Numerical Optimal Transport

http://optimaltransport.github.io

Density Fitting

Gabriel Peyré

www.numerical-tours.com

ÉCOLE NORMALE SUPÉRIEURE

Weak vs Strong Topology

Random vectors

$$\mathbb{P}(X \in A)$$

Convergence in law:

$\forall$ set A

$$\mathbb{P}(X_n \in A) \xrightarrow{n \to +\infty} \mathbb{P}(X \in A)$$

$=$

Radon measures

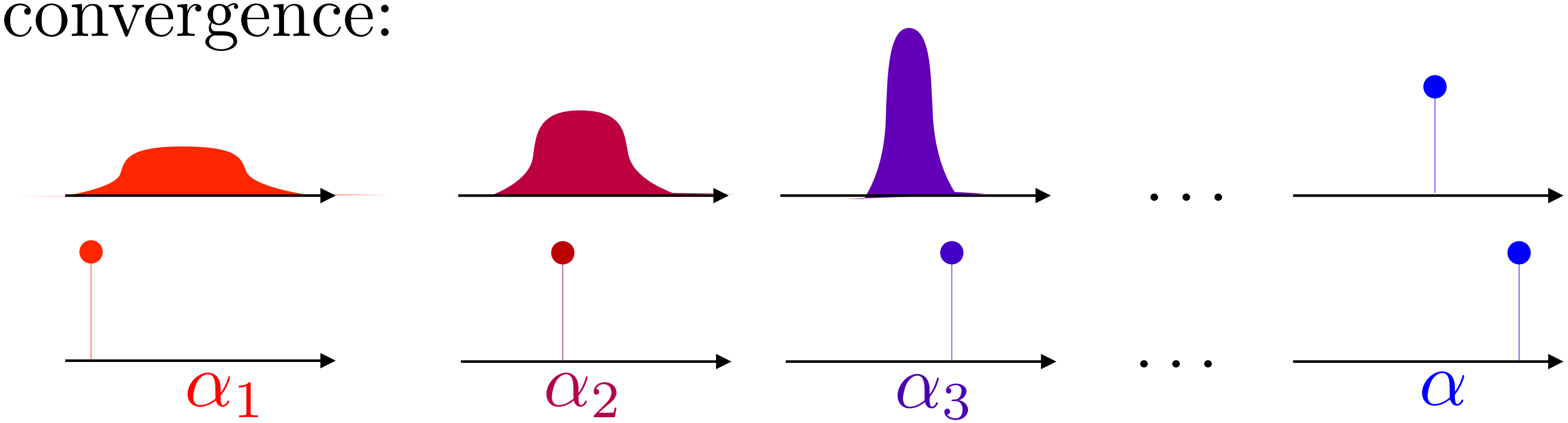
$$\int_A \mathrm{d}\alpha(x)$$

Weak* convergence:

$\forall$ continuous function f

$$\int f \mathrm{d}\alpha_n \xrightarrow{n \to +\infty} \int f \mathrm{d}\alpha$$

Weak convergence:



Key question: quantifying weak convergence.

Central Limit Theorem

Central limit theorem: If $\mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1$ and $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$

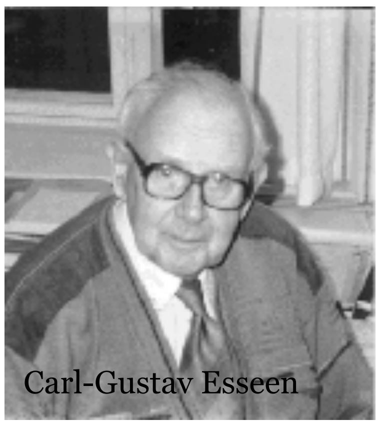
$$Y_n \overset{\text{def.}}{=} \frac{X_1 + \ldots + X_n}{\sqrt{n}} \xrightarrow{\text{law}} \mathcal{N}(0,1)$$

Central Limit Theorem

Central limit theorem: If $\mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1$ and $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$

$$Y_n \overset{\text{def.}}{=} \frac{X_1 + \ldots + X_n}{\sqrt{n}} \xrightarrow{\text{law}} \mathcal{N}(0,1)$$

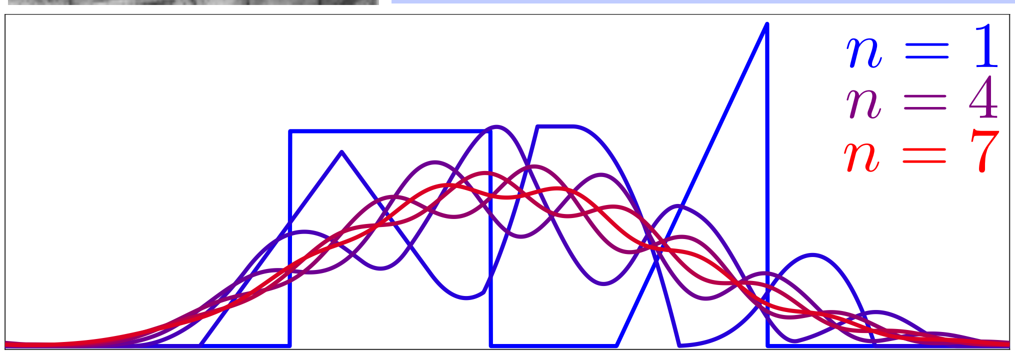
Kolmogorov-Smirnov distance: $d_{KS}(X,Y) \overset{\text{def.}}{=} \max_t |\mathbb{P}(X \leqslant t) - \mathbb{P}(Y \leqslant t)|$

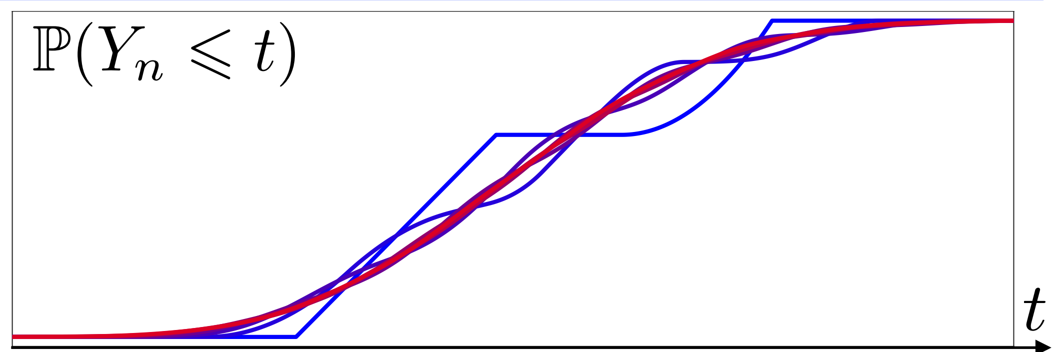


Metrizes convergence in law: $X \overset{\text{law}}{\to} Y \Leftrightarrow d_{\text{KS}}(X,Y) \to 0$

Theorem:
[Berry 1941]
[Esseen, 1942]

$$d_{\text{KS}}(Y_n, \mathcal{N}(0,1)) \leqslant \frac{C\mathbb{E}(|X|^3)}{\sqrt{n}} \qquad C \leqslant 1/2$$



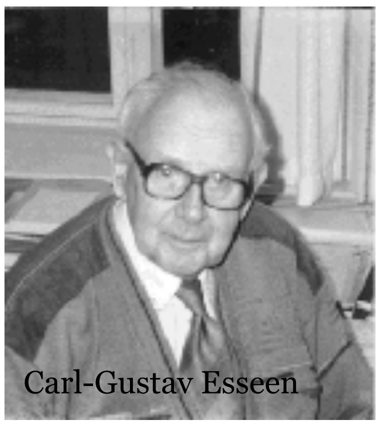


Central Limit Theorem

Central limit theorem: If $\mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1$ and $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$

$$Y_n \overset{\text{def.}}{=} \frac{X_1 + \ldots + X_n}{\sqrt{n}} \overset{\text{law}}{\longrightarrow} \mathcal{N}(0,1)$$

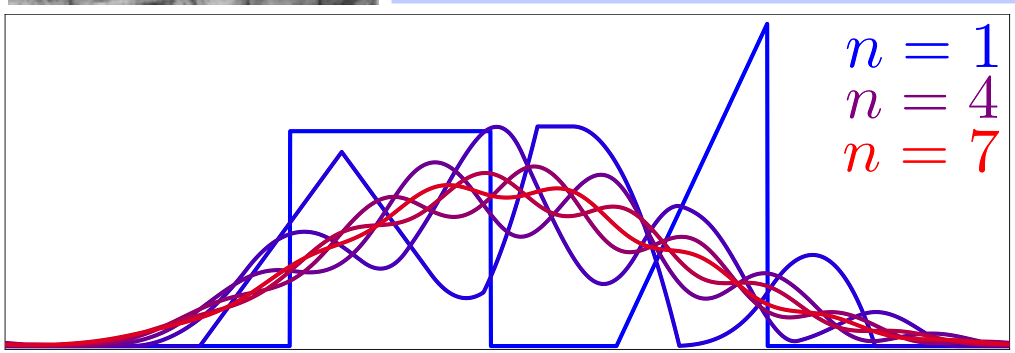
Kolmogorov-Smirnov distance: $d_{KS}(X,Y) \overset{\text{def.}}{=} \max_t |\mathbb{P}(X \leqslant t) - \mathbb{P}(Y \leqslant t)|$

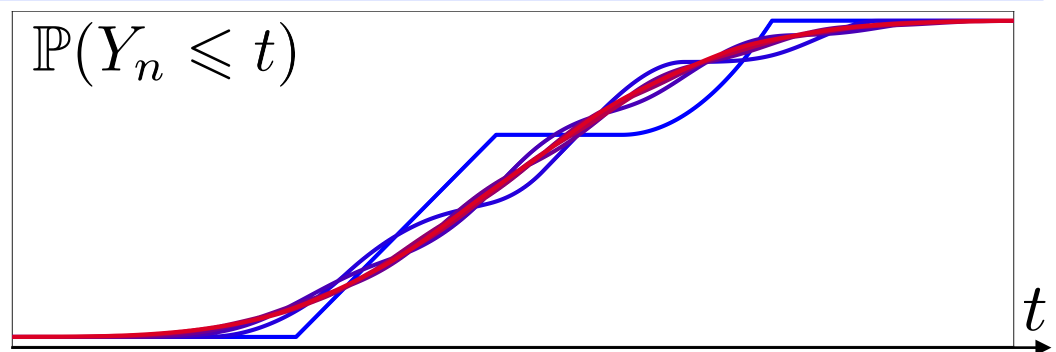


Metrizes convergence in law: $X \overset{\text{law}}{\to} Y \Leftrightarrow d_{\text{KS}}(X,Y) \to 0$

Theorem:
[Berry 1941]
[Esseen, 1942]

$$d_{\text{KS}}(Y_n, \mathcal{N}(0,1)) \leqslant \frac{C\mathbb{E}(|X|^3)}{\sqrt{n}} \qquad C \leqslant 1/2$$





Multi-dimensional extension: use W_1 in place of d_{KS}!

Overview

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

Strong Norms

Reference measure $\mathrm{d}x$ on $\mathcal{X}$.

L^p norms on densities:

$$D(\alpha,\beta) \overset{\text{def.}}{=} \left(\int_{\mathcal{X}} (\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x) - \frac{\mathrm{d}\beta}{\mathrm{d}x}(x))^p \mathrm{d}x \right)^{1/p} = \left\| \frac{\mathrm{d}\alpha}{\mathrm{d}x} - \frac{\mathrm{d}\beta}{\mathrm{d}x} \right\|_{L^p(\mathrm{d}x)}$$

$\rightarrow$ defined only if $\alpha \ll \mathrm{d}x$ and $\beta \ll \mathrm{d}x$.

Strong Norms

Reference measure $\mathrm{d}x$ on $\mathcal{X}$.

L^p norms on densities:

$$D(\alpha,\beta) \overset{\text{def.}}{=} \left(\int_{\mathcal{X}} \left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x) - \frac{\mathrm{d}\beta}{\mathrm{d}x}(x)\right)^p \mathrm{d}x \right)^{1/p} = \left\| \frac{\mathrm{d}\alpha}{\mathrm{d}x} - \frac{\mathrm{d}\beta}{\mathrm{d}x} \right\|_{L^p(\mathrm{d}x)}$$

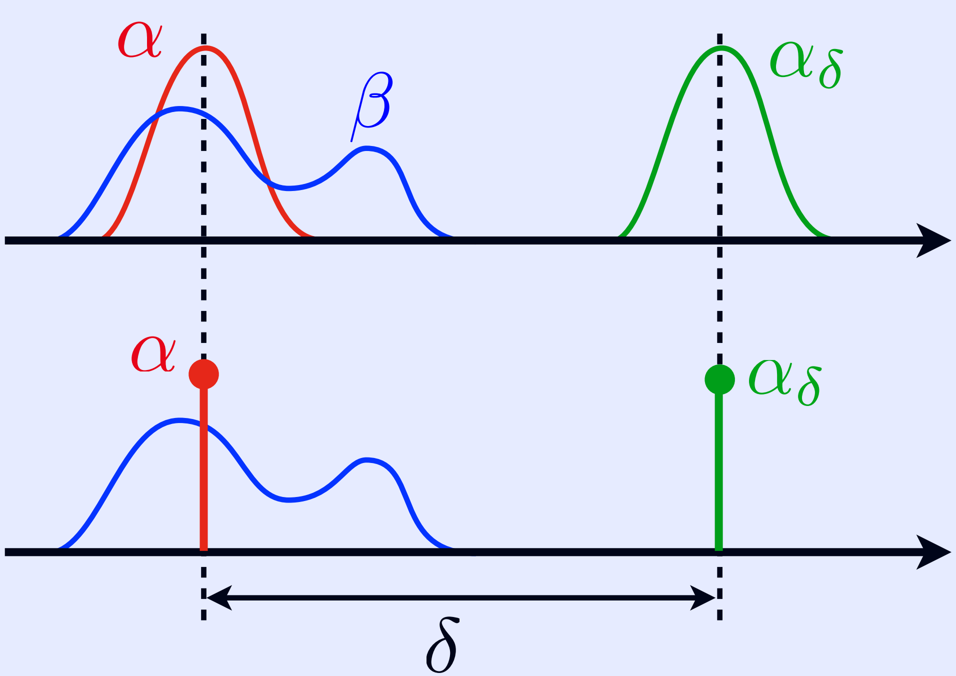
$\rightarrow$ defined only if $\alpha \ll \mathrm{d}x$ and $\beta \ll \mathrm{d}x$.

Metrizes the strong topology.

$$\alpha_\delta \xrightarrow{\text{weak}} \alpha$$

$$D(\alpha, \alpha_\delta) \approx \text{cst}$$

$$\mathrm{W}_p(\alpha, \alpha_\delta) = \delta$$



Csiszar Divergence

Comparing $\frac{d\alpha}{dx} \leftrightarrow \frac{d\beta}{dx}$ → $\frac{d\alpha}{d\beta} \leftrightarrow 1$

Csiszár φ-divergence: $\mathcal{D}_\varphi(\alpha|\beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{d\alpha}{d\beta}\right) d\beta + \varphi'_\infty \alpha^\perp(\mathcal{X})$

φ convex, $\varphi(1) = 0$, $\varphi \geqslant 0$ → Important if $\alpha(\mathcal{X}) \neq \beta(\mathcal{X})$.

Proposition: $\mathcal{D}_\varphi \geqslant 0$ is convex, $\mathcal{D}_\varphi(\alpha|\beta) = 0 \Leftrightarrow \alpha = \beta$.

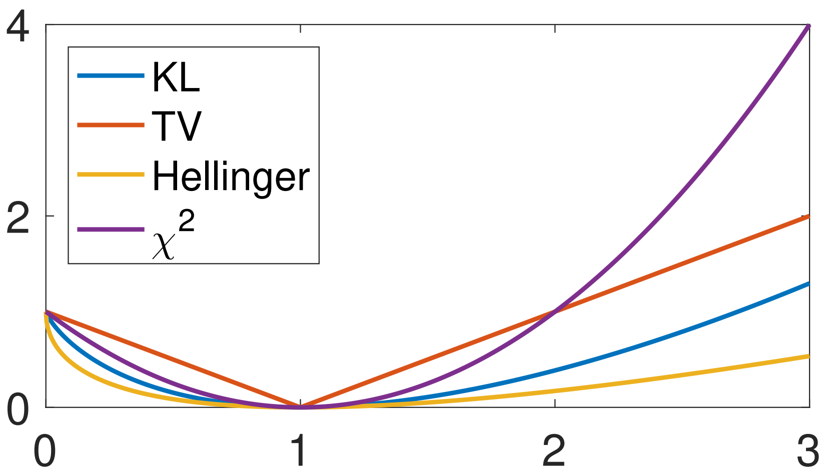
Csiszar Divergence

Comparing $\frac{\mathrm{d}\alpha}{\mathrm{d}x} \leftrightarrow \frac{\mathrm{d}\beta}{\mathrm{d}x} \longrightarrow \frac{\mathrm{d}\alpha}{\mathrm{d}\beta} \leftrightarrow 1$

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

φ convex, $\varphi(1) = 0$, $\varphi \geqslant 0$ $\longrightarrow$ Important if $\alpha(\mathcal{X}) \neq \beta(\mathcal{X})$.

Proposition: $\mathcal{D}_\varphi \geqslant 0$ is convex, $\mathcal{D}_\varphi(\alpha|\beta) = 0 \Leftrightarrow \alpha = \beta$.

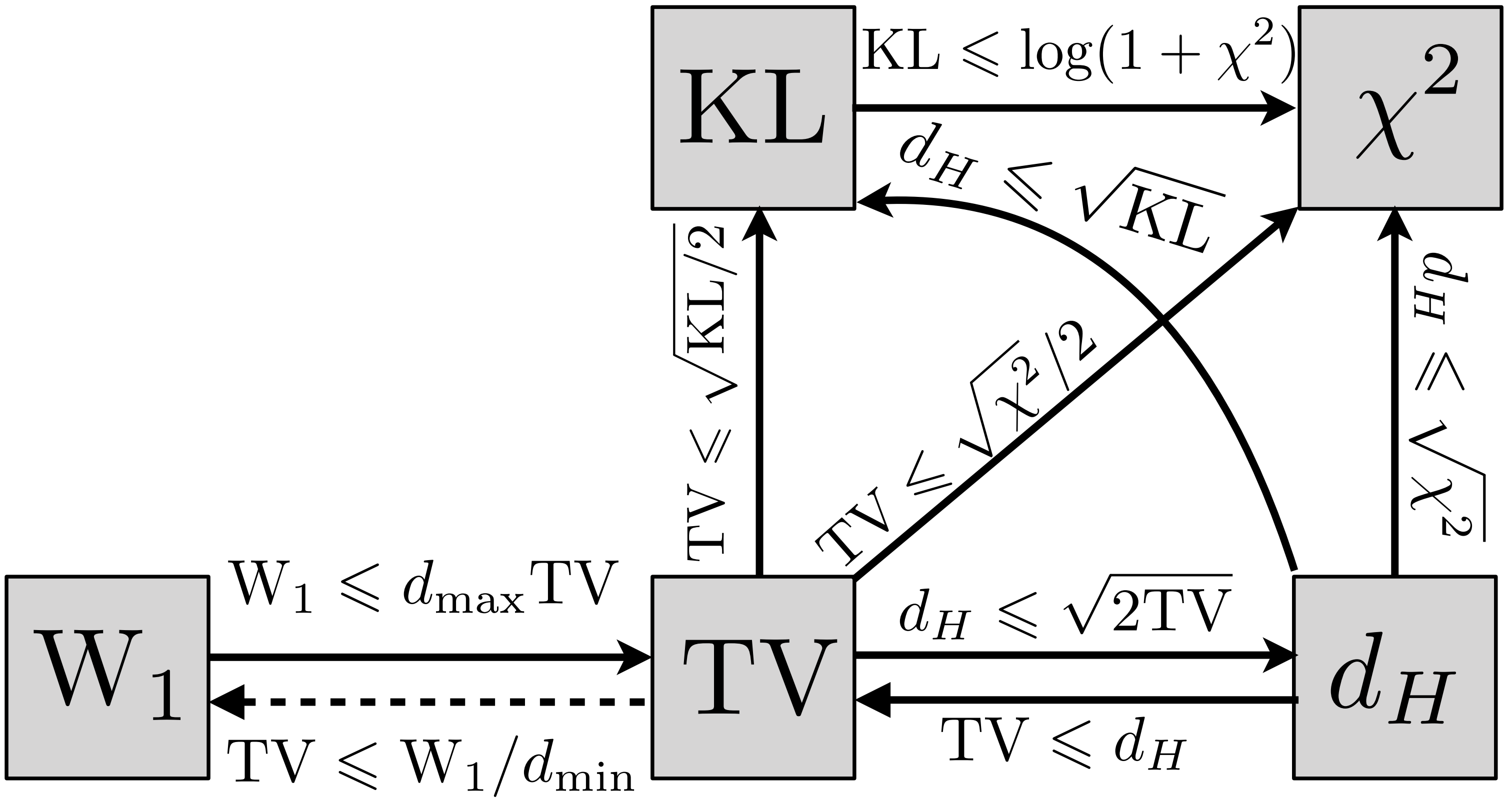


$\lvert s-1\rvert^2$	χ^2
$\lvert s-1\rvert$	TV norm
$s\log(s) - s + 1$	Generalized KL
$\lvert\sqrt{s}-1\rvert^2$	Hellinger distance
$s\log(s)$	KL

$$\|\alpha - \beta\|_{\mathrm{TV}} = \left\| \frac{\mathrm{d}\alpha}{\mathrm{d}x} - \frac{\mathrm{d}\beta}{\mathrm{d}x} \right\|_{L^1(\mathrm{d}x)}$$

$$d_{\mathrm{H}}(\alpha, \beta) = \left\| \sqrt{\frac{\mathrm{d}\alpha}{\mathrm{d}x}} - \sqrt{\frac{\mathrm{d}\beta}{\mathrm{d}x}} \right\|^2_{L^2(\mathrm{d}x)}$$

Equivalence and non-equivalence



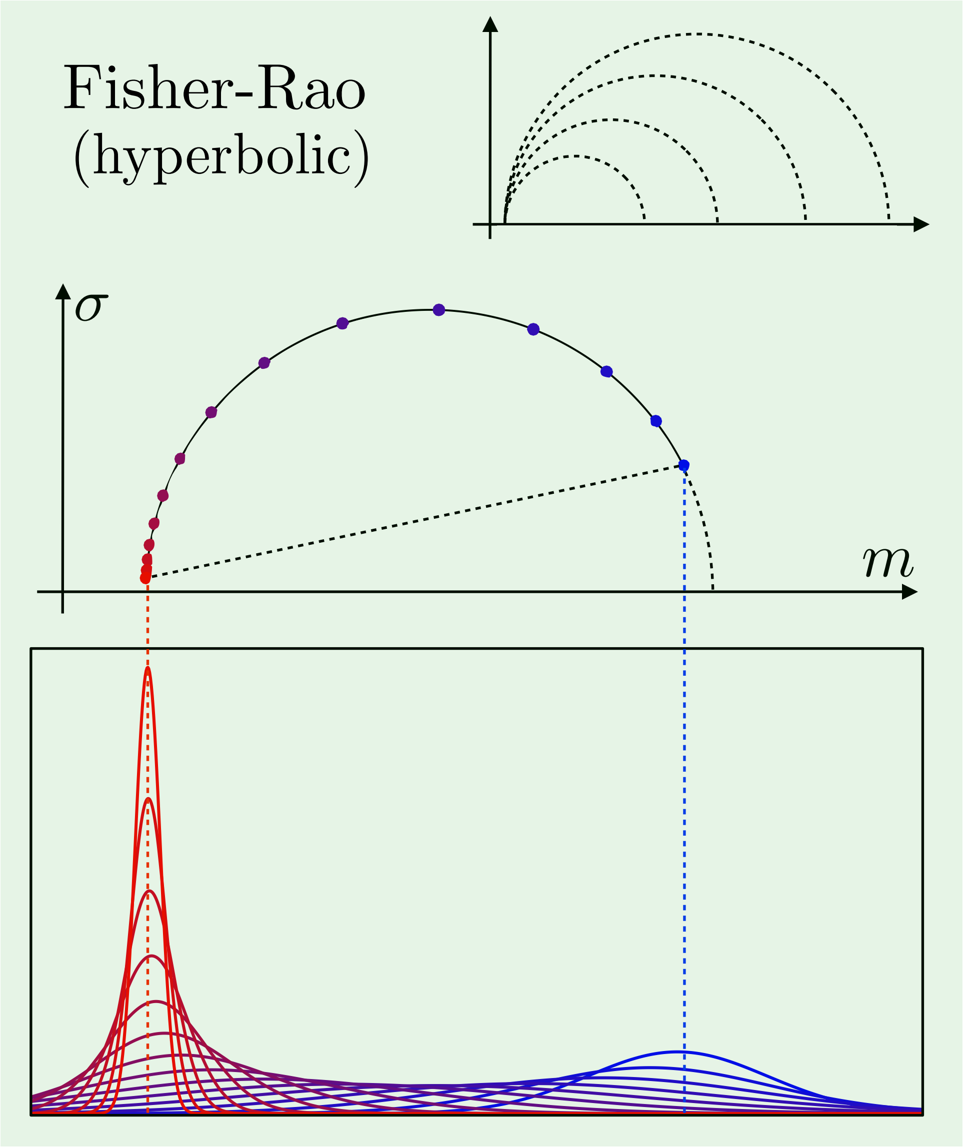
$d_{\max} = \sup_{(x,x')} d(x,x')$ $\quad$ $d_{\min} \stackrel{\text{def.}}{=} \min_{x \neq x'} d(x,x')$

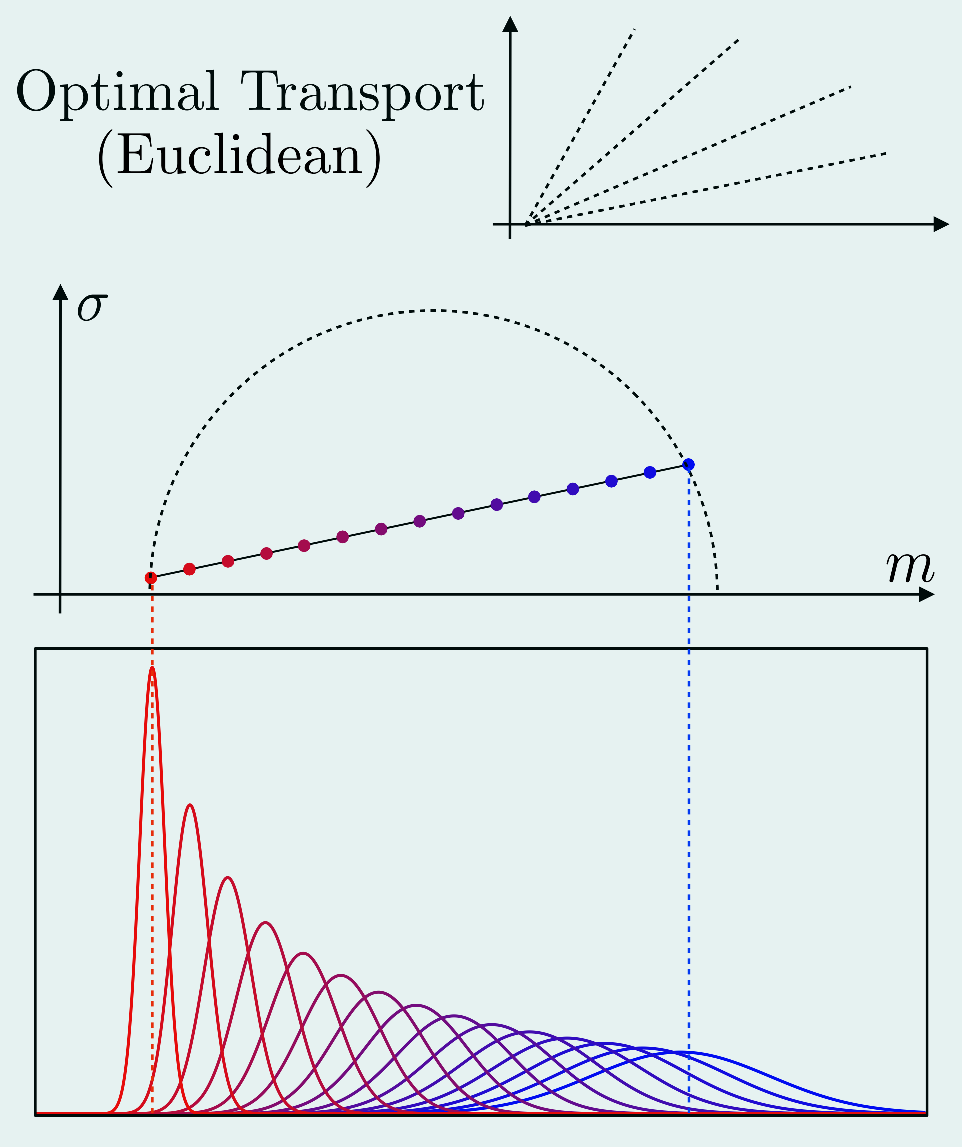
OT vs. KL (Fisher-Rao)

$$\mathcal{X} = \mathbb{R} \qquad \alpha = \mathcal{N}(m_\alpha, \sigma_\alpha), \quad \beta = \mathcal{N}(m_\beta, \sigma_\beta)$$

$$\mathrm{KL}(\alpha|\beta) = \frac{1}{2}\left(\frac{\sigma_\alpha^2}{\sigma_\beta^2} + \log\left(\frac{\sigma_\beta^2}{\sigma_\alpha^2}\right) + \frac{|m_\alpha - m_\beta|}{\sigma_\beta^2} - 1\right)$$

$$\mathrm{W}_2^2(\alpha,\beta) = |m_\alpha - m_\beta|^2 + |\sigma_\alpha - \sigma_\beta|^2$$





Overview

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

Dual Norms

Dual norms: (aka Integral Probability Metrics)

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

Dual Norms

Dual norms: (aka Integral Probability Metrics)

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

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

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

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

Negative Sobolev: $B = \left\{f \; ; \; k = 0, \ldots, s, \; \|\partial^k f\|_{L^2(\mathbb{R}^d)} \leqslant 1\right\}$

Dual Norms

Dual norms: (aka Integral Probability Metrics)

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

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

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

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

Negative Sobolev: $B = \left\{f \; ; \; k = 0, \ldots, s, \; \|\partial^k f\|_{L^2(\mathbb{R}^d)} \leqslant 1\right\}$

Proposition: If span(B) is dense in $\mathcal{C}(\mathcal{X})$,

$$\|\alpha\|_B \to 0 \quad \Longleftrightarrow \quad \alpha \overset{\text{weak}}{\longrightarrow} 0$$

Dual Norms

Dual norms: (aka Integral Probability Metrics)

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

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

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

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

Negative Sobolev: $B = \left\{f \; ; \; k = 0, \ldots, s, \; \|\partial^k f\|_{L^2(\mathbb{R}^d)} \leqslant 1\right\}$

Proposition: If span(B) is dense in $\mathcal{C}(\mathcal{X})$,

$$\|\alpha\|_B \to 0 \quad \Longleftrightarrow \quad \alpha \overset{\text{weak}}{\longrightarrow} 0$$

$\|\delta_x - \delta_y\|_{\text{TV}} = 2 \quad \text{if} \quad x \neq y$ → $f \in B$ needs to regular (e.g. $s > d/2$)

Hilbertian Norms on Measures

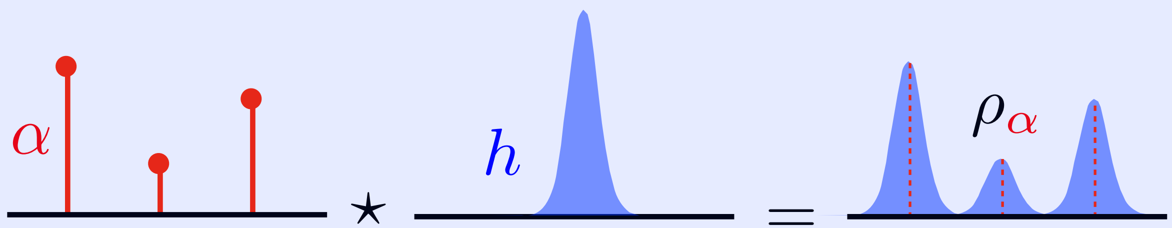
In $\mathcal{X} = \mathbb{R}^d$, smoothing with convolution:

$$\alpha \xrightarrow{\star h} \alpha \star h = \rho_\alpha \mathrm{d}x \qquad \rho_\alpha(x) \overset{\text{def.}}{=} \int_{\mathbb{R}^d} h(x-y)\mathrm{d}\alpha(y)$$

$$\alpha = \sum_i \mathbf{a}_i \delta_{x_i}$$

$$\Big\downarrow \star h$$

$$\rho_\alpha = \sum_i \mathbf{a}_i h(\cdot - x_i)$$

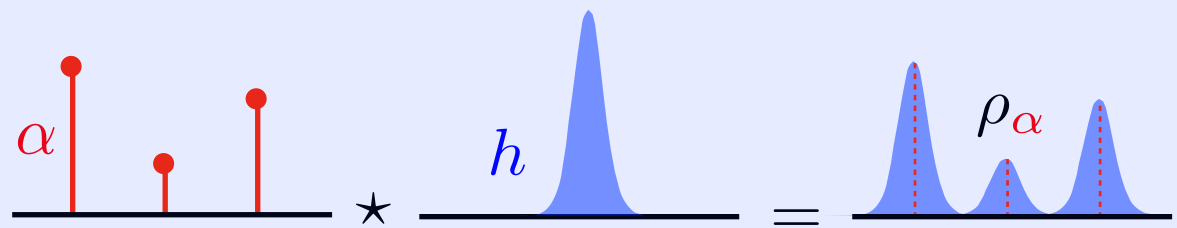


Hilbertian Norms on Measures

In $\mathcal{X} = \mathbb{R}^d$, smoothing with convolution:

$$\alpha \xrightarrow{\star h} \alpha \star h = \rho_\alpha \mathrm{d}x \qquad \rho_\alpha(x) \overset{\text{def.}}{=} \int_{\mathbb{R}^d} h(x-y)\mathrm{d}\alpha(y)$$

$$\alpha = \textstyle\sum_i \mathbf{a}_i \delta_{x_i} \; \xrightarrow{\star h} \; \rho_\alpha = \textstyle\sum_i \mathbf{a}_i h(\cdot - x_i)$$



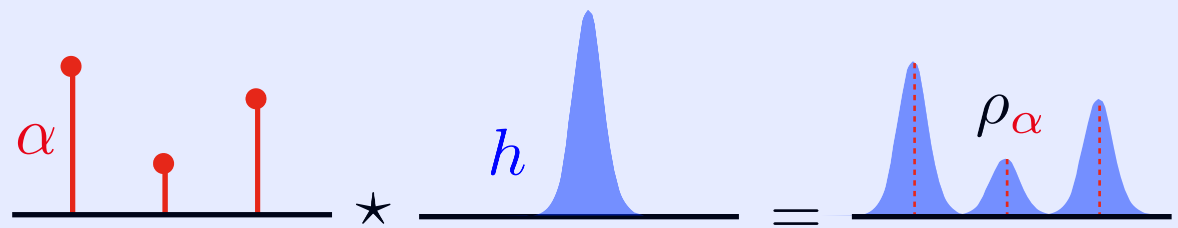
Hilbertian norm: $\|\alpha - \beta\|_k^2 \overset{\text{def.}}{=} \|\rho_\alpha - \rho_\beta\|_{L^2(\mathrm{d}x)}^2$

Hilbertian Norms on Measures

In $\mathcal{X} = \mathbb{R}^d$, smoothing with convolution:

$$\alpha \xrightarrow{\star h} \alpha \star h = \rho_\alpha \mathrm{d}x \qquad \rho_\alpha(x) \stackrel{\text{def.}}{=} \int_{\mathbb{R}^d} h(x-y)\mathrm{d}\alpha(y)$$

$$\alpha = \sum_i \mathbf{a}_i \delta_{x_i} \quad \xrightarrow{\star h} \quad \rho_\alpha = \sum_i \mathbf{a}_i h(\cdot - x_i)$$

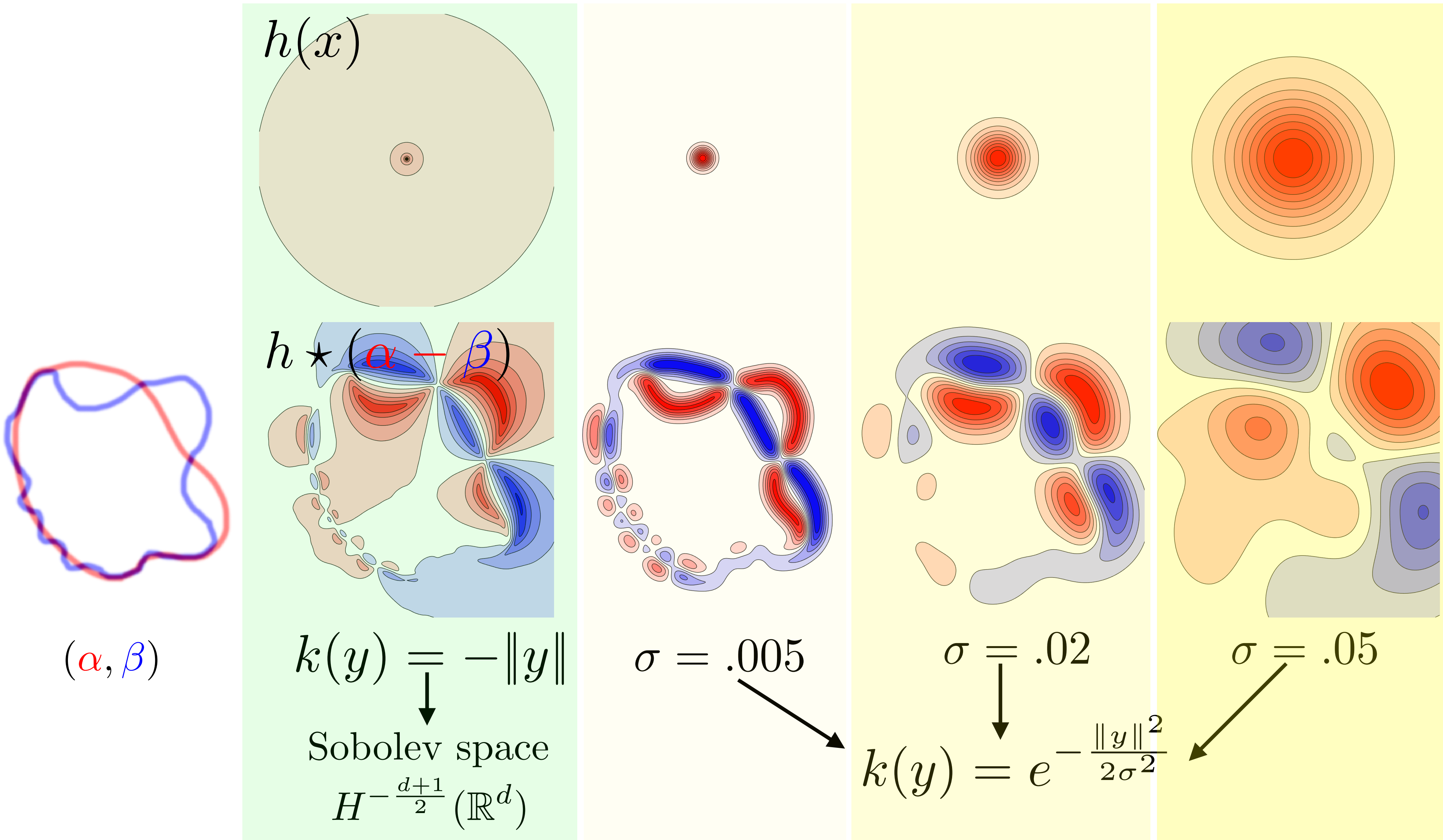


Hilbertian norm: $\|\alpha - \beta\|_k^2 \stackrel{\text{def.}}{=} \|\rho_\alpha - \rho_\beta\|_{L^2(\mathrm{d}x)}^2$

Kernel expression:
$$\|\xi\|_k^2 = \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} h(x-y)\mathrm{d}\xi(y) \right)^2 \mathrm{d}x$$
$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(y-y')\mathrm{d}\xi(y')\mathrm{d}\xi(y)$$

Correlation kernel: $k(y) \stackrel{\text{def.}}{=} \int_{\mathbb{R}^d} h(x-y)h(x)\mathrm{d}x$

Comparison of Kernels



Maximum Mean Discrepancies

$$\|\xi\|_k^2 = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(y - y')\mathrm{d}\xi(y')\mathrm{d}\xi(y)$$

Theorem: if $\hat{k}(\omega) > 0$, $\| \cdot \|_k$ metrizes weak convergence.

Maximum Mean Discrepancies

$$\|\xi\|_k^2 = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(y - y') \mathrm{d}\xi(y') \mathrm{d}\xi(y)$$

Theorem: if $\hat{k}(\omega) > 0$, $\|\cdot\|_k$ metrizes weak convergence.

$\rightarrow$ Extends to general $\mathcal{X}$ using positive kernels (MMD).

MMD:
$$\|\xi\|_k^2 \stackrel{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} k(x,y)\mathrm{d}\xi(x)\mathrm{d}\xi(y)$$
$$= \mathbb{E}(k(X,Y)),\ (X,Y) \sim \xi \text{ indep.}$$

Maximum Mean Discrepancies

$$\|\xi\|_k^2 = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(y-y')\mathrm{d}\xi(y')\mathrm{d}\xi(y)$$

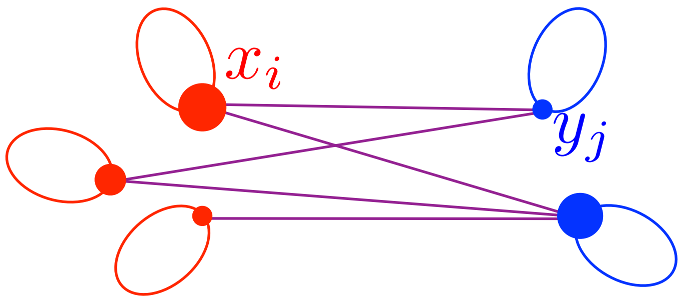
Theorem: if $\hat{k}(\omega) > 0$, $\|\cdot\|_k$ metrizes weak convergence.

$\rightarrow$ Extends to general $\mathcal{X}$ using positive kernels (MMD).

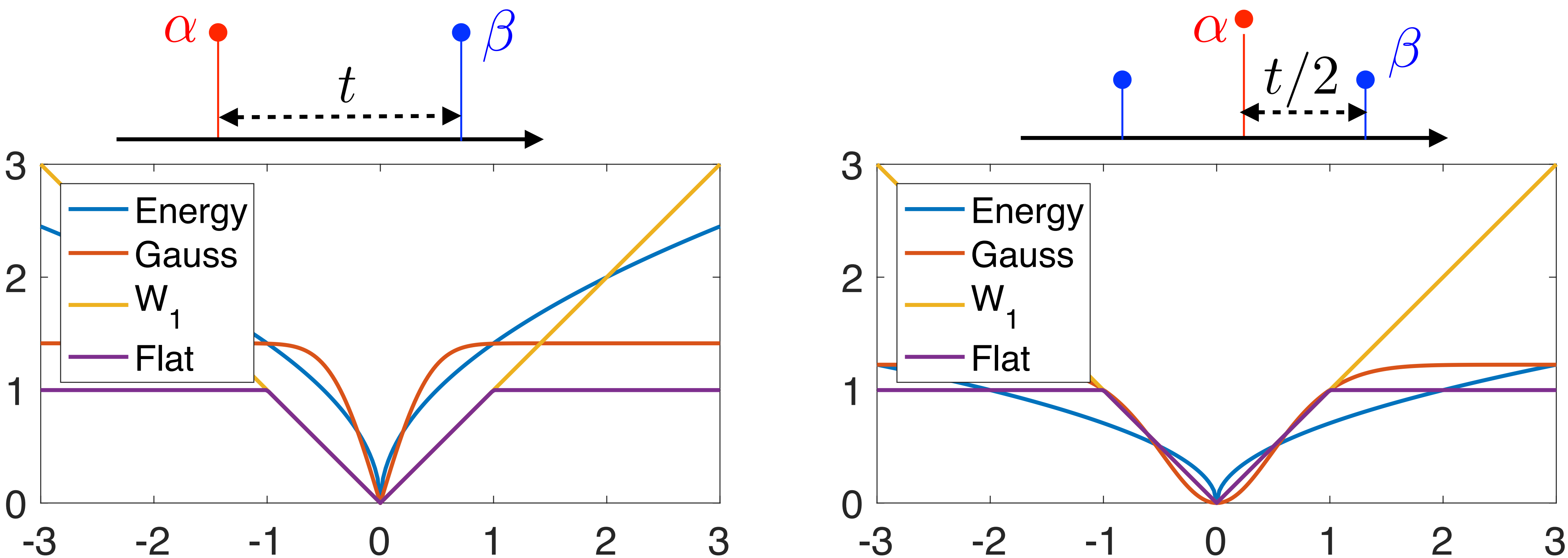
MMD:
$$\|\xi\|_k^2 \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} k(x,y)\mathrm{d}\xi(x)\mathrm{d}\xi(y)$$
$$= \mathbb{E}(k(X,Y)),\ (X,Y)\sim\xi \text{ indep.}$$

$$\alpha = \sum_{i=1}^n \mathbf{a}_i\delta_{x_i} \quad \beta = \sum_{j=1}^m \mathbf{b}_j\delta_{y_j}$$

$$\|\alpha-\beta\|^2 = \sum_{i,i'} \mathbf{a}_i\mathbf{a}_{i'}k(x_i,x_{i'}) - 2\sum_{i,j}\mathbf{a}_i\mathbf{b}_j k(x_i,y_j) + \sum_{j,j'}\mathbf{b}_j\mathbf{b}_{j'}k(y_j,y_{j'})$$



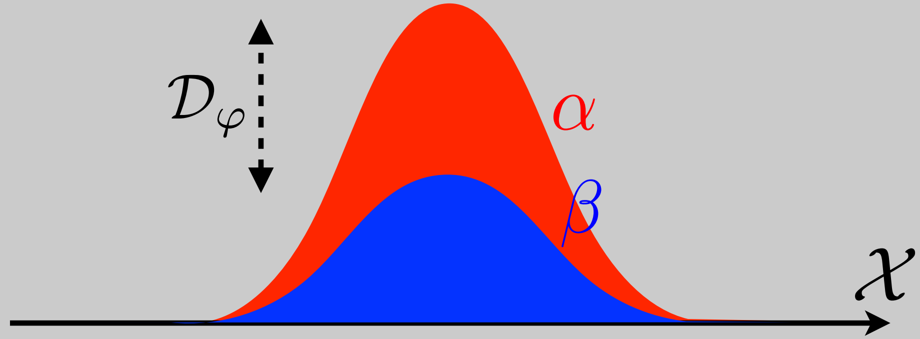
Comparison of Dual Norms



Csiszar Divergence vs Dual Norms

Csiszár divergences:

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

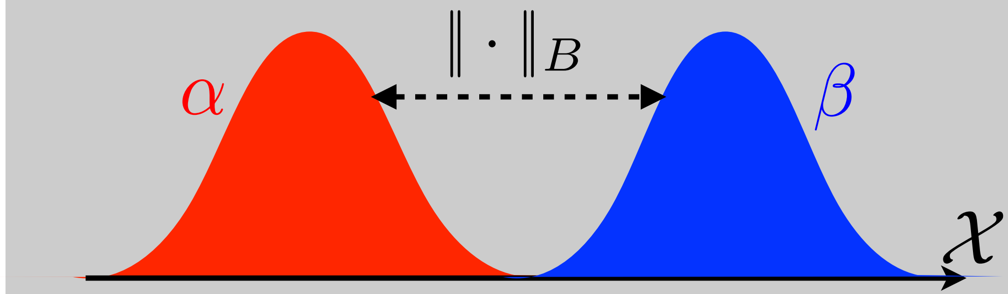


Strong topology

$\longrightarrow$ 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, …

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

OT Loss for Diffeomorphic Registration

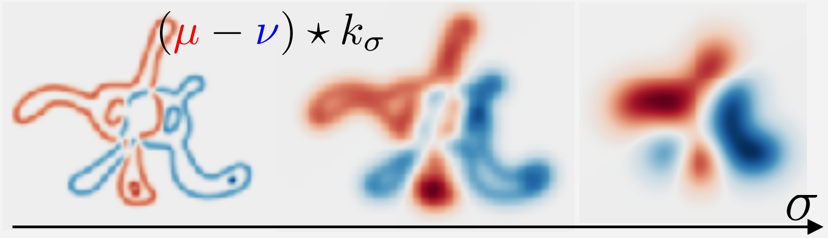
Joint work with J. Feydy, B. Charier, F-X. Vialard.

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

loss — regularity

Hilbertian loss (MMD/RKHS):

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

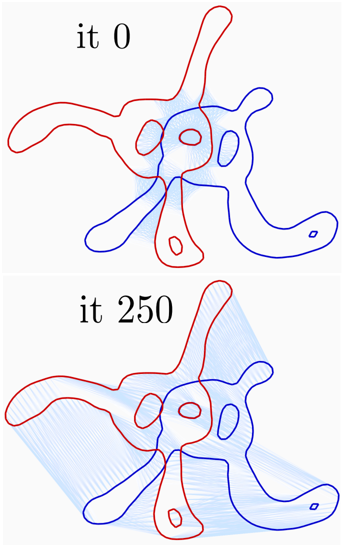


OT Loss for Diffeomorphic Registration

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

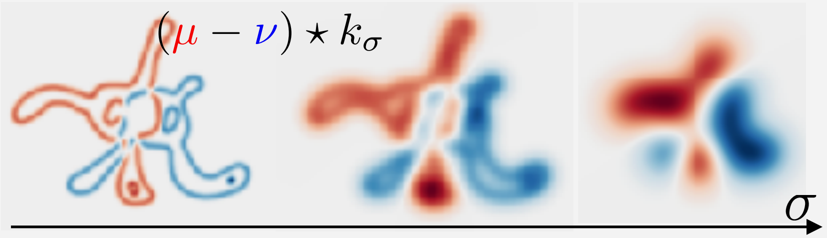
Shape registration: $\min_{\varphi \text{ diffeo}} D(\varphi(\mu), \nu) + R(\varphi)$

loss — regularity



Hilbertian loss (MMD/RKHS):

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



Sinkhorn divergence:

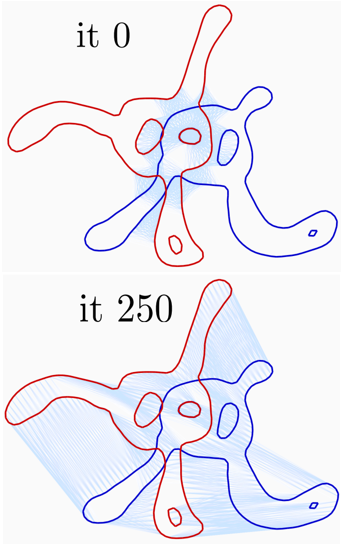
$$D(\mu, \nu) = \bar{W}_\varepsilon(\mu, \nu)$$

OT Loss for Diffeomorphic Registration

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

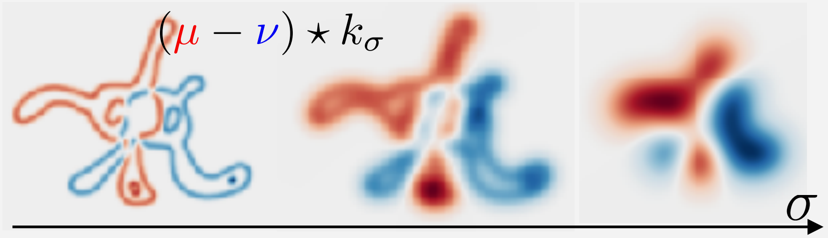
Shape registration: $\min_{\varphi \text{ diffeo}} D(\varphi(\mu), \nu) + R(\varphi)$

loss regularity



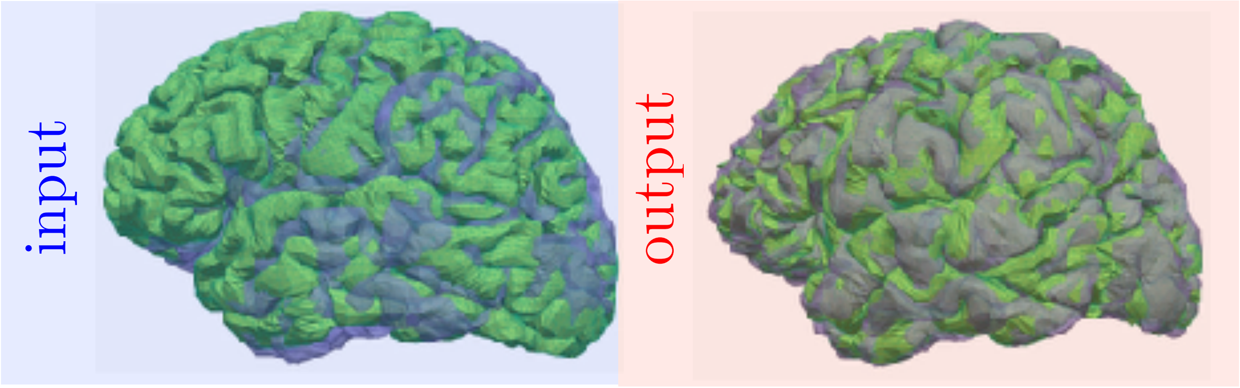
Hilbertian loss (MMD/RKHS):

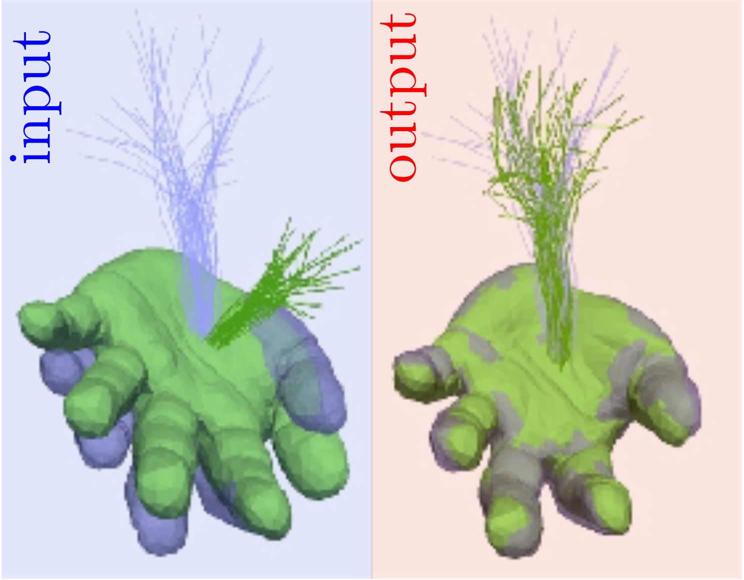
$$D(\mu, \nu) = \|k_\sigma \star (\mu - \nu)\|_{L^2}^2$$



Sinkhorn divergence:

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



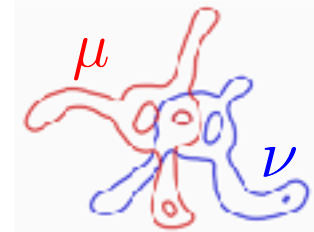


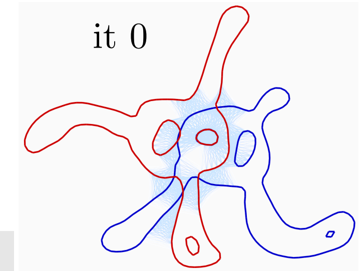
OT Loss for Diffeomorphic Registration

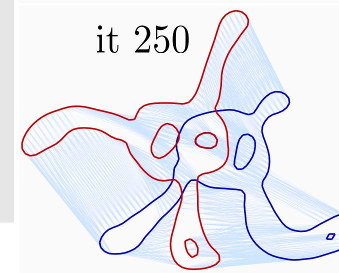
Joint work with J. Feydy, B. Charier, F-X. Vialard.

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

loss   regularity

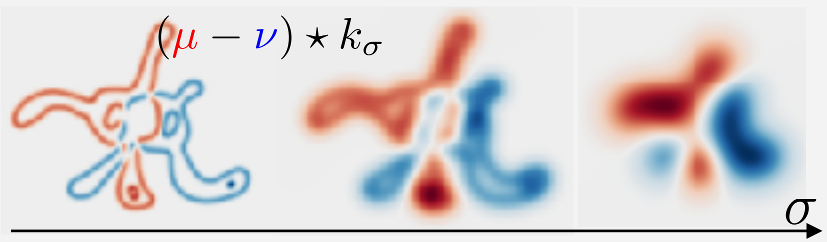






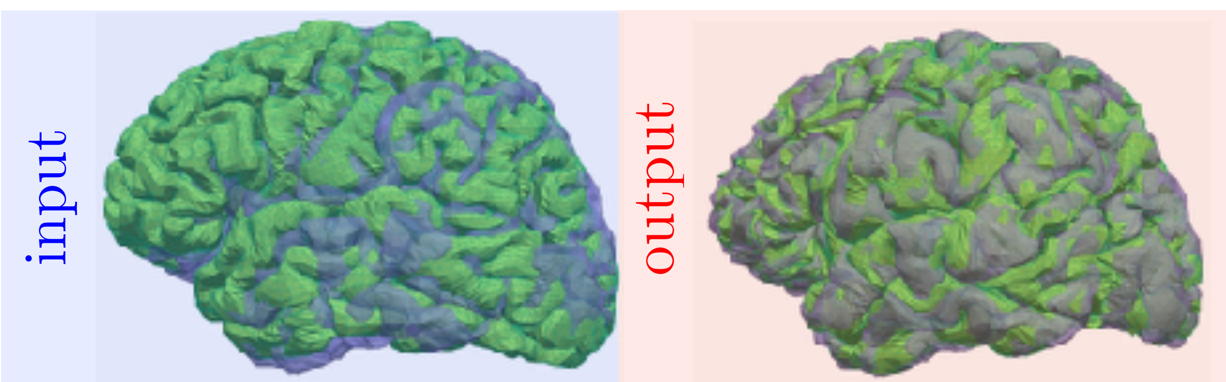
Hilbertian loss (MMD/RKHS):

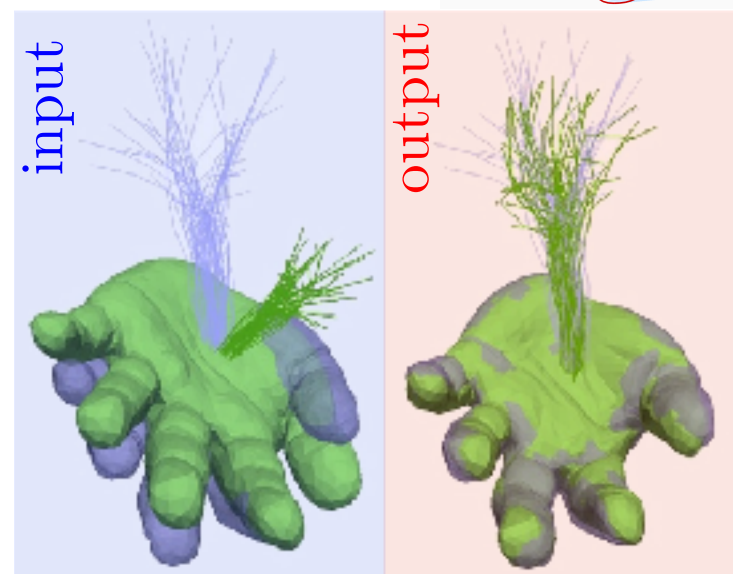
$$D(\mu, \nu) = \|k_\sigma \star (\mu - \nu)\|^2_{L^2}$$



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.

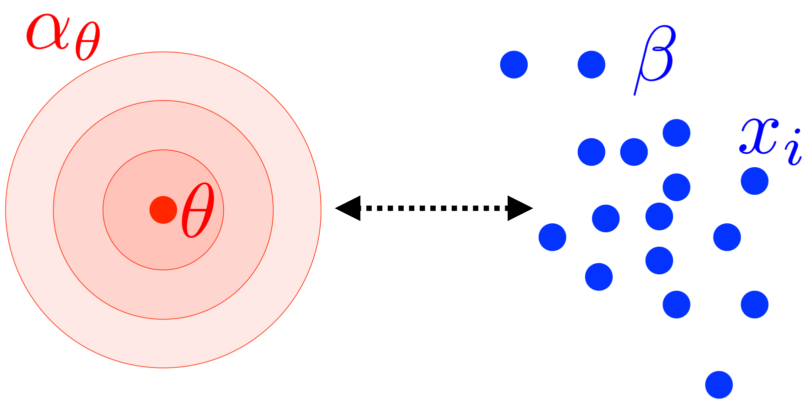
Overview

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

Density Fitting and Generative Models

Observations: $\beta \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$

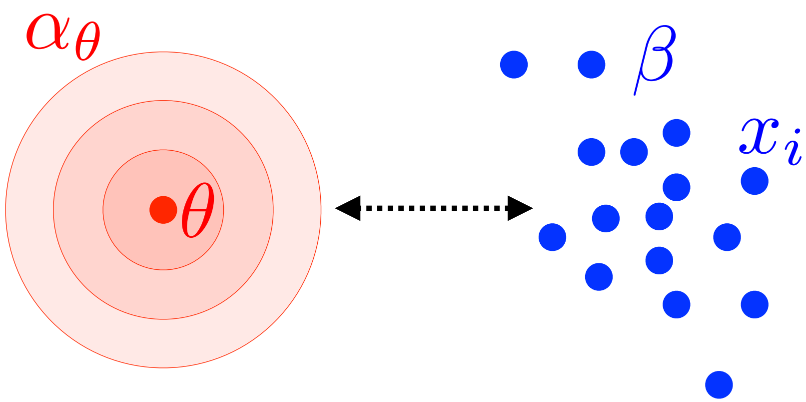
Parametric model: $\theta \mapsto \alpha_\theta$



Density Fitting and Generative Models

Observations: $\beta \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Parametric model: $\theta \mapsto \alpha_\theta$



Density fitting: $\mathrm{d}\alpha_\theta(x) = \rho_\theta(x)\mathrm{d}x$

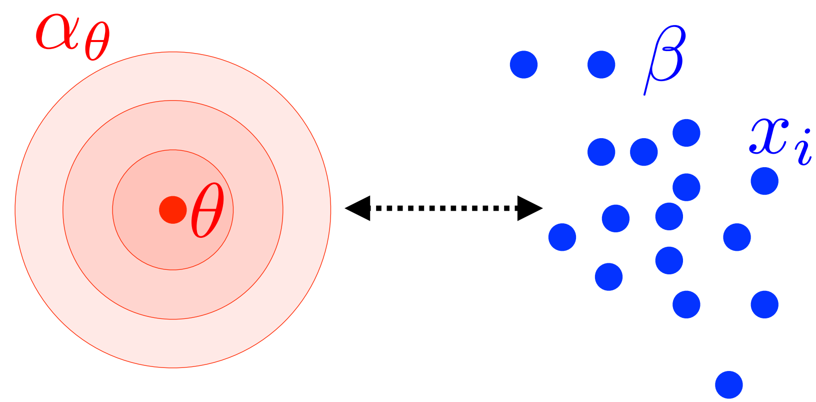
$$\min_\theta \widehat{\mathrm{KL}}(\beta|\alpha_\theta) \overset{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i))$$

Maximum likelihood (MLE)

Density Fitting and Generative Models

Observations: $\beta \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Parametric model: $\theta \mapsto \alpha_\theta$



Density fitting: $\mathrm{d}\alpha_\theta(x) = \rho_\theta(x)\mathrm{d}x$

$$\min_\theta \widehat{\mathrm{KL}}(\beta|\alpha_\theta) \overset{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i))$$

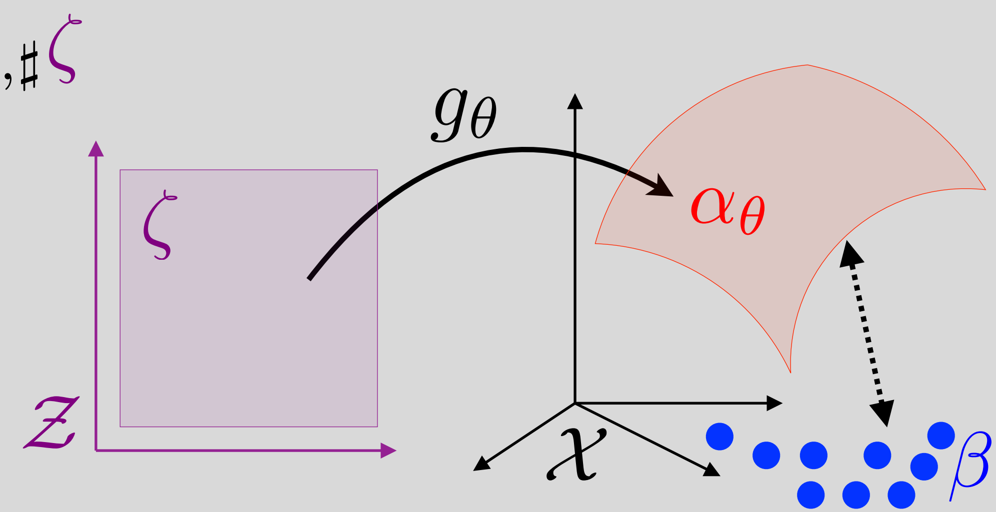
Maximum likelihood (MLE)

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

$$\widehat{\mathrm{KL}}(\beta|\alpha_\theta) = +\infty$$

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

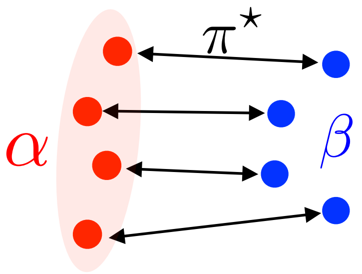


Loss Functions for Measures

Density fitting: $\min_{\theta} D(\alpha_\theta, \beta)$ $\beta = \frac{1}{n}\sum_i \delta_{x_i}$

Optimal Transport Distances

$$\mathrm{W}_p^p(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int d(x,y)^p \mathrm{d}\pi(x,y)$$



Loss Functions for Measures

Density fitting: $\min_{\theta} D(\alpha_\theta, \beta)$ $\beta = \frac{1}{n}\sum_i \delta_{x_i}$

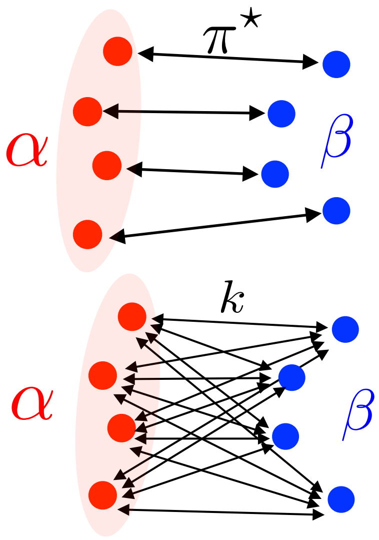
Optimal Transport Distances

$$W_p^p(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int d(x,y)^p \mathrm{d}\pi(x,y)$$

Maximum Mean Discrepancy (MMD)

$$\|\alpha - \beta\|_k^2 \stackrel{\text{def.}}{=} \int k(x,y)\mathrm{d}(\alpha(x) - \beta(x))\mathrm{d}(\alpha(y) - \beta(y))$$

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

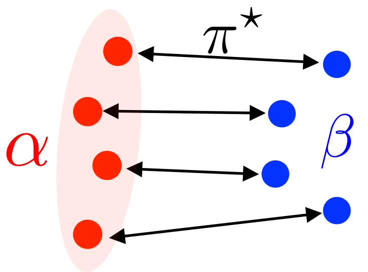


Loss Functions for Measures

Density fitting: $\min_\theta D(\alpha_\theta, \beta)$ $\quad \beta = \frac{1}{n}\sum_i \delta_{x_i}$

Optimal Transport Distances

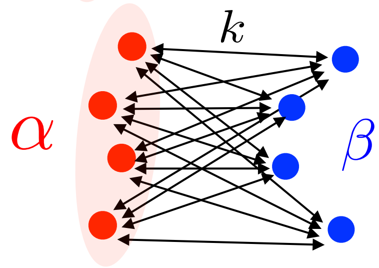
$$\mathrm{W}_p^p(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int d(x,y)^p \mathrm{d}\pi(x,y)$$



Maximum Mean Discrepancy (MMD)

$$\|\alpha - \beta\|_k^2 \stackrel{\text{def.}}{=} \int k(x,y)\mathrm{d}(\alpha(x) - \beta(x))\mathrm{d}(\alpha(y) - \beta(y))$$

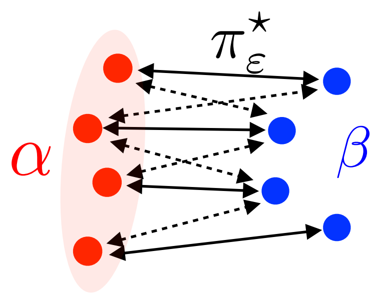
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Genevay, Peyré, Cuturi 17]

$$\mathrm{W}_{\varepsilon,p}^p(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int d^p \mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta)$$

$$\bar{W}_{p,\varepsilon}^p(\alpha, \beta)^p \stackrel{\text{def.}}{=} \mathrm{W}_{p,\varepsilon}^p(\alpha, \beta)^p - \frac{1}{2}\mathrm{W}_{p,\varepsilon}^p(\alpha, \beta)^p - \frac{1}{2}\mathrm{W}_{p,\varepsilon}^p(\alpha, \beta)^p$$

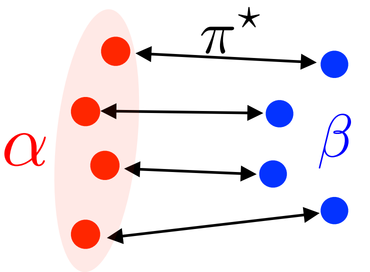


Loss Functions for Measures

Density fitting: $\min_{\theta} D(\alpha_\theta, \beta)$ $\beta = \frac{1}{n}\sum_i \delta_{x_i}$

Optimal Transport Distances

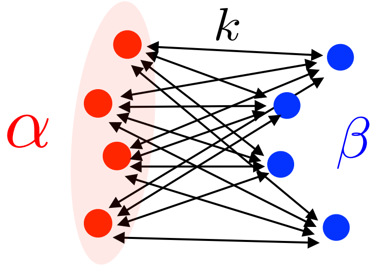
$$\mathrm{W}_p^p(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int d(x,y)^p \mathrm{d}\pi(x,y)$$



Maximum Mean Discrepancy (MMD)

$$\|\alpha - \beta\|_k^2 \overset{\text{def.}}{=} \int k(x,y)\mathrm{d}(\alpha(x)-\beta(x))\mathrm{d}(\alpha(y)-\beta(y))$$

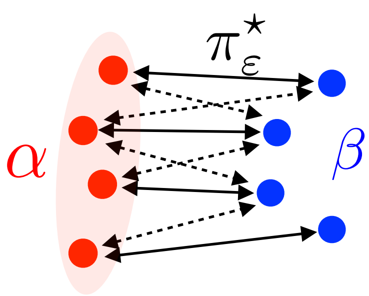
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Genevay, Peyré, Cuturi 17]

$$\mathrm{W}_{\varepsilon,p}^p(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int d^p \mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi|\alpha\otimes\beta)$$

$$\bar{W}_{p,\varepsilon}^p(\alpha,\beta)^p \overset{\text{def.}}{=} \mathrm{W}_{p,\varepsilon}^p(\alpha,\beta)^p - \frac{1}{2}\mathrm{W}_{p,\varepsilon}^p(\alpha,\beta)^p - \frac{1}{2}\mathrm{W}_{p,\varepsilon}^p(\alpha,\beta)^p$$



Theorem: [Genevay, P, C, 17] $\bar{W}_{\varepsilon,p}^p(\alpha,\beta) \xrightarrow{\varepsilon\to 0} \mathrm{W}_p^p(\alpha,\beta)$, $\bar{W}_{\varepsilon,p}^p(\alpha,\beta) \xrightarrow{\varepsilon\to +\infty} \|\alpha-\beta\|_k^2$

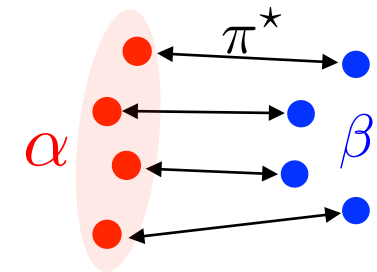
for $k(x,y) = -d(x,y)^p$

Loss Functions for Measures

Density fitting: $\min_{\theta} D(\alpha_\theta, \beta)$ $\quad \beta = \frac{1}{n}\sum_i \delta_{x_i}$

Optimal Transport Distances

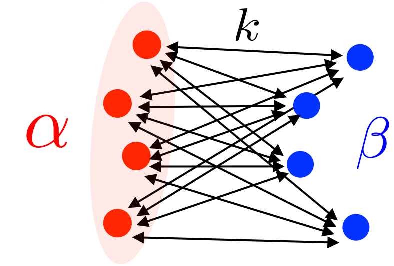
$$\mathrm{W}_p^p(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int d(x,y)^p \mathrm{d}\pi(x,y)$$



Maximum Mean Discrepancy (MMD)

$$\|\alpha-\beta\|_k^2 \overset{\text{def.}}{=} \int k(x,y)\mathrm{d}(\alpha(x)-\beta(x))\mathrm{d}(\alpha(y)-\beta(y))$$

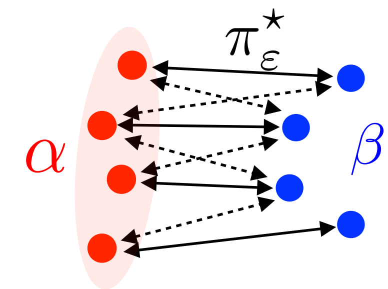
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Genevay, Peyré, Cuturi 17]

$$\mathrm{W}_{\varepsilon,p}^p(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int d^p \mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi|\alpha\otimes\beta)$$

$$\bar{W}_{p,\varepsilon}^p(\alpha,\beta)^p \overset{\text{def.}}{=} \mathrm{W}_{p,\varepsilon}^p(\alpha,\beta)^p - \frac{1}{2}\mathrm{W}_{p,\varepsilon}^p(\alpha,\alpha)^p - \frac{1}{2}\mathrm{W}_{p,\varepsilon}^p(\alpha,\beta)^p$$



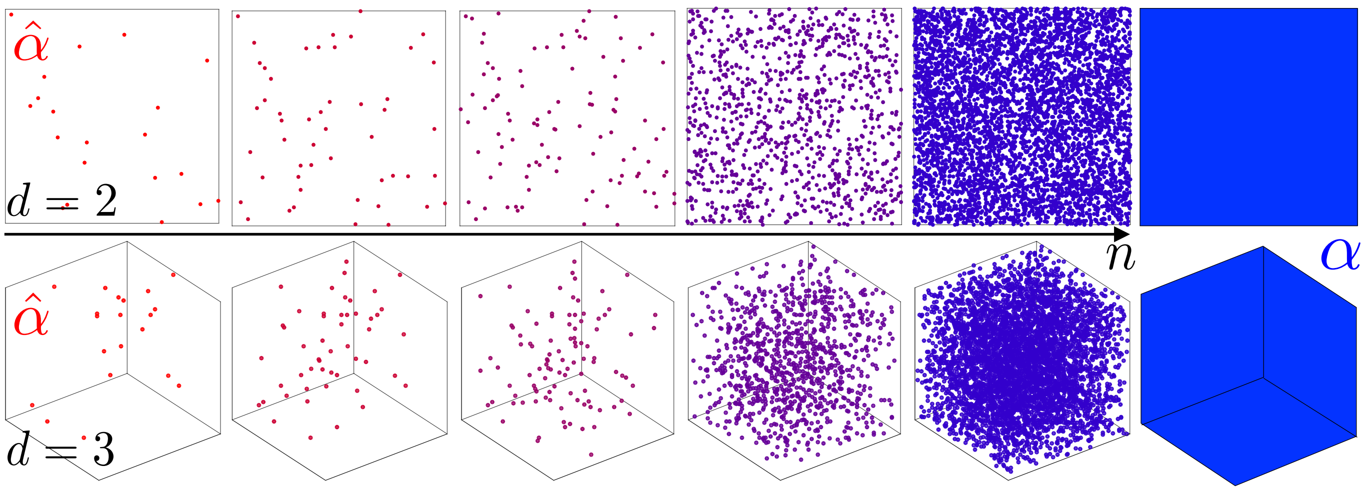
Theorem: [Genevay, P, C, 17] $\quad \bar{W}_{\varepsilon,p}^p(\alpha,\beta) \xrightarrow{\varepsilon\to 0} \mathrm{W}_p^p(\alpha,\beta)$, $\quad \bar{W}_{\varepsilon,p}^p(\alpha,\beta) \xrightarrow{\varepsilon\to+\infty} \|\alpha-\beta\|_k^2$ $\quad$ for $k(x,y) = -d(x,y)^p$

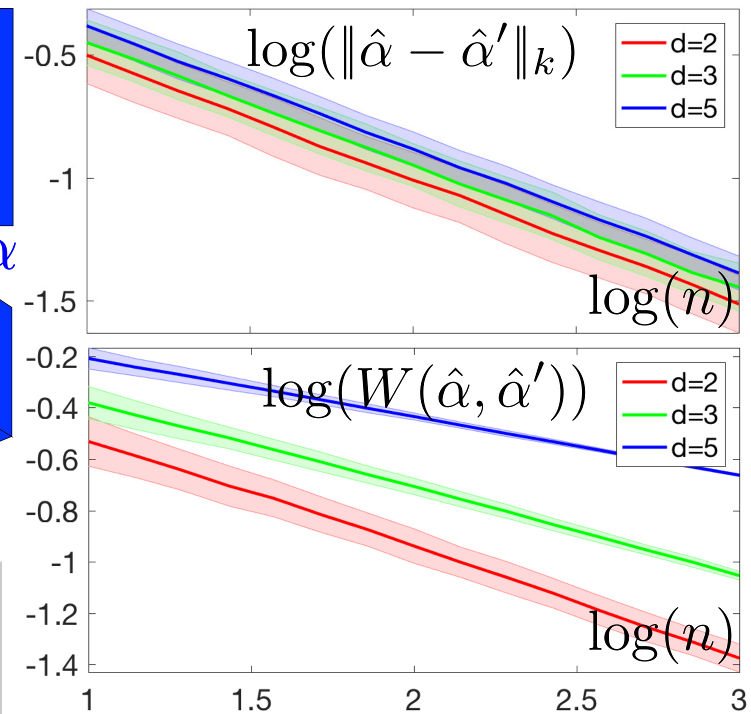
Best of both worlds:

$\rightarrow$ 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).

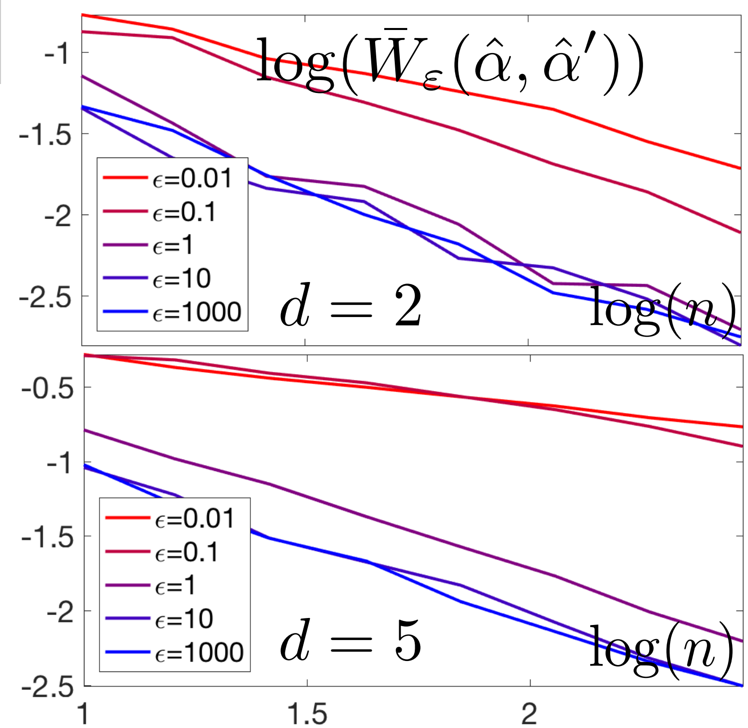
Sample Complexity



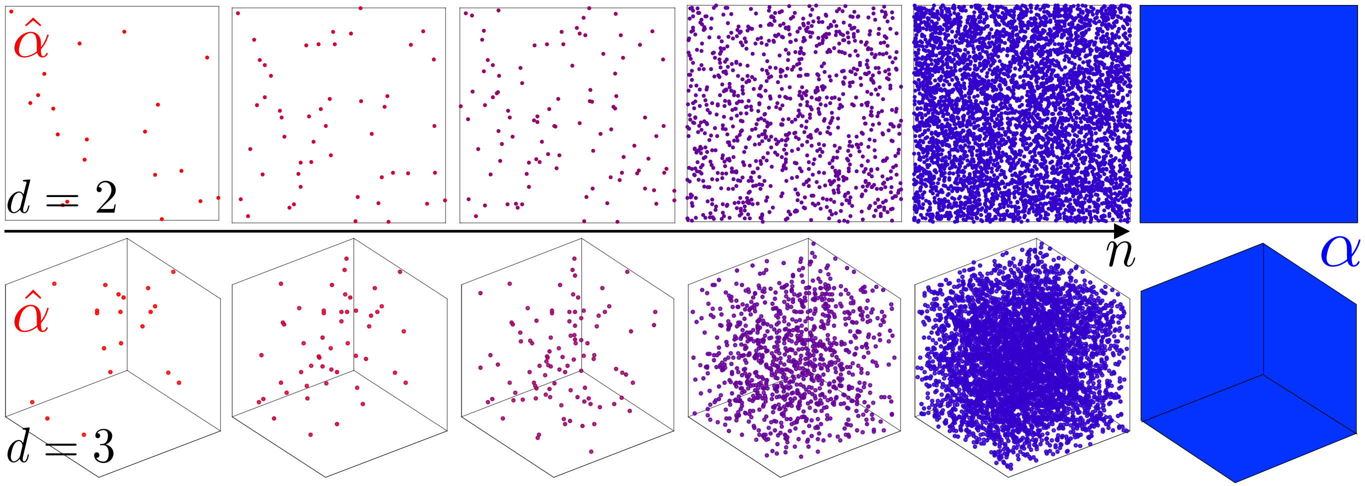


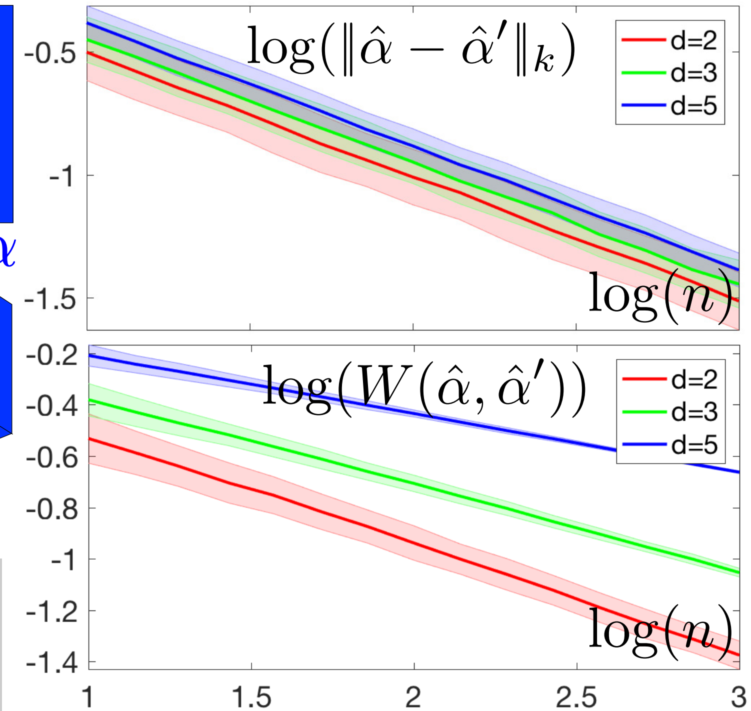
Theorem: $\mathbb{E}(|W(\hat{\alpha}, \hat{\beta}) - W(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$

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



Sample Complexity

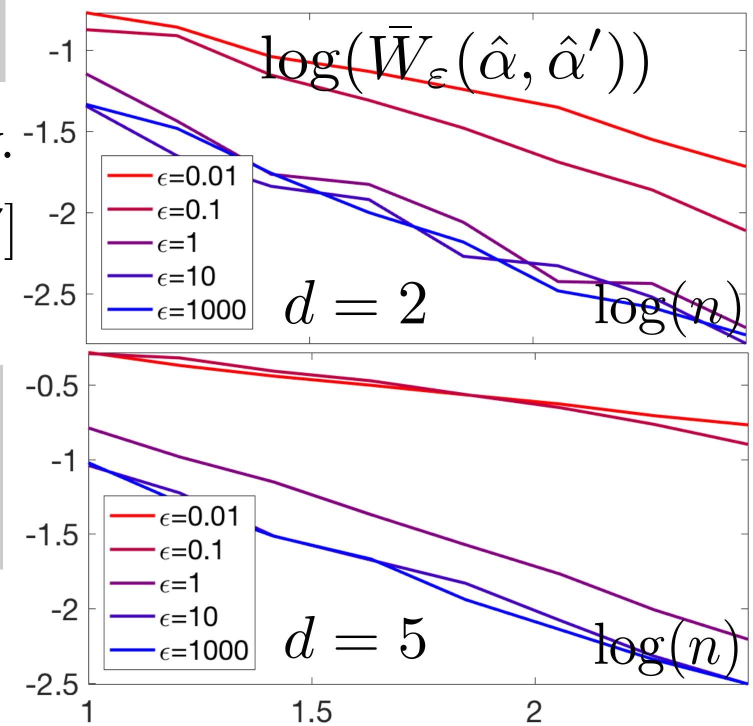




Theorem:
$$\mathbb{E}(|W(\hat\alpha,\hat\beta)-W(\alpha,\beta)|)=O(n^{-\frac{1}{d}})$$
$$\mathbb{E}(|\|\hat\alpha-\hat\beta\|_k-\|\alpha-\beta\|_k|)=O(n^{-\frac{1}{2}})$$

Optimal transport: suffers from curse of dimensionality.

→ Adapt to support dimensionality [Weed, Bach 2017]



Open problem: sample complexity of $\bar W_\varepsilon$?

Overview

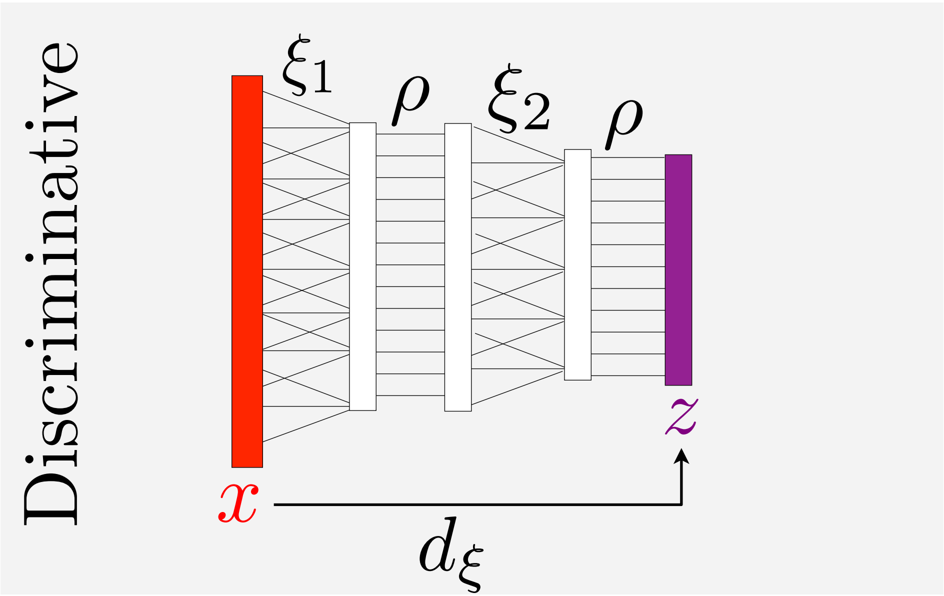
- Csiszar Divergences
- Dual Norms and MMD
- Minimum Kantorovitch Estimators
- **Deep Generative Models Fitting**

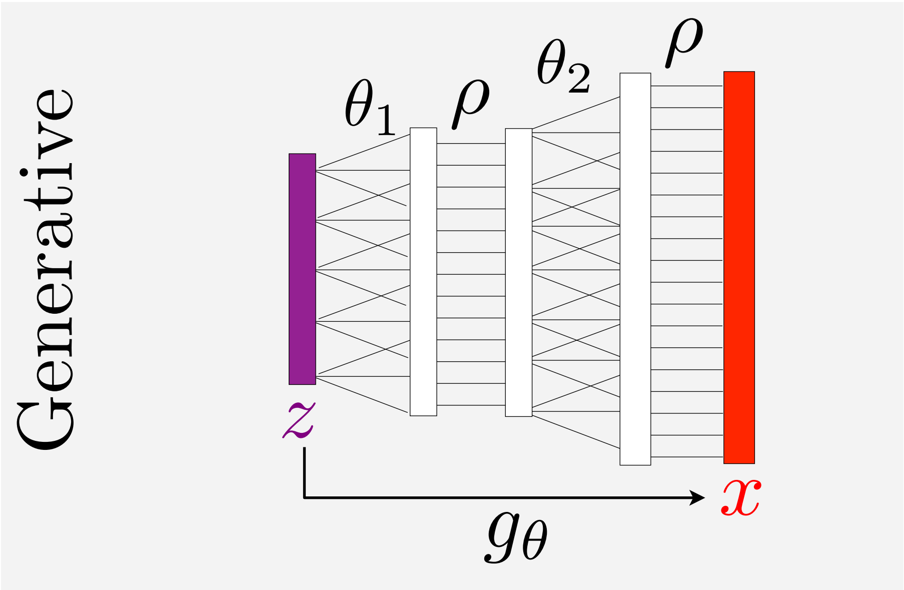
Deep Discriminative vs Generative Models

Deep networks:

$$d_\xi(x) = \rho(\xi_K(\ldots\rho(\xi_2(\rho(\xi_1(x)\ldots)$$

$$g_\theta(z) = \rho(\theta_K(\ldots\rho(\theta_2(\rho(\theta_1(z)\ldots)$$



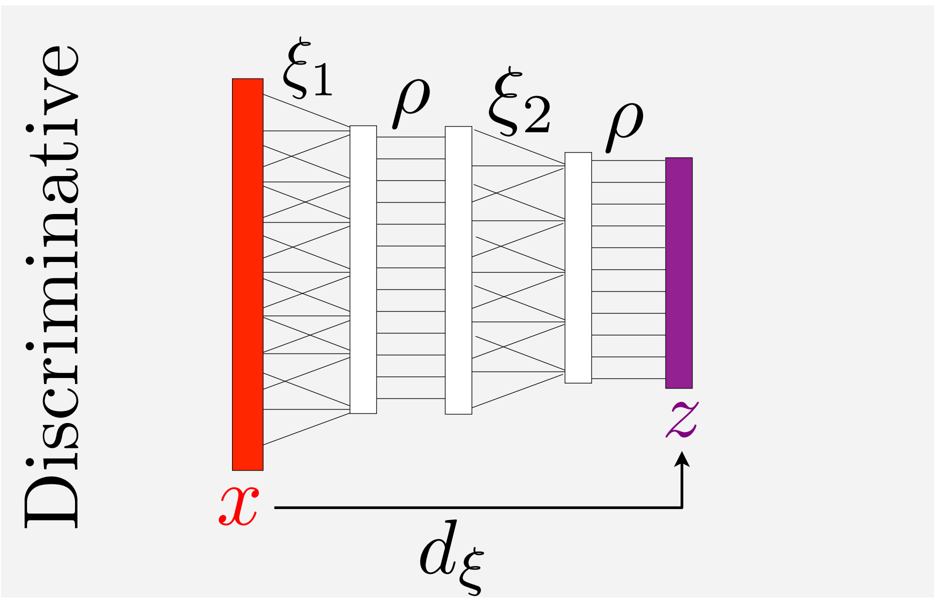


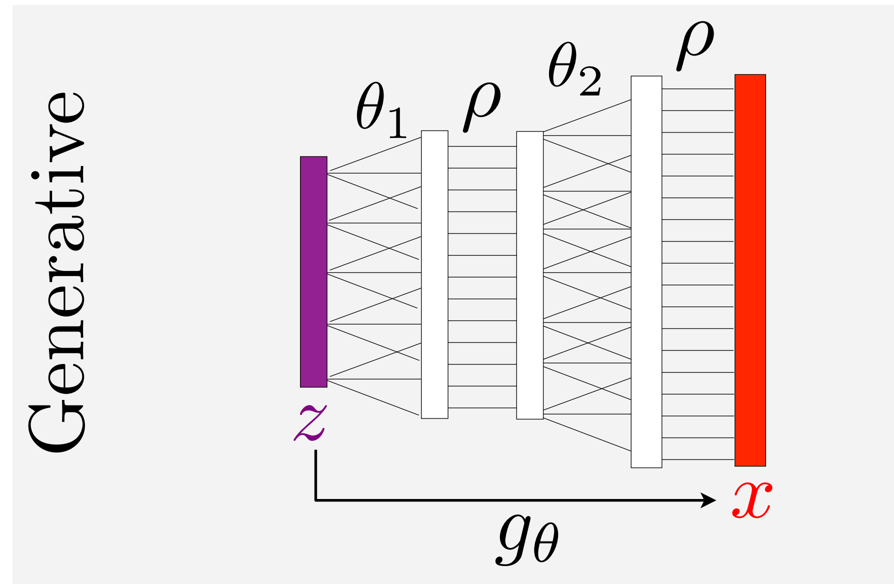
Deep Discriminative vs Generative Models

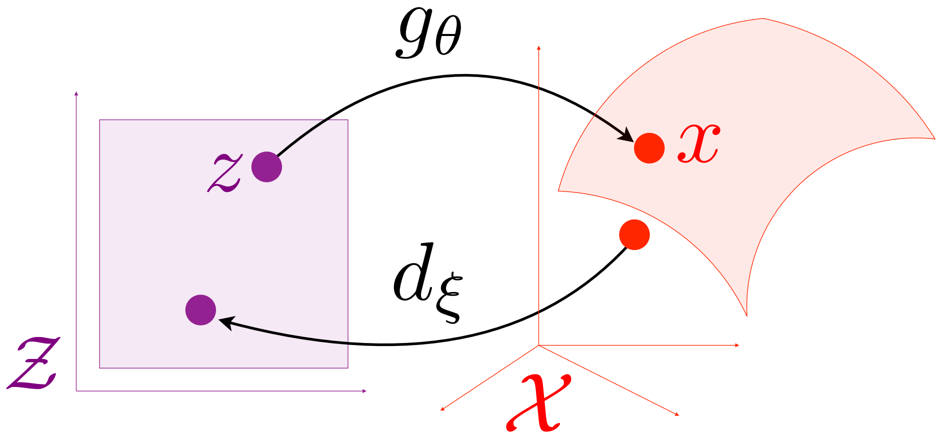
Deep networks:

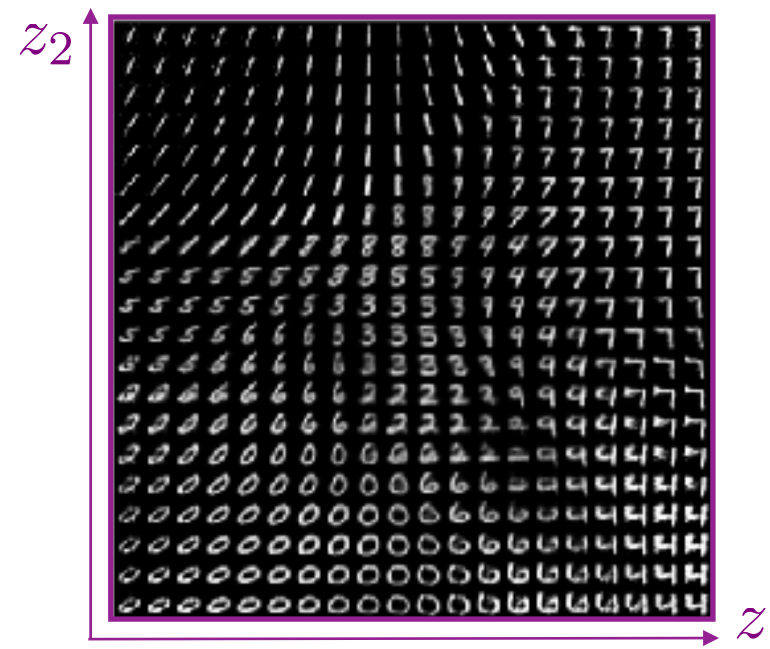
$$d_{\xi}(x) = \rho(\xi_K(\ldots\rho(\xi_2(\rho(\xi_1(x)\ldots)$$

$$g_{\theta}(z) = \rho(\theta_K(\ldots\rho(\theta_2(\rho(\theta_1(z)\ldots)$$

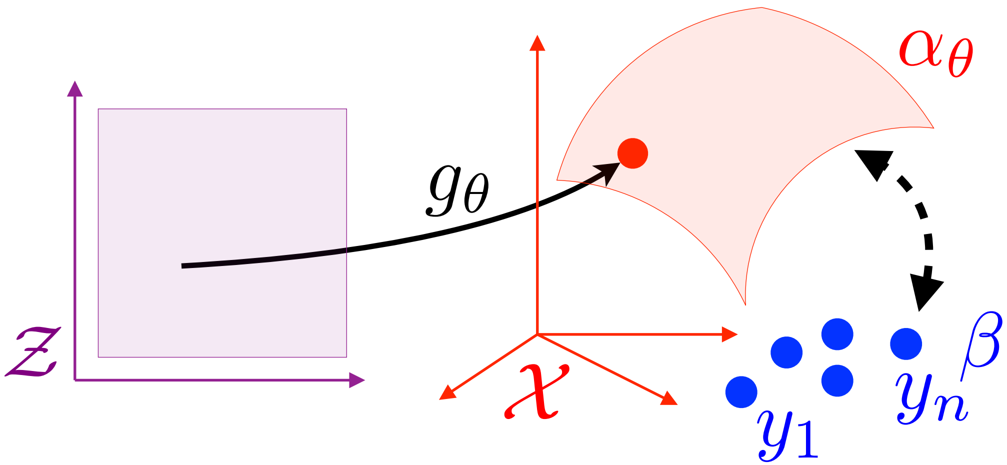








Training Architecture



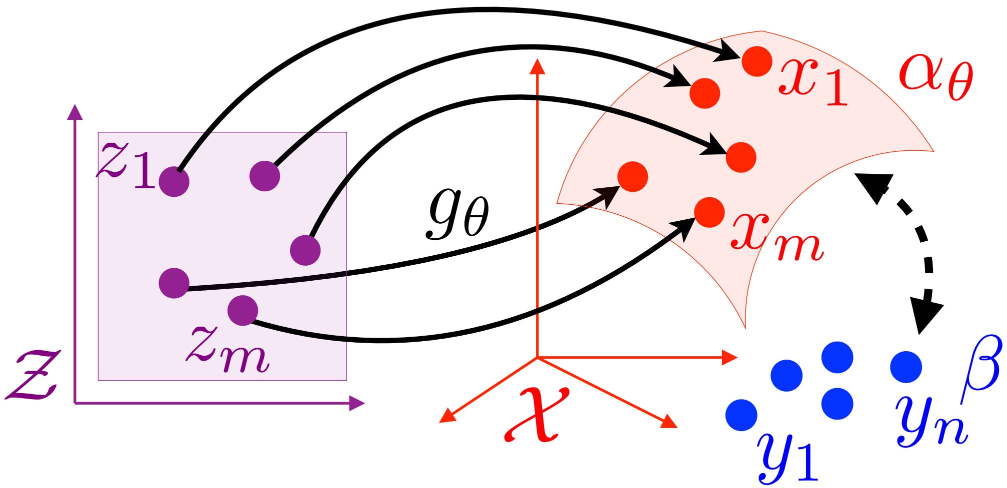
$$\min_\theta E(\theta) \stackrel{\text{def.}}{=} \bar{W}_\varepsilon(\alpha_\theta, \beta)$$

Stochastic gradient descent

$$\theta^{(\ell)} = \theta^{(\ell)} - \tau_\ell \nabla \hat{E}_L(\theta)$$

$$\hat{E}(\theta) \stackrel{\text{def.}}{=} \bar{W}_\varepsilon^L(\tfrac{1}{m} \textstyle\sum_i g_\theta(z_i), \beta)$$

Training Architecture



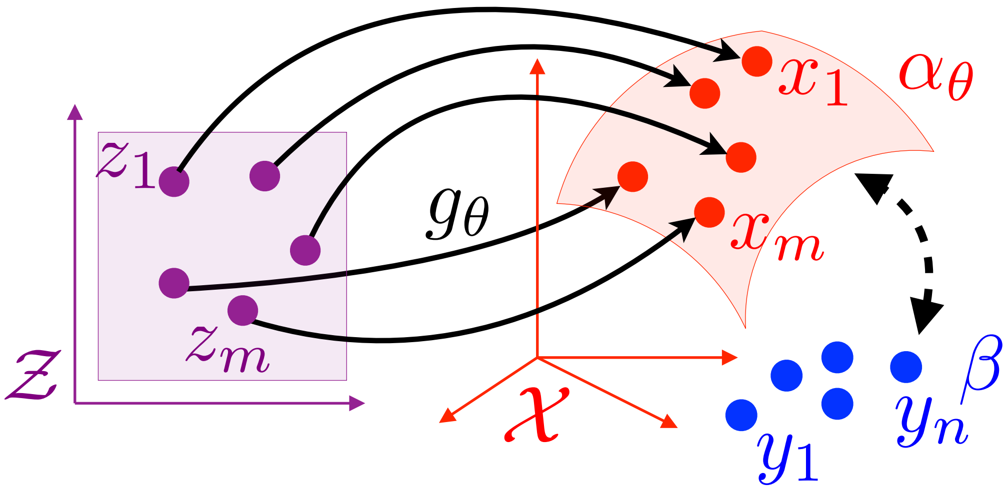
$$\min_\theta E(\theta) \overset{\text{def.}}{=} \bar{W}_\varepsilon(\alpha_\theta, \beta)$$

Stochastic gradient descent

$$\theta^{(\ell)} = \theta^{(\ell)} - \tau_\ell \nabla \hat{E}_L(\theta)$$

$$\hat{E}(\theta) \overset{\text{def.}}{=} \bar{W}_\varepsilon^L(\tfrac{1}{m} \textstyle\sum_i g_\theta(z_i), \beta)$$

Training Architecture

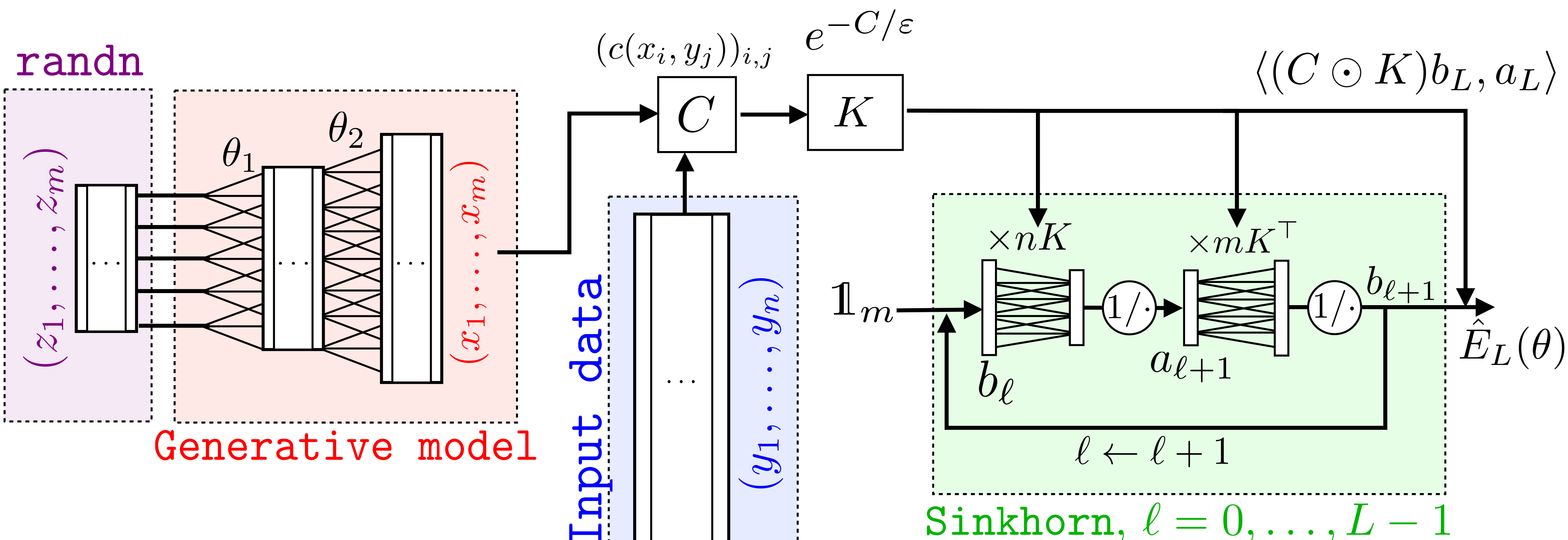


$$\min_\theta E(\theta) \overset{\text{def.}}{=} \bar{W}_\varepsilon(\alpha_\theta, \beta)$$

Stochastic gradient descent

$$\theta^{(\ell)} = \theta^{(\ell)} - \tau_\ell \nabla \hat{E}_L(\theta)$$

$$\hat{E}(\theta) \overset{\text{def.}}{=} \bar{W}_\varepsilon^L\left(\frac{1}{m}\sum_i g_\theta(z_i), \beta\right)$$



Automatic Differentiation

Setup: $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ computable in K operations.

```
def ForwardNN(A,b,Z):
    X = []
    X.append(Z)
    for r in arange(0,R):
        X.append( rhoF( A[r].dot(X[r]) + tile(b[r],[1,Z.shape[1]]) ) )
    return X
```

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \ldots)$ and their derivatives cost $O(1)$.

Question: What is the complexity of computing $\nabla\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^n$?

Automatic Differentiation

Setup: $\mathcal{E} : \mathbb{R}^n \rightarrow \mathbb{R}$ computable in K operations.

```
def ForwardNN(A,b,Z):
    X = []
    X.append(Z)
    for r in arange(0,R):
        X.append( rhoF( A[r].dot(X[r]) + tile(b[r],[1,Z.shape[1]]) ) )
    return X
```

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \ldots)$ and their derivatives cost $O(1)$.

Question: What is the complexity of computing $\nabla \mathcal{E} : \mathbb{R}^n \rightarrow \mathbb{R}^n$?

Finite differences:
$$\nabla \mathcal{E}(\theta) \approx \frac{1}{\varepsilon}(\mathcal{E}(\theta + \varepsilon\delta_1) - \mathcal{E}(\theta), \ldots \mathcal{E}(\theta + \varepsilon\delta_1) - \mathcal{E}(\theta))$$
$K(n+1)$ operations, intractable for large n.

Automatic Differentiation

Setup: $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ computable in K operations.

```
def ForwardNN(A,b,Z):
    X = []
    X.append(Z)
    for r in arange(0,R):
        X.append( rhoF( A[r].dot(X[r]) + tile(b[r],[1,Z.shape[1]]) ) )
    return X
```

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \ldots)$ and their derivatives cost $O(1)$.

Question: What is the complexity of computing $\nabla\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^n$?

Finite differences: $$\nabla\mathcal{E}(\theta) \approx \frac{1}{\varepsilon}(\mathcal{E}(\theta + \varepsilon\delta_1) - \mathcal{E}(\theta), \ldots \mathcal{E}(\theta + \varepsilon\delta_1) - \mathcal{E}(\theta))$$

$K(n+1)$ operations, intractable for large n.

Theorem: there is an algorithm to compute $\nabla\mathcal{E}$ in $O(K)$ operations. [Seppo Linnainmaa, 1970]

This algorithm is reverse mode automatic differentiation

```
def BackwardNN(A,b,X):
    gx = lossG(X[R],Y) # initialise the gradient
    for r in arange(R-1,-1,-1):
        M = rhoG( A[r].dot(X[r]) + tile(b[r],[1,n]) ) * gx
        gx = A[r].transpose().dot(M)
        gA[r] = M.dot(X[r].transpose())
        gb[r] = MakeCol(M.sum(axis=1))
    return [gA,gb]
```

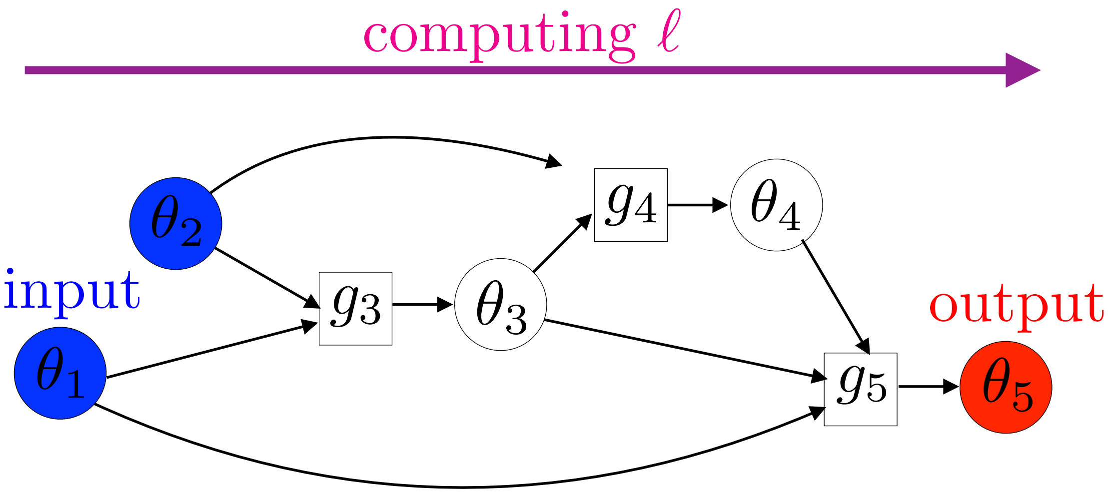
Computational Graph

Computational Graph

Computer program $\Leftrightarrow$ directed acyclic graph $\Leftrightarrow$ linear ordering of nodes $(\theta_r)_r$

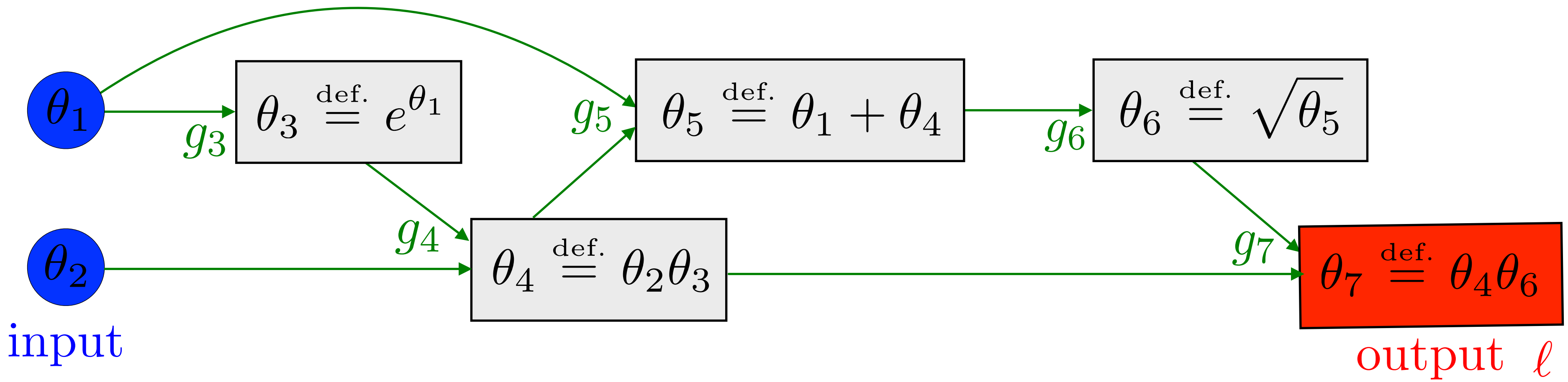
forward

function $\ell(\theta_1, \ldots, \theta_M)$
  for $r = M+1, \ldots, R$
    $\theta_r = g_r(\theta_{\text{Parents}(r)})$
  return θ_R



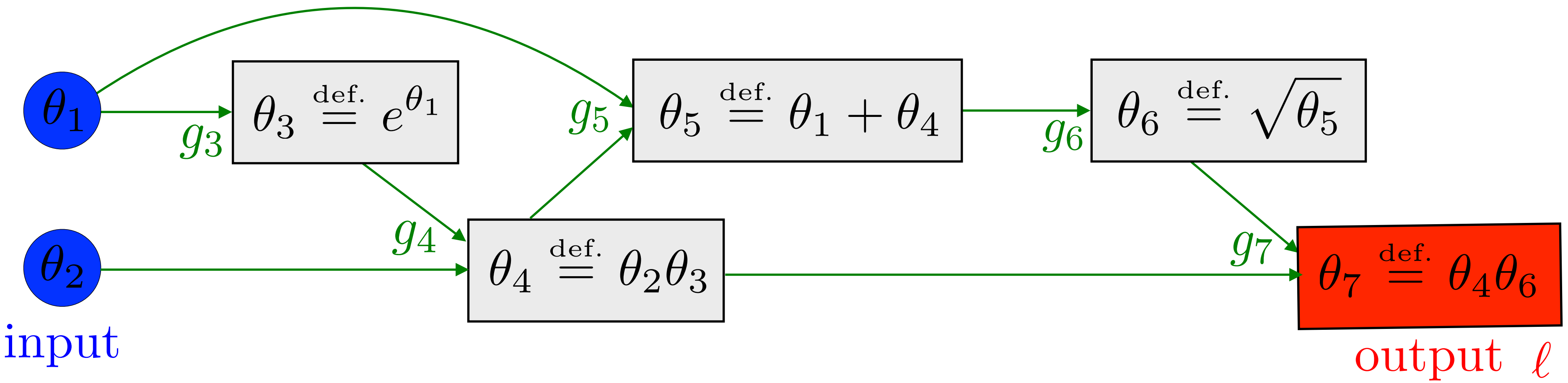
Example

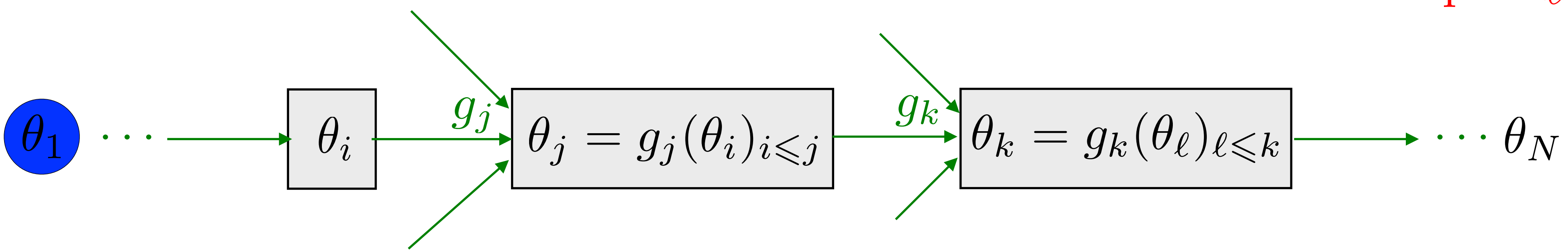
$$\ell(\theta_1, \theta_2) \stackrel{\text{def.}}{=} \theta_2 e^{\theta_1} \sqrt{\theta_1 + \theta_2 e^{\theta_1}}$$



Example

$$\ell(\theta_1, \theta_2) \overset{\text{def.}}{=} \theta_2 e^{\theta_1} \sqrt{\theta_1 + \theta_2 e^{\theta_1}}$$





Chain rules:

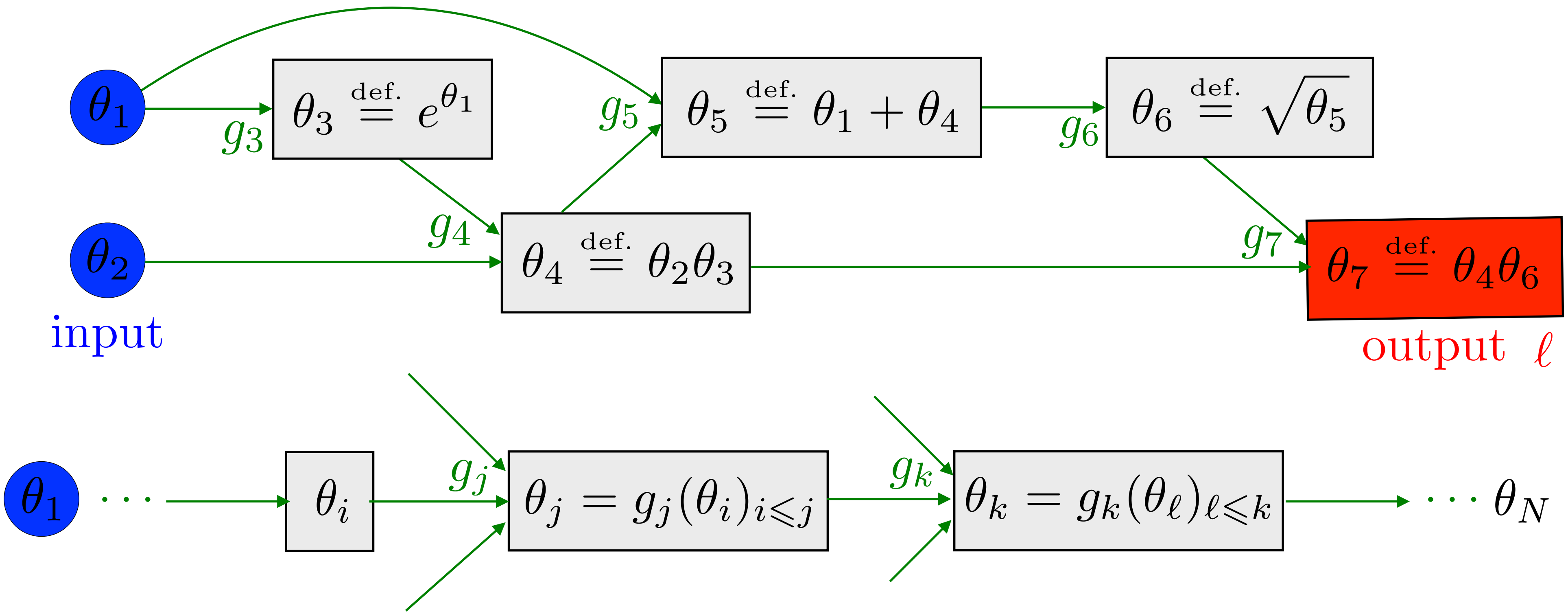
$$\text{“}\ \frac{\partial \theta_j}{\partial \theta_1} = \sum_{i \in \text{Parent}(j)} \frac{\partial \theta_j}{\partial \theta_i} \frac{\partial \theta_i}{\partial \theta_1}\ \text{”}$$

$\frac{\partial \theta_j}{\partial \theta_i} \rightarrow \partial_i g_j(\theta)$

“Classical” evaluation: **forward**.
Complexity $\sim$ #inputs.

Example

$$\ell(\theta_1, \theta_2) \overset{\text{def.}}{=} \theta_2 e^{\theta_1} \sqrt{\theta_1 + \theta_2 e^{\theta_1}}$$



Chain rules:

$$“\ \frac{\partial \theta_j}{\partial \theta_1} = \sum_{i \in \text{Parent}(j)} \frac{\partial \theta_j}{\partial \theta_i} \frac{\partial \theta_i}{\partial \theta_1}\ ”$$

$\frac{\partial \theta_j}{\partial \theta_i} \to \partial_i g_j(\theta)$

“Classical” evaluation: **forward**.
Complexity $\sim$ #inputs.

$$“\ \frac{\partial \theta_N}{\partial \theta_j} = \sum_{k \in \text{Child}(j)} \frac{\partial \theta_N}{\partial \theta_k} \frac{\partial \theta_k}{\partial \theta_j}\ ”$$

$\frac{\partial \theta_N}{\partial \theta_j} \to \nabla_j \ell(\theta)$

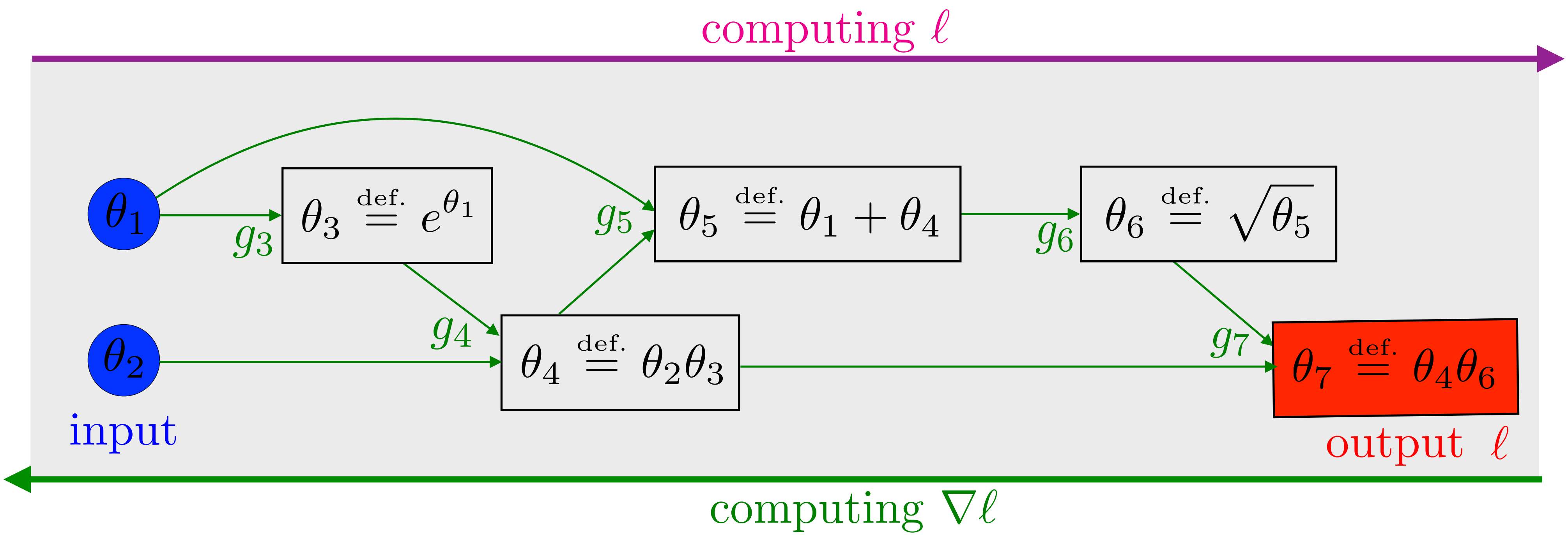
$\frac{\partial \theta_N}{\partial \theta_k} \to \nabla_k \ell(\theta)$

$\frac{\partial \theta_k}{\partial \theta_j} \to \partial_j g_k(\theta)$

Backward evaluation.
Complexity $\sim$ #outputs (1 for grad).

Backward Automatic Differentiation

$$\ell(\theta_1, \theta_2) \overset{\text{def.}}{=} \theta_2 e^{\theta_1} \sqrt{\theta_1 + \theta_2 e^{\theta_1}}$$



forward

function $\ell(\theta_1, \ldots, \theta_M)$
  for $r = M+1, \ldots, R$
    $\theta_r = g_r(\theta_{\text{Parents}(r)})$
  return θ_R

backward

function $\nabla\ell(\theta_1, \ldots, \theta_M)$
  $\nabla_R \ell = 1$
  for $r = R-1, \ldots, 1$
    $\nabla_r \ell = \sum_{s \in \text{Child}(r)} \partial_r g_s(\theta) \, \nabla_s \ell$
  return $(\nabla_1 \ell, \ldots, \nabla_M \ell)$

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

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