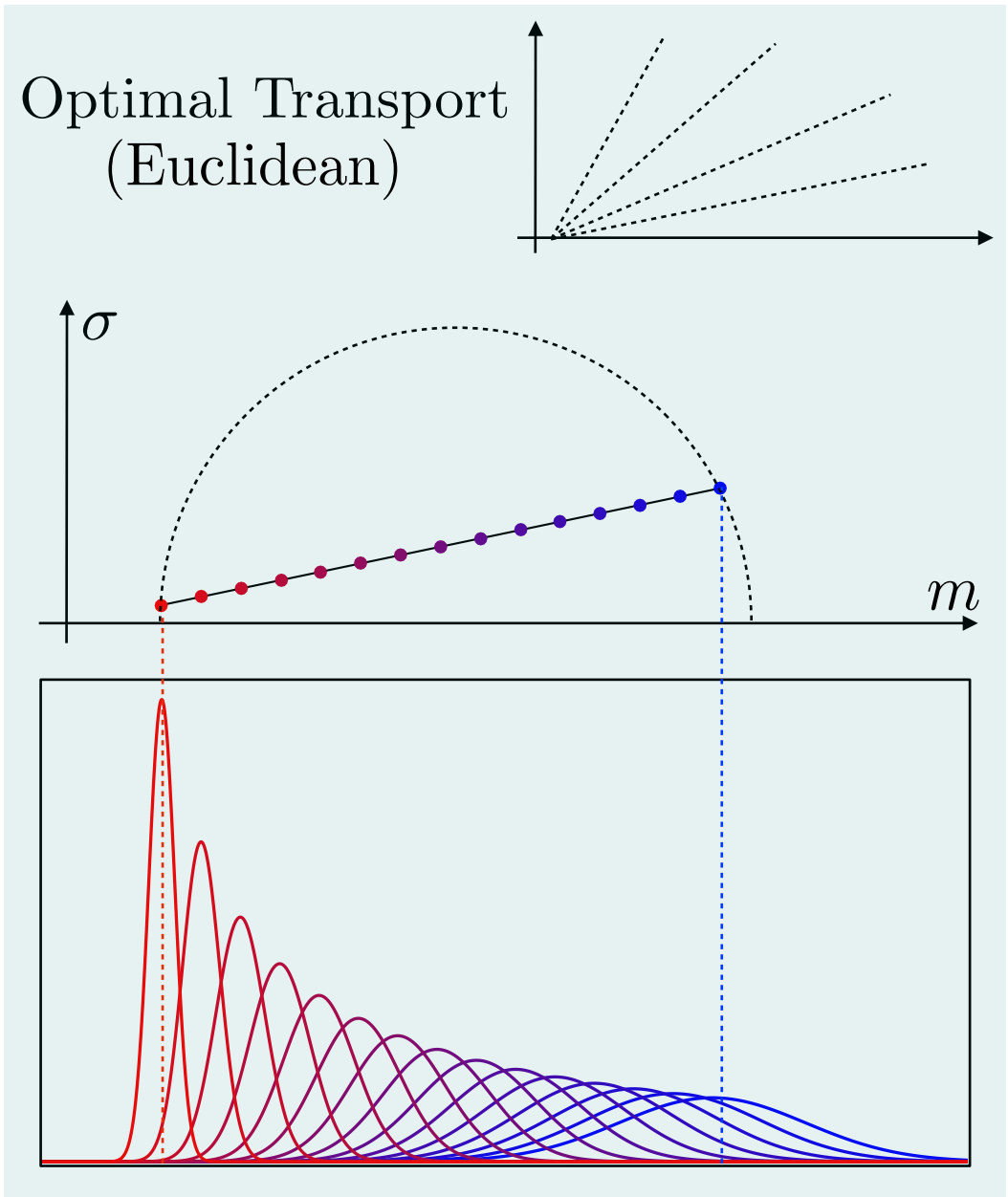




Csiszár divergences, a unifying way to define losses between arbitrary positive measures (discrete & densities). https://en.wikipedia.org/wiki/F-divergence …

OT vs. KL (Fisher-Rao)





Overview

- Csiszar Divergences
- **Dual Norms and MMD**
- Minimum Kantorovitch Estimators
- Deep Generative Models Fitting

Dual Norms

Dual norms: (aka Integral Probability Metrics)

$$\|\alpha - \beta\|_B \stackrel{\text{def.}}{=} \max \left\{ \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x)) \; ; \; f \in B \right\}$$

Wasserstein 1: $B = \{f \; ; \; \|\nabla f\|_\infty \leqslant 1\}$.

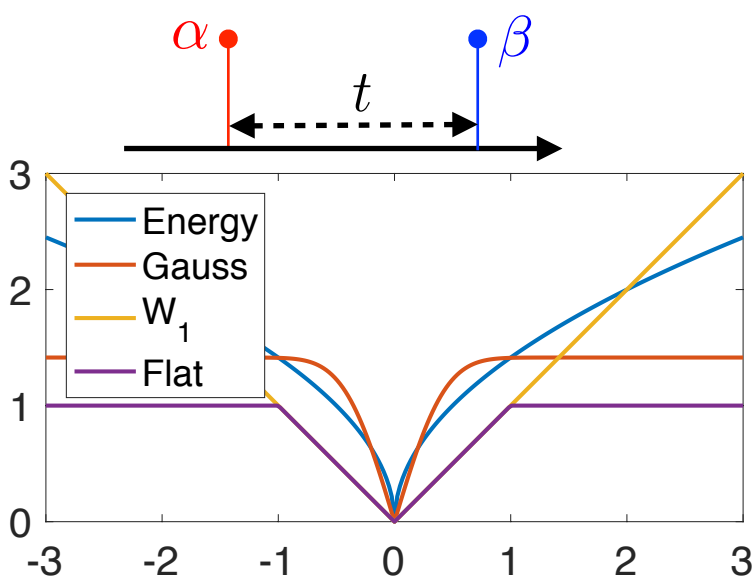
Flat norm: $B = \{f \; ; \; \|f\|_\infty \leqslant 1, \|\nabla f\|_\infty \leqslant 1\}$.

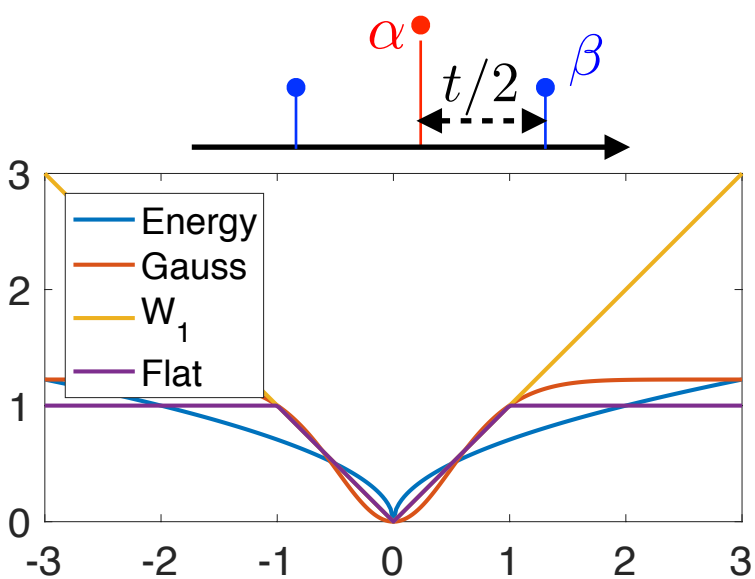
RKHS: $B = \left\{f \; ; \; \|f\|_k^2 \leqslant 1\right\}$.

$$\|\alpha - \beta\|_B^2 = \int k(x,x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int k(x,x')\mathrm{d}\beta(y)\mathrm{d}\beta(y') - 2\int k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$$

Energy distance: $k(x,y) = -\|x-y\|$

Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$

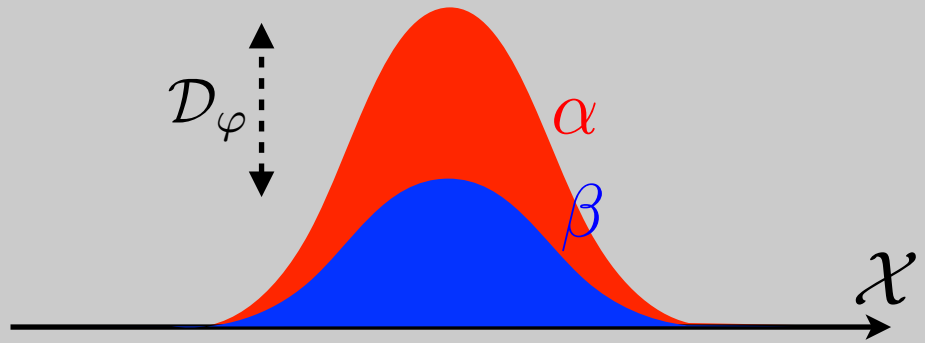




Csiszar Divergence vs Dual Norms

Csiszár divergences:

$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta$$

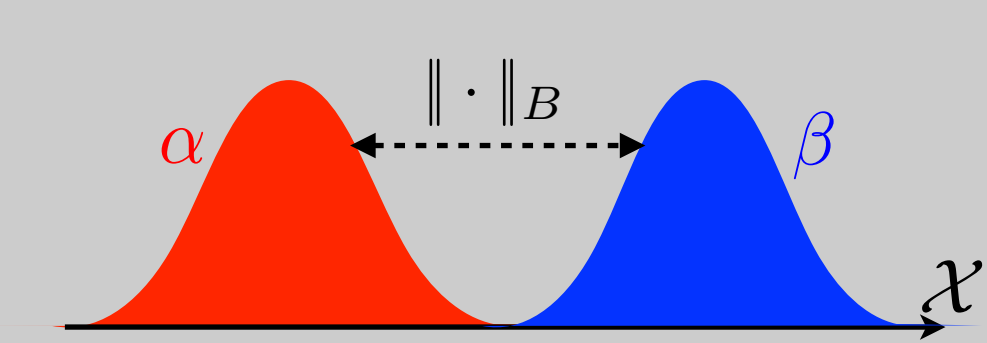


Strong topology

$\longrightarrow$ KL, TV, χ^2, Hellinger ...

Dual norms:

$$\|\alpha - \beta\|_B \overset{\text{def.}}{=} \max_{f \in B} \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x))$$



Weak topology

$\longrightarrow$ W_1, flat, RKHS*, energy dist, ...

RKHS Norms aka Maximum Mean Discrepancy

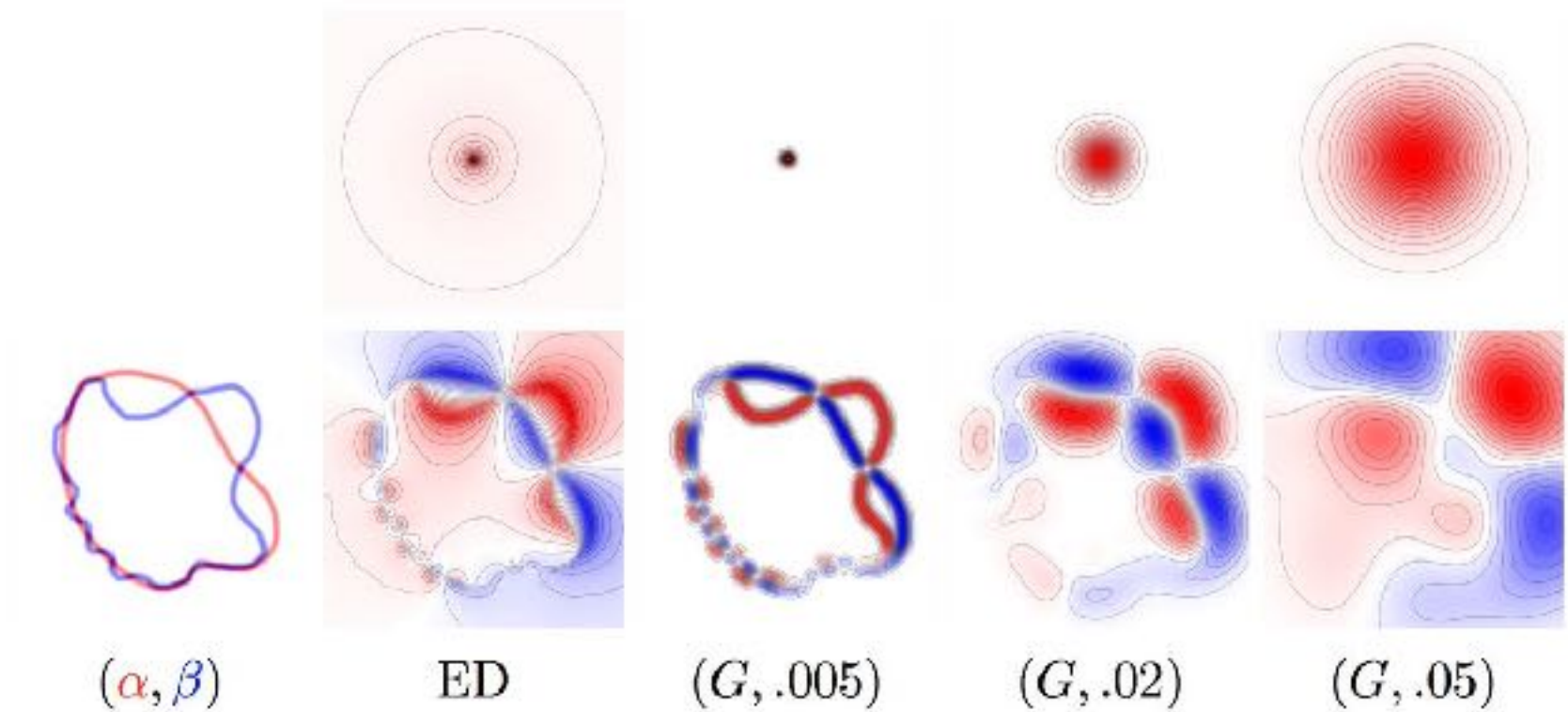


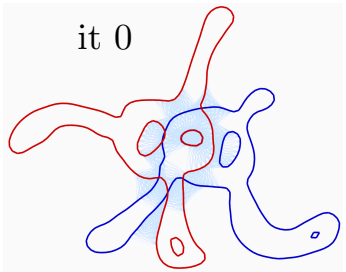
Figure 8.4: Top row: display of ψ such that $\|\alpha - \beta\|_k = \|\psi \star (\alpha - \beta)\|_{L^2(\mathbb{R}^2)}$, formally defined over Fourier as $\hat{\psi}(\omega) = \sqrt{\hat{\varphi}(\omega)}$ where $k^*(x, x') = \varphi(x - x')$. Bottom row: display of $\psi \star (\alpha - \beta)$. (G,$\sigma$) stands for Gaussian kernel of variance σ^2 and ED for Energy Distance kernel (in which case $\psi(x) = 1/\sqrt{\|x\|}$).

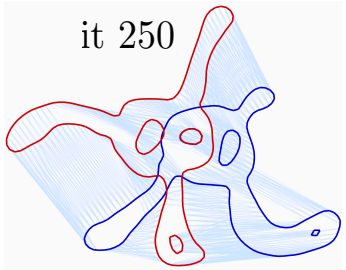
OT Loss for Diffeomorphic Registration

Joint work with J. Feydy, B. Charier, F-X. Vialard.

Shape registration: $\min_{\varphi \text{ diffeo}} D(\varphi(\mu), \nu) + R(\varphi)$

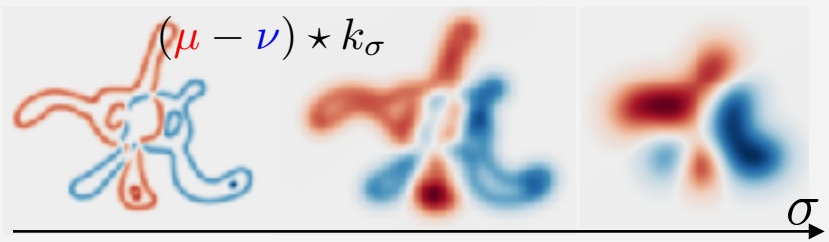
loss regularity





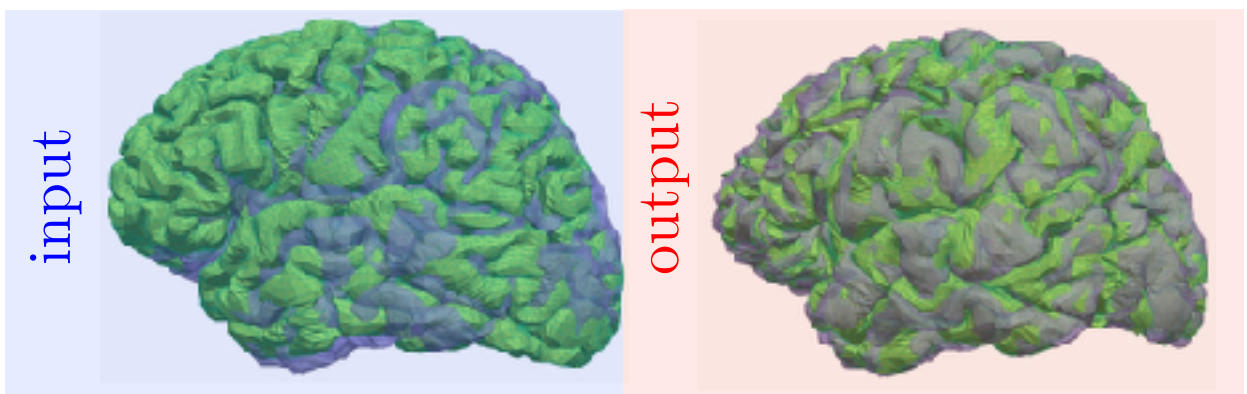
Hilbertian loss (MMD/RKHS):

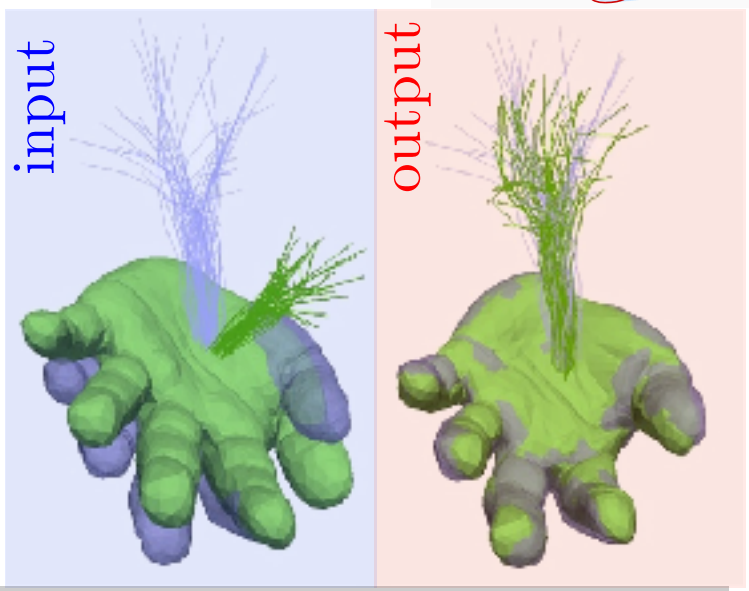
$$D(\mu, \nu) = \|k_\sigma \star (\mu - \nu)\|_{L^2}^2$$



Sinkhorn divergence:

$$D(\mu, \nu) = \bar{W}_\varepsilon(\mu, \nu)$$





→ Do not use OT for registration . . . but as a loss.

→ Sinkhorn's iterates "propagate" a small bandwidth kernel.

→ Automatic differentation: game changer for advanced loss and models.

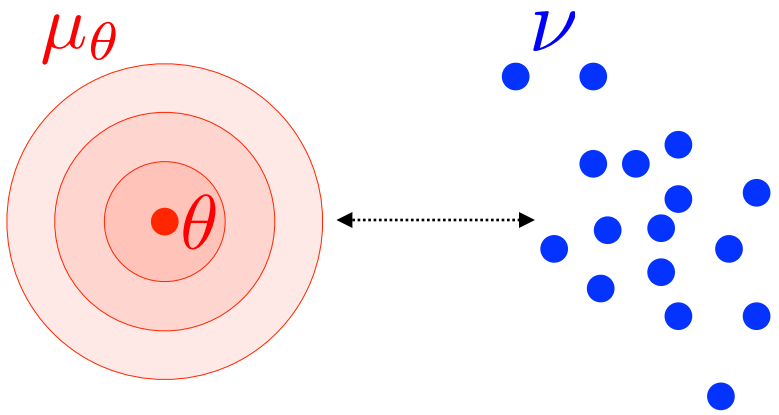
Overview

- Csiszar Divergences
- Dual Norms and MMD
- **Minimum Kantorovitch Estimators**
- Deep Generative Models Fitting

Density Fitting and Generative Models

Observations: $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Parametric model: $\theta \mapsto \mu_\theta$



Density fitting: $\mathrm{d}\mu_\theta(y) = f_\theta(y)\mathrm{d}y$

$$\min_\theta \widehat{\mathrm{KL}}(\nu|\mu_\theta) \overset{\text{def.}}{=} -\sum_j \log(f_\theta(y_j))$$

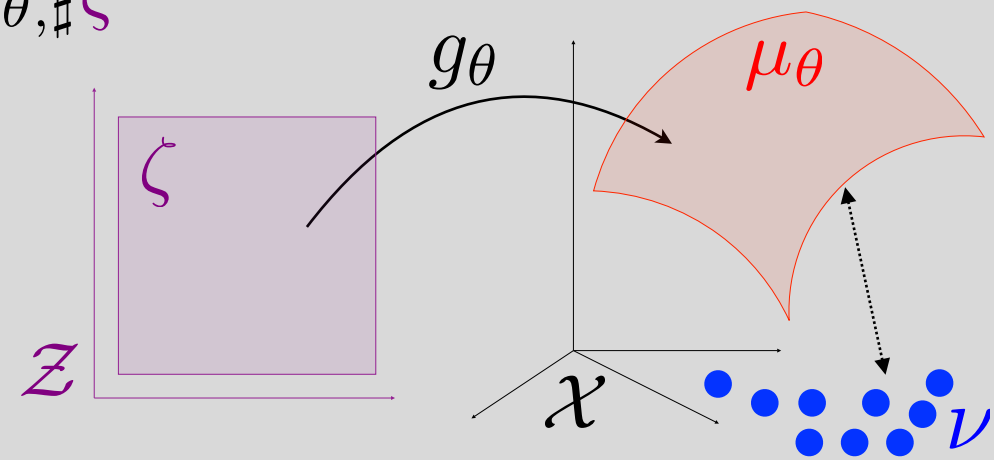
Maximum likelihood (MLE)

Generative model fit: $\mu_\theta = g_{\theta,\sharp}\zeta$

$$\widehat{\mathrm{KL}}(\nu|\mu_\theta) = +\infty$$

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

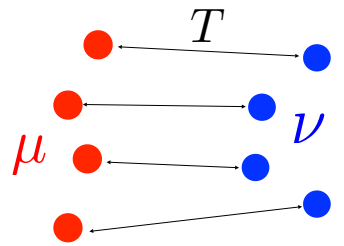


Loss Functions for Measures

Density fitting: $\min_\theta D(\mu_\theta, \nu)$ $\quad \nu = \frac{1}{P}\sum_j \delta_{y_j}$ $\quad \mu = \frac{1}{N}\sum_i \delta_{x_i}$

Optimal Transport Distances

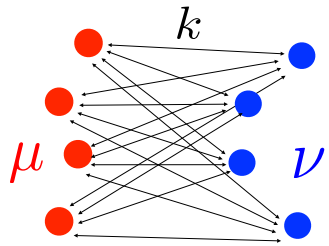
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j}\|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

$$\|\mu-\nu\|_k^2 \overset{\text{def.}}{=} \frac{1}{N^2}\sum_{i,i'} k(x_i,x_{i'}) + \frac{1}{P^2}\sum_{j,j'} k(y_j,y_{j'}) - \frac{2}{NP}\sum_{i,j} k(x_i,y_j)$$

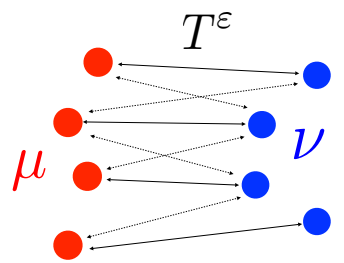
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Cuturi 13]

$$W_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} \sum_{i,j} T^\varepsilon_{i,j}\|x_i - y_j\|^p$$

$$\bar{W}_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} W_\varepsilon(\mu,\nu)^p - \tfrac{1}{2}W_\varepsilon(\mu,\mu)^p - \tfrac{1}{2}W_\varepsilon(\nu,\nu)^p$$



Theorem: [Ramdas, G.Trillos, Cuturi 17] $\quad \bar{W}_\varepsilon(\mu,\nu)^p \underset{\varepsilon\to+\infty}{\overset{\varepsilon\to 0}{\longrightarrow}} \begin{matrix} W(\mu,\nu)^p \\ \|\mu-\nu\|_k^2 \end{matrix}$ for $k(x,y) = -\|x-y\|^p$

Best of both worlds:

$\rightarrow$ cross-validate ε

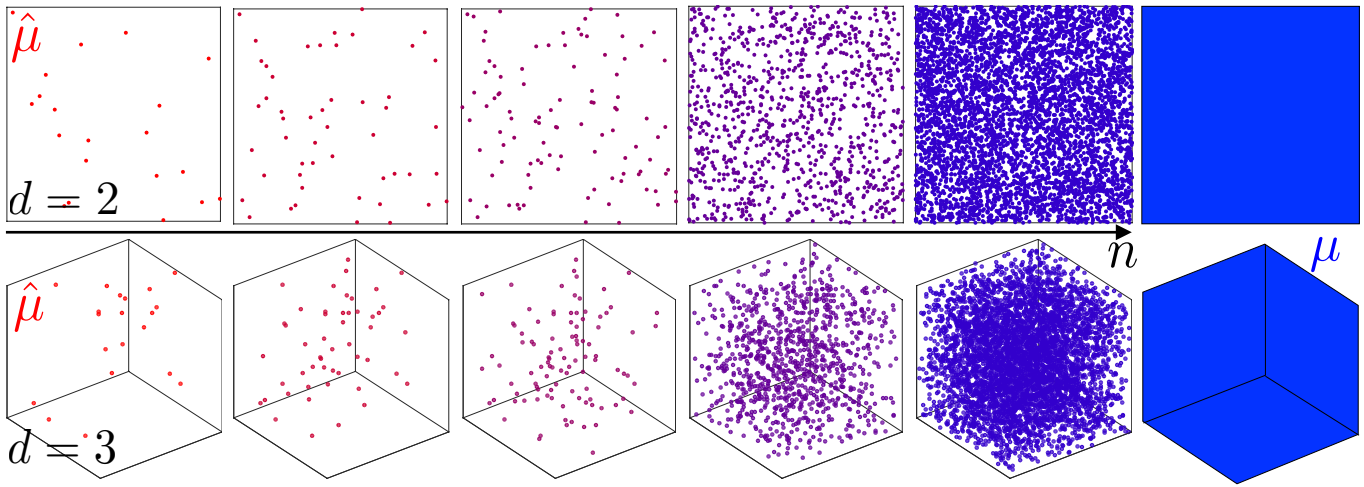
– Scale free (no σ, no heavy tail kernel).

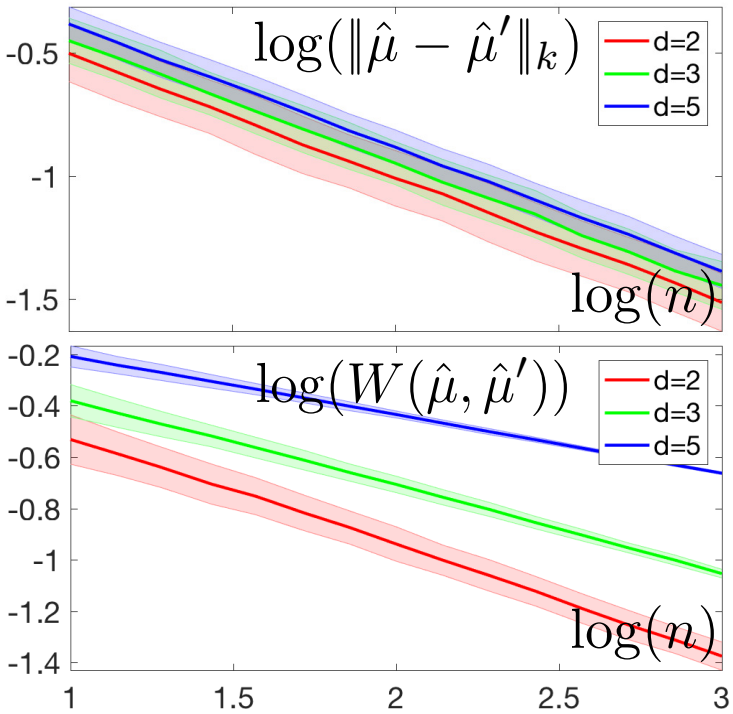
– Non-Euclidean, arbitrary ground distance.

– Less biased gradient.

– No curse of dimension (low sample complexity).

Sample Complexity





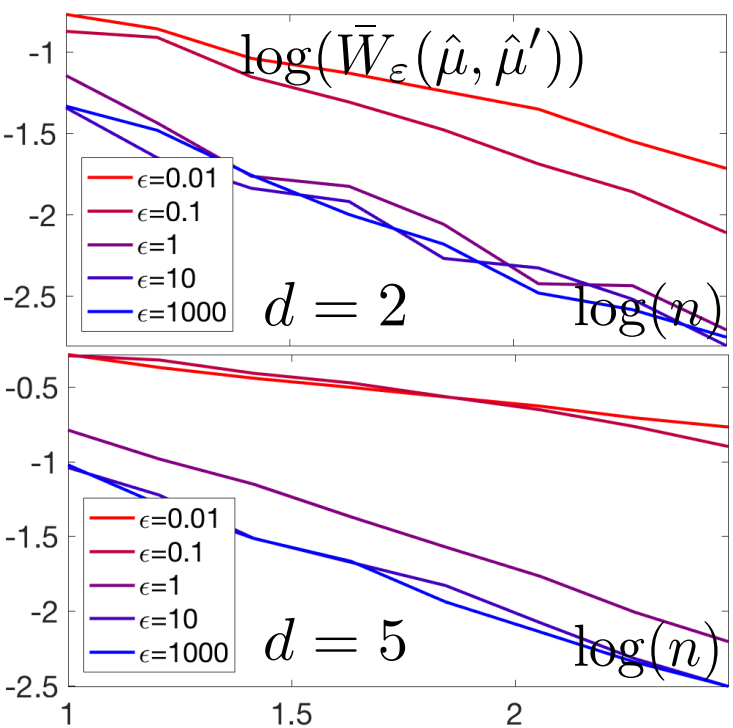
Theorem: $\mathbb{E}(|W(\hat{\mu}, \hat{\nu}) - W(\mu, \nu)|) = O(n^{-\frac{1}{d}})$

$\mathbb{E}(|\|\hat{\mu} - \hat{\nu}\|_k - \|\mu - \nu\|_k|) = O(n^{-\frac{1}{2}})$

Optimal transport: suffers from curse of dimensionality.

$\rightarrow$ Adapt to support dimensionality [Weed, Bach 2017]

Open problem: sample complexity of $\bar{W}_\varepsilon$?



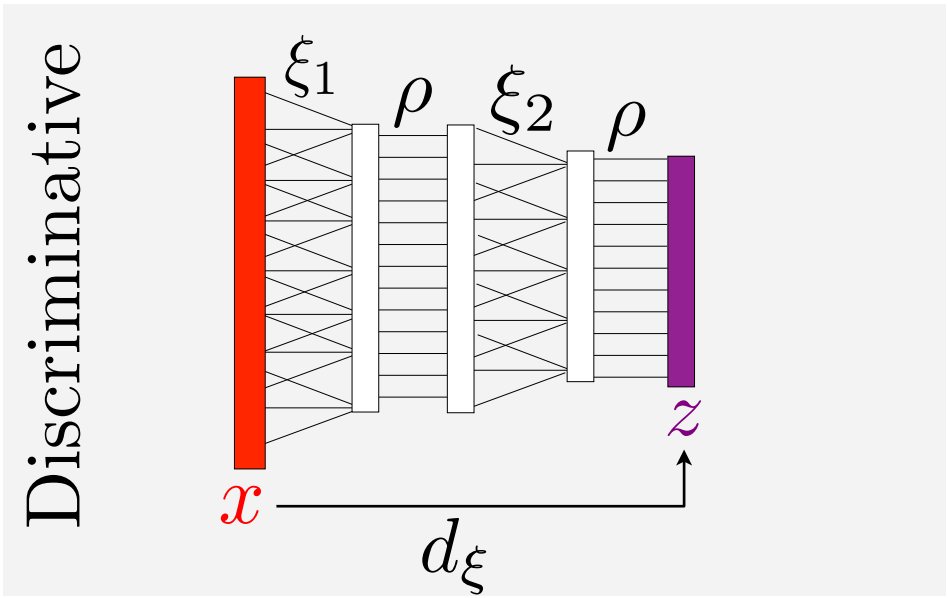
Overview

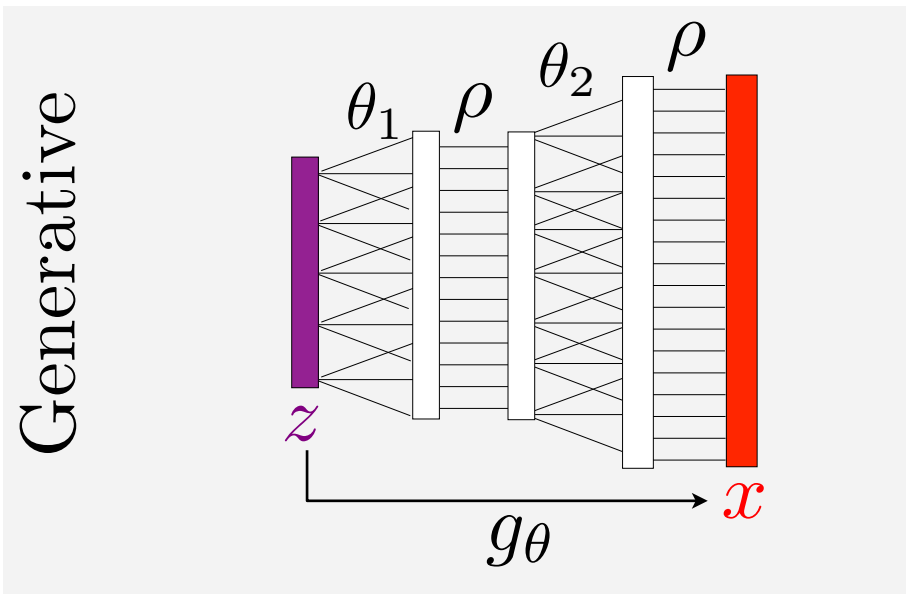
- Csiszar Divergences
- Dual Norms and MMD
- Minimum Kantorovitch Estimators
- **Deep Generative Models Fitting**

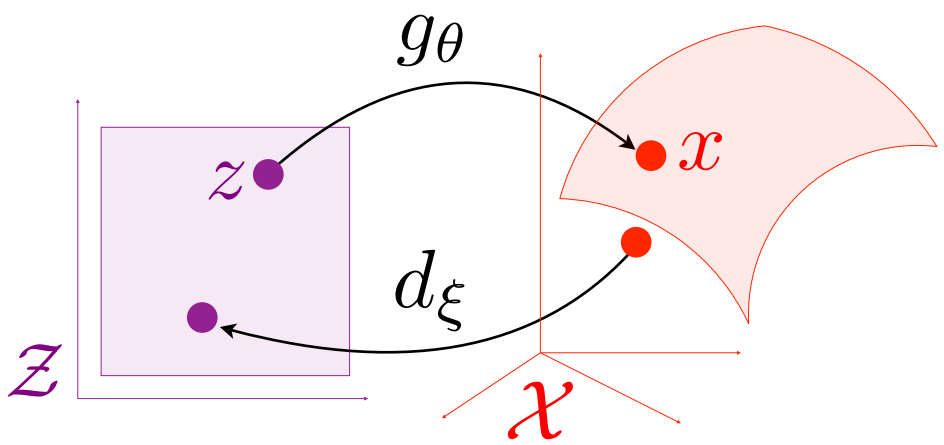
Deep Discriminative vs Generative Models

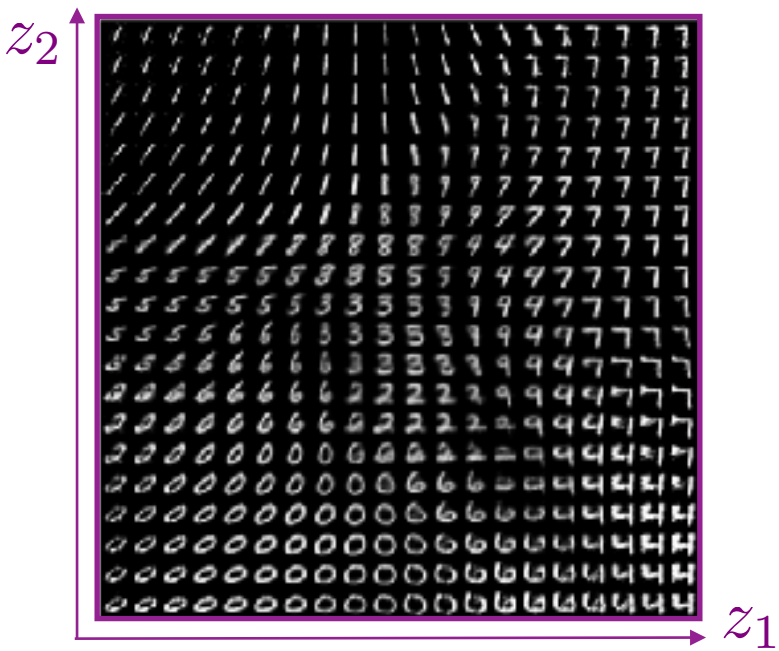
Deep networks:

$$d_\xi(x) = \rho(\xi_K(\ldots\rho(\xi_2(\rho(\xi_1(x)\ldots)$$
$$g_\theta(z) = \rho(\theta_K(\ldots\rho(\theta_2(\rho(\theta_1(z)\ldots)$$









Examples of Image Generation

Inputs

Small ε

Large ε

$\rightarrow$ Need to learn the metric $c(x, y) = \|d_\xi(x) - d_\xi(y)\|^p$ ($\sim$ GANs)

$\rightarrow$ Performance evaluation of generative models is an open problem.

Progressive Growing of GANs for Improved Quality, Stability, and Variation
Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, ICLR 2018