Numerical Optimal Transport

http://optimaltransport.github.io

Entropic Regularization

Gabriel Peyré

www.numerical-tours.com

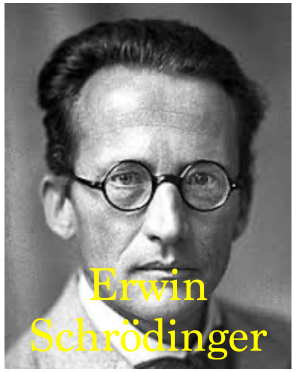
Overview

- Entropic Regularization and Sinkhorn
- Convergence Analysis
- Sinkhorn Divergences
- Generative Model Fitting

Entropic Regularization

Schrödinger's problem: [1931]

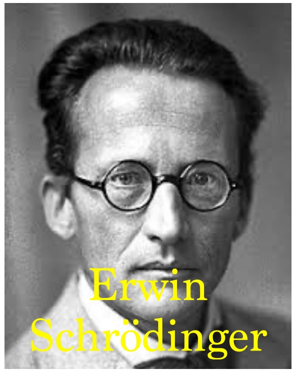
$$\min_{\mathbf{P}\mathbb{1}=\mathbf{a},\mathbf{P}^\top\mathbb{1}=\mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$

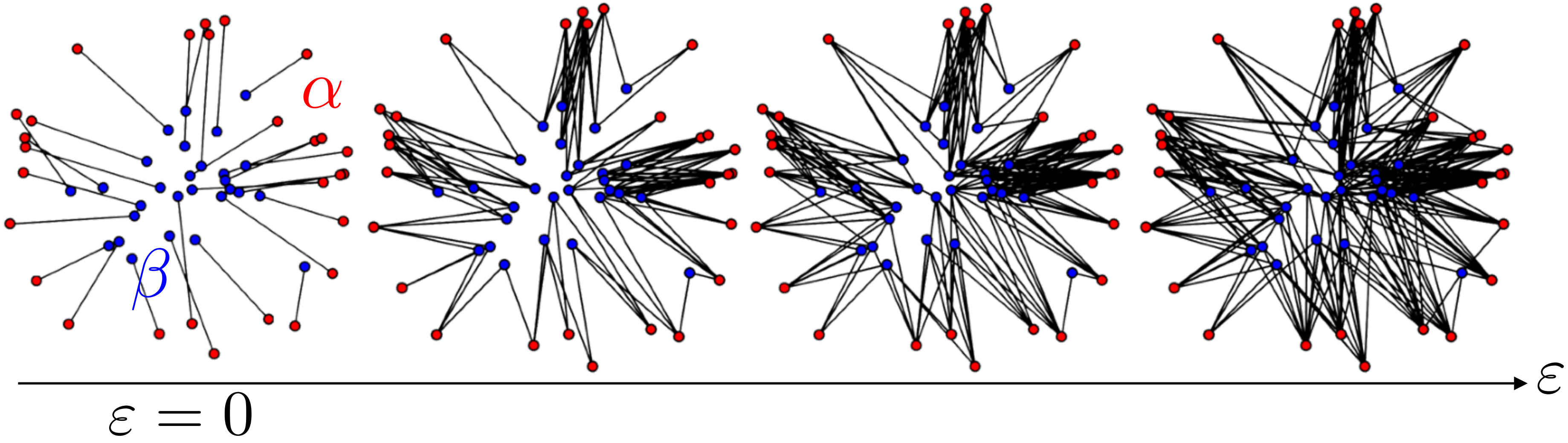


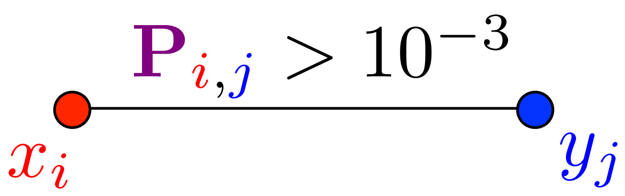
Entropic Regularization

Schrödinger's problem: [1931]

$$\min_{\mathbf{P}\mathbb{1}=\mathbf{a},\mathbf{P}^\top\mathbb{1}=\mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$







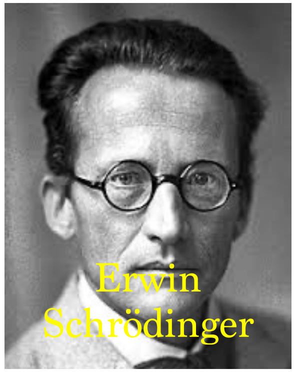
Entropic Regularization: General Case

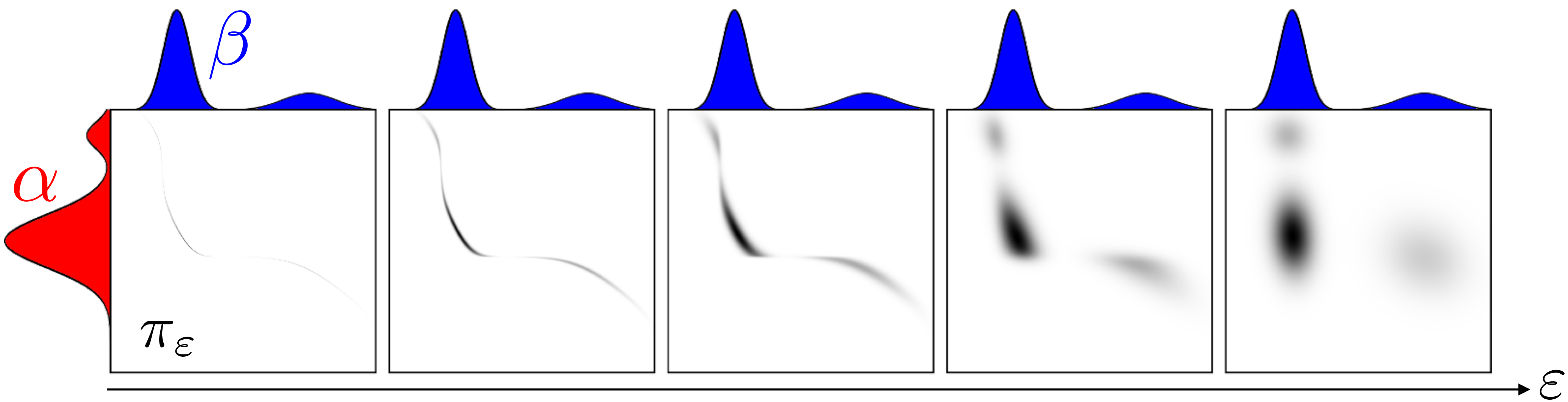
$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$

Relative-entropy: $\mathrm{KL}(\pi|\alpha \otimes \beta) \overset{\text{def.}}{=} \int_{\mathcal{X}^2} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}(x,y) \right) \mathrm{d}\pi(x,y)$

Schrödinger's problem: [1931]

$$\mathrm{W}^p_{\varepsilon,p}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha, \pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta)$$





Probabilistic Interpretation

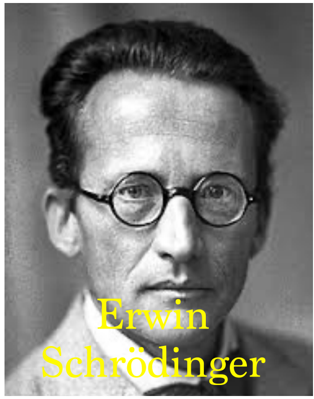
Relative-entropy: $\mathrm{KL}(\pi|\alpha \otimes \beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}^2} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y)\right)\mathrm{d}\pi(x,y)$

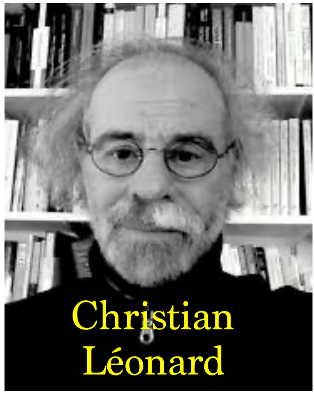
Schrödinger's problem: [1931]

$$W^p_{\varepsilon,p}(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta)$$

$$\min_{(X,Y)} \{\mathbb{E}(c(X,Y)) + \varepsilon \mathrm{I}(X,Y) \; ; \; X \sim \alpha, Y \sim \beta\}$$

Mutual information



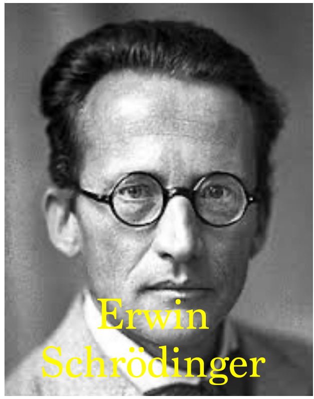


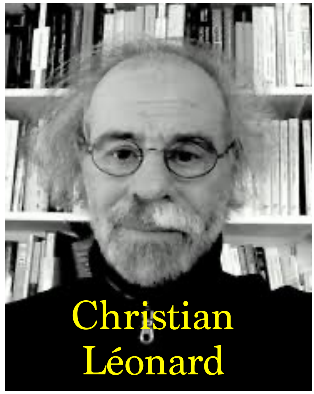
Probabilistic Interpretation

Relative-entropy: $\mathrm{KL}(\pi|\alpha \otimes \beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}^2} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y)\right)\mathrm{d}\pi(x,y)$

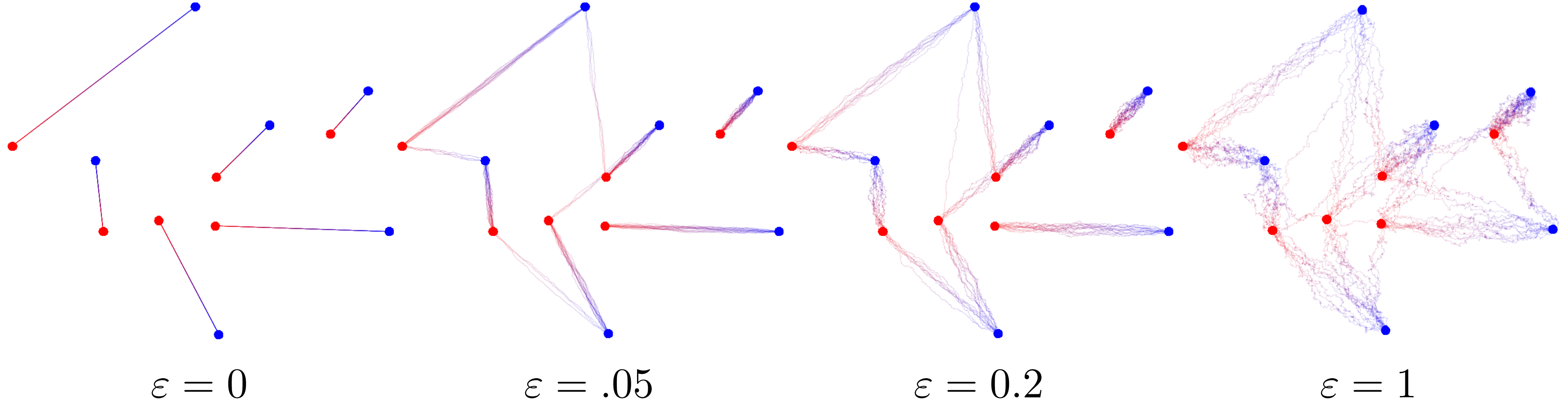
Schrödinger's problem: [1931]

$$W_{\varepsilon,p}^p(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\alpha\otimes\beta)$$

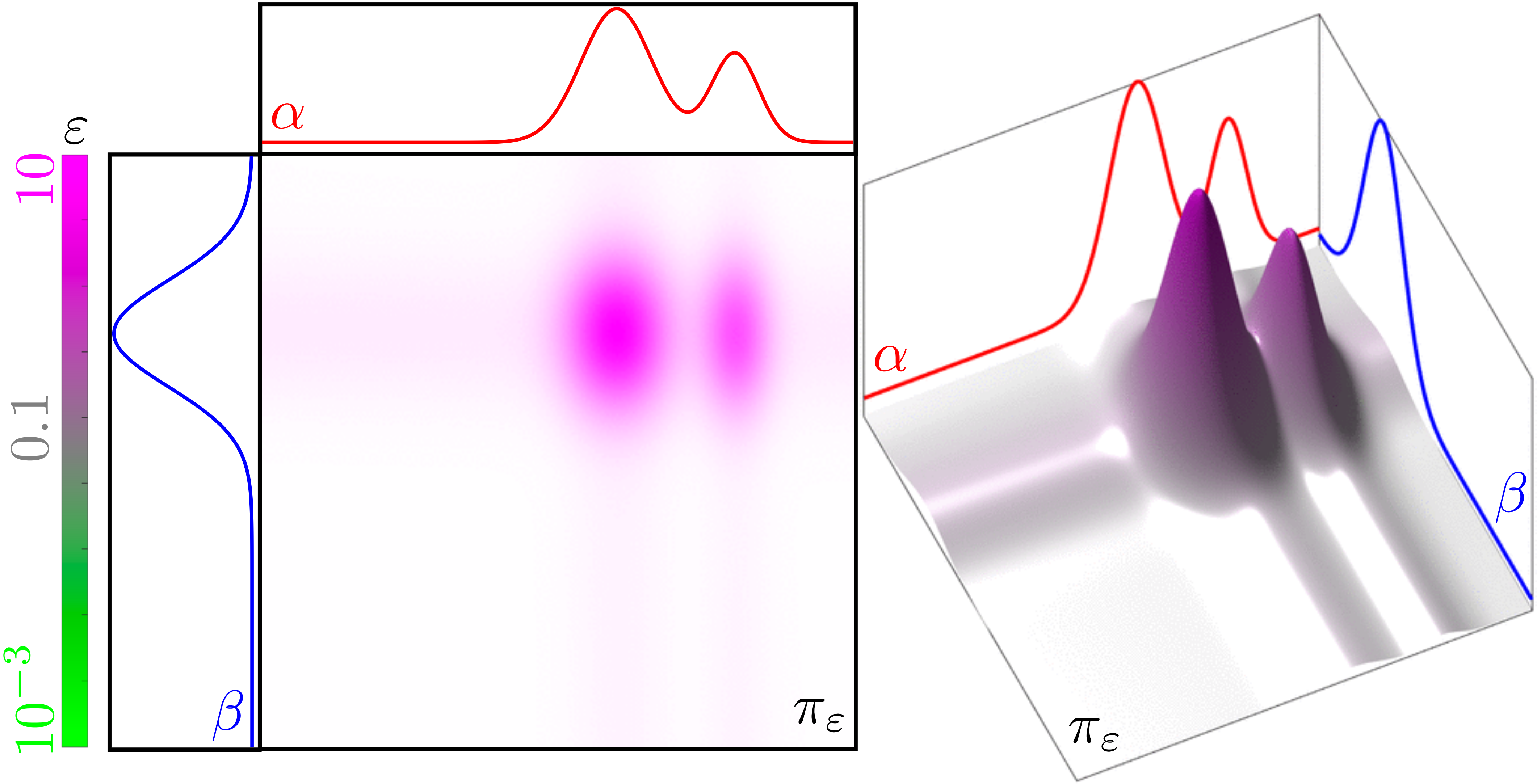




$$\min_{(X,Y)} \{\mathbb{E}(c(X,Y)) + \varepsilon \underbrace{\mathrm{I}(X,Y)}_{\text{Mutual information}} \; ; \; X \sim \alpha, Y \sim \beta\}$$



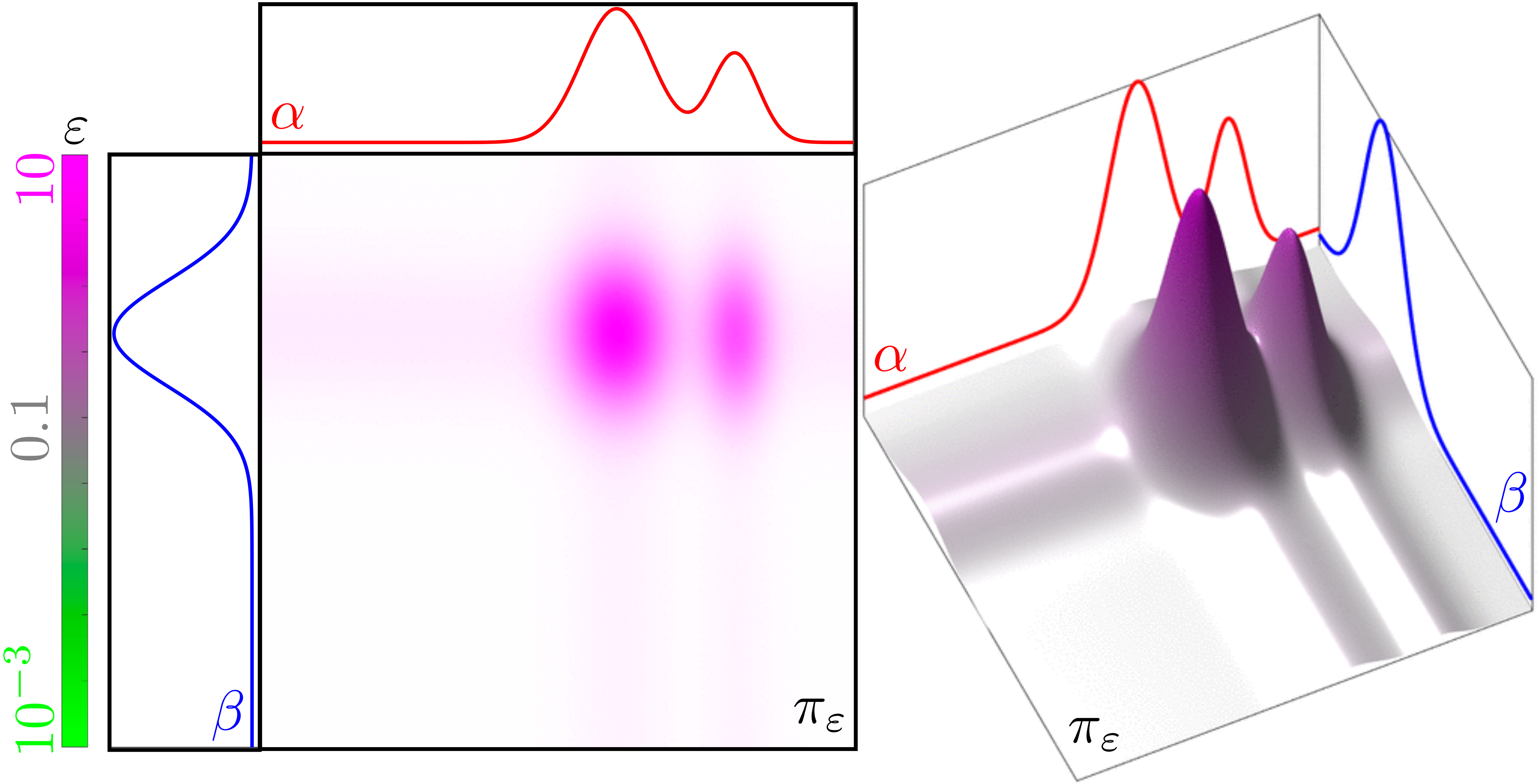
Impact of Regularization



$$\pi_\varepsilon = \operatorname{argmin}_\pi \left\{ \int_{\mathbb{R}^2} \left(\|x-y\|^2 + \varepsilon \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}(x,y) \right) \right) \mathrm{d}\pi(x,y) + \; ; \; \pi_1 = \alpha, \pi_2 = \beta \right\}$$

Theorem: $\pi_\varepsilon \xrightarrow{\varepsilon \to +\infty} \alpha \otimes \beta$ $\qquad \pi_\varepsilon \xrightarrow{\varepsilon \to 0} \pi_{\mathrm{OT}}$

Impact of Regularization



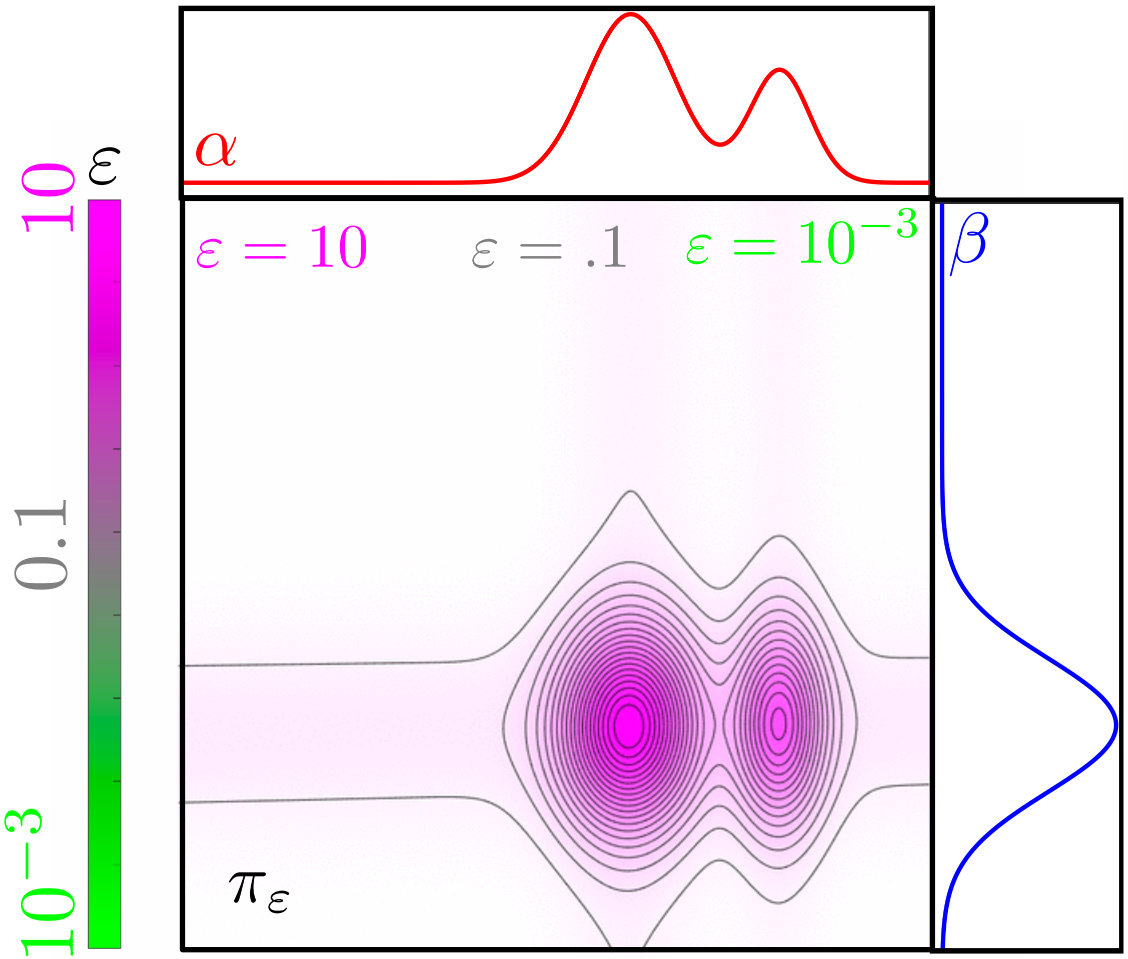
$$\pi_\varepsilon = \text{argmin}_\pi \left\{ \int_{\mathbb{R}^2} \left(\|x-y\|^2 + \varepsilon \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}(x,y) \right) \right) \mathrm{d}\pi(x,y) + \; ; \; \pi_1 = \alpha, \pi_2 = \beta \right\}$$

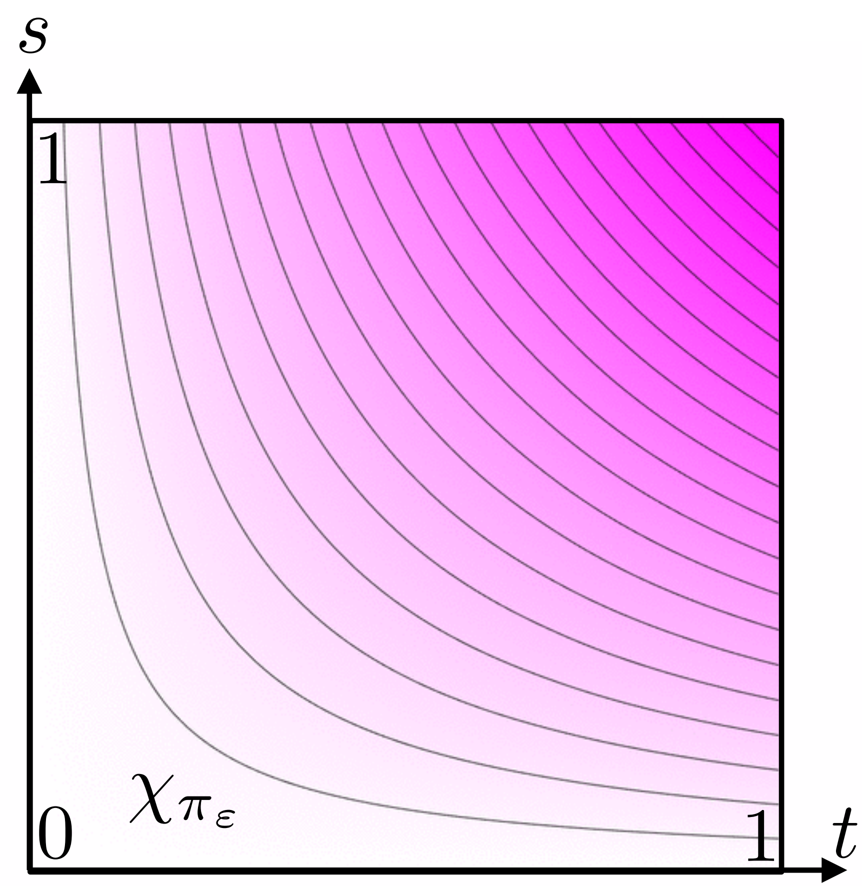
Theorem: $\pi_\varepsilon \xrightarrow{\varepsilon \to +\infty} \alpha \otimes \beta$ $\qquad$ $\pi_\varepsilon \xrightarrow{\varepsilon \to 0} \pi_{\mathrm{OT}}$

Impact of Regularization

Cumulative: $C_\pi(x,y) \overset{\text{def.}}{=} \int_{-\infty}^{x}\int_{-\infty}^{y} \mathrm{d}\pi(x,y)$

Copula: $\chi_\pi(s,t) \overset{\text{def.}}{=} C_\pi(C_\alpha^{-1}(s), C_\beta^{-1}(t))$





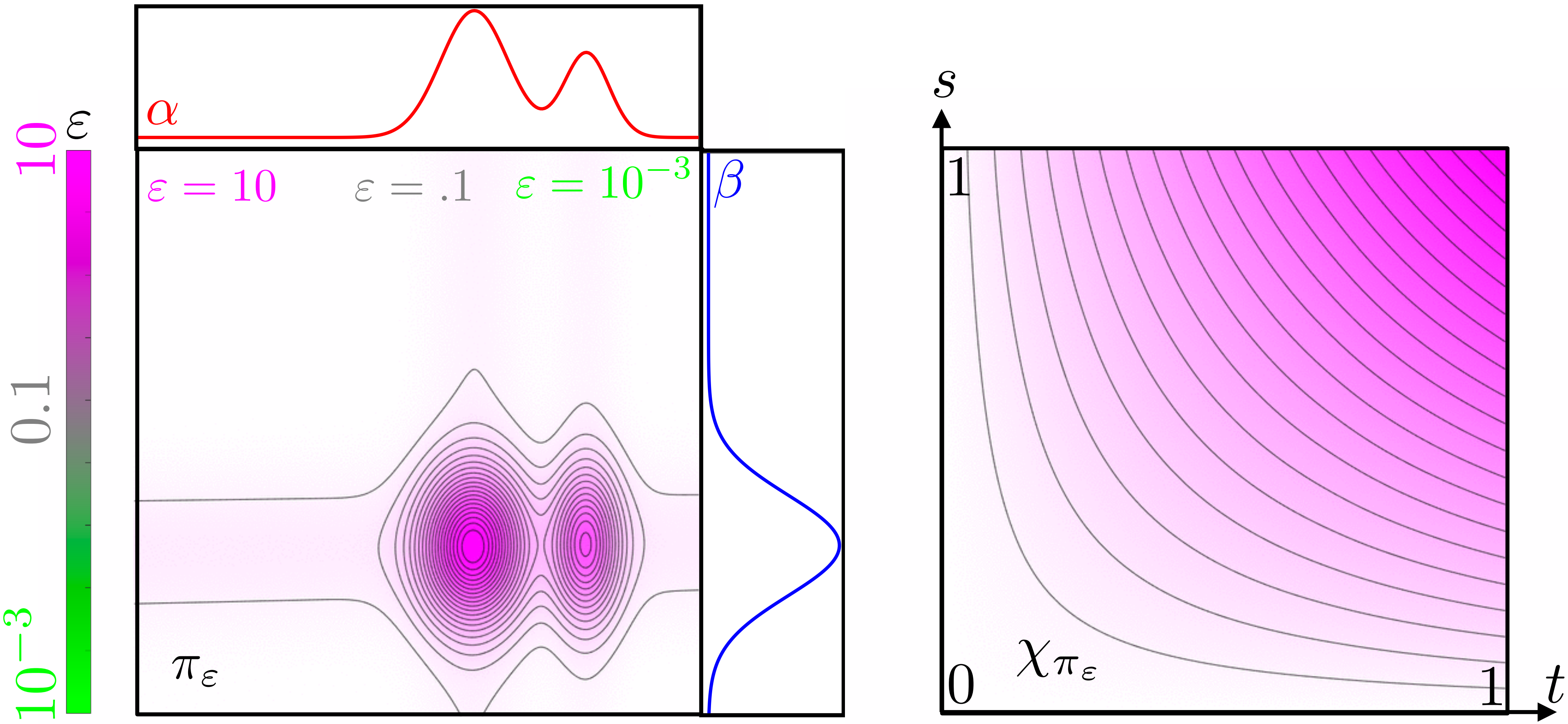
Theorem: $\chi_{\pi_\varepsilon}(s,t) \xrightarrow{\varepsilon \to 0} \min(s,t)$ (dependence)

$\chi_{\pi_\varepsilon}(s,t) \xrightarrow{\varepsilon \to +\infty} st$ (independence)

Impact of Regularization

Cumulative: $C_\pi(x,y) \overset{\text{def.}}{=} \int_{-\infty}^{x}\int_{-\infty}^{y} \mathrm{d}\pi(x,y)$

Copula: $\chi_\pi(s,t) \overset{\text{def.}}{=} C_\pi(C_\alpha^{-1}(s), C_\beta^{-1}(t))$



Theorem: $\chi_{\pi_\varepsilon}(s,t) \xrightarrow{\varepsilon \to 0} \min(s,t)$ (dependence)

$\chi_{\pi_\varepsilon}(s,t) \xrightarrow{\varepsilon \to +\infty} st$ (independence)

Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) \; ; \; \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b} \right\}$$

Proposition:

$$\mathbf{P} \text{ solution} \Leftrightarrow \begin{cases} \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \;\; \mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j \end{cases}$$

$$\mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$$

Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) \; ; \; \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b} \right\}$$

Proposition: $\mathbf{P}$ solution $\Leftrightarrow \begin{cases} \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \; \mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j \end{cases}$ $\quad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

$$\mathbf{P} = \operatorname{diag}(\mathbf{u}) \mathbf{K} \operatorname{diag}(\mathbf{v}) \implies \mathbf{a} = \mathbf{P}\mathbb{1} = \operatorname{diag}(\mathbf{u})(\mathbf{K}\mathbf{v}) = \mathbf{u} \odot (\mathbf{K}\mathbf{v})$$

Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

Col. constraint: $\mathbf{v} \odot (\mathbf{K}^\top \mathbf{u}) = \mathbf{b}$

Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) \;;\; \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b} \right\}$$

Proposition: $\mathbf{P}$ solution $\Leftrightarrow \begin{cases} \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \;\; \mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j \end{cases}$ $\quad \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

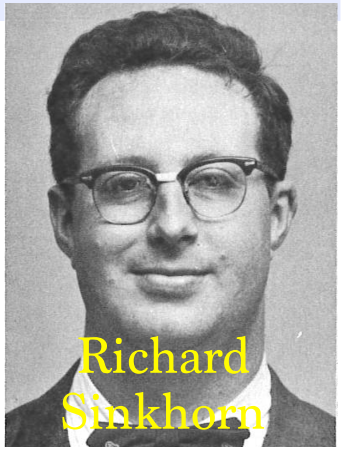
$$\mathbf{P} = \operatorname{diag}(\mathbf{u}) \mathbf{K} \operatorname{diag}(\mathbf{v}) \implies \mathbf{a} = \mathbf{P}\mathbb{1} = \operatorname{diag}(\mathbf{u})(\mathbf{K}\mathbf{v}) = \mathbf{u} \odot (\mathbf{K}\mathbf{v})$$

Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

Col. constraint: $\mathbf{v} \odot (\mathbf{K}^\top \mathbf{u}) = \mathbf{b}$

Sinkhorn iterations: $\mathbf{u} \leftarrow \dfrac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$ $\quad \mathbf{v} \leftarrow \dfrac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$

Theorem: [Sinkhorn 1964] $(\mathbf{u}, \mathbf{v})$ converges.



Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \textstyle\sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) \; ; \; \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b} \right\}$$

Proposition:

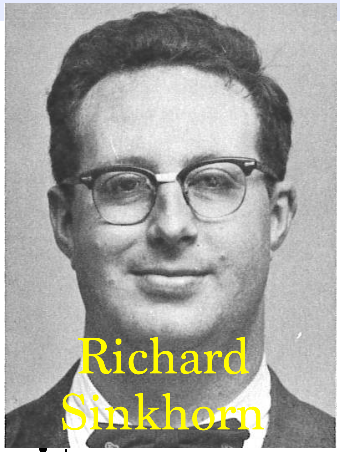
$$\mathbf{P} \text{ solution} \Leftrightarrow \begin{cases} \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top\mathbb{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \;\; \mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j \end{cases} \qquad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$$

$$\mathbf{P} = \operatorname{diag}(\mathbf{u}) \mathbf{K} \operatorname{diag}(\mathbf{v}) \implies \mathbf{a} = \mathbf{P}\mathbb{1} = \operatorname{diag}(\mathbf{u})(\mathbf{K}\mathbf{v}) = \mathbf{u} \odot (\mathbf{K}\mathbf{v})$$

Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

Col. constraint: $\mathbf{v} \odot (\mathbf{K}^\top \mathbf{u}) = \mathbf{b}$

Sinkhorn iterations: $\mathbf{u} \leftarrow \dfrac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$ $\quad \mathbf{v} \leftarrow \dfrac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$



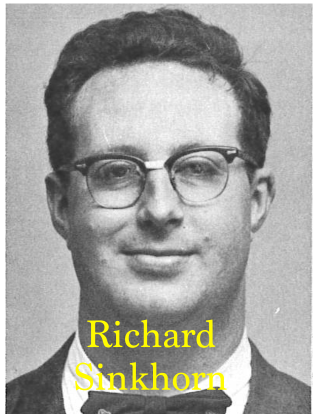
Theorem: [Sinkhorn 1964] $\quad (\mathbf{u}, \mathbf{v})$ converges.

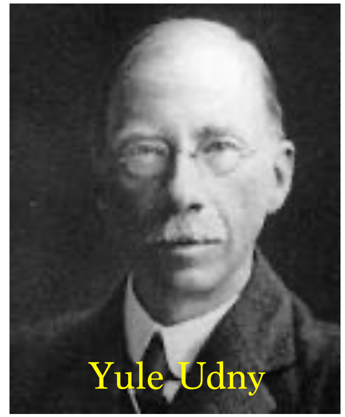
Matrix/vector multiplications: $\rightarrow O(n^2/\varepsilon^2)$ complexity.

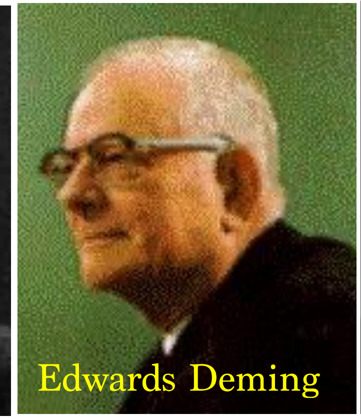
$\rightarrow$ Parallelizable on GPUs.

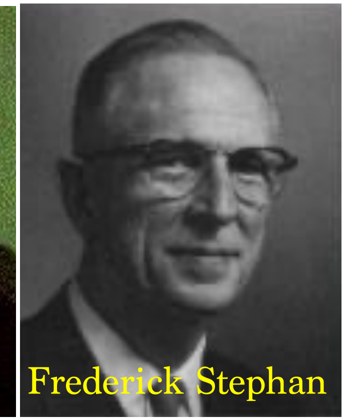
$\rightarrow$ Convolution on regular grids, separable kernels.

Sinkhorn, IPFP, RAS, …









Many names:

Sinkhorn algorithm

DAD scaling

Iterative proportional fitting

Biproportional fitting

RAS algorithm

Matrix scaling

Gravity model

Udny 1912

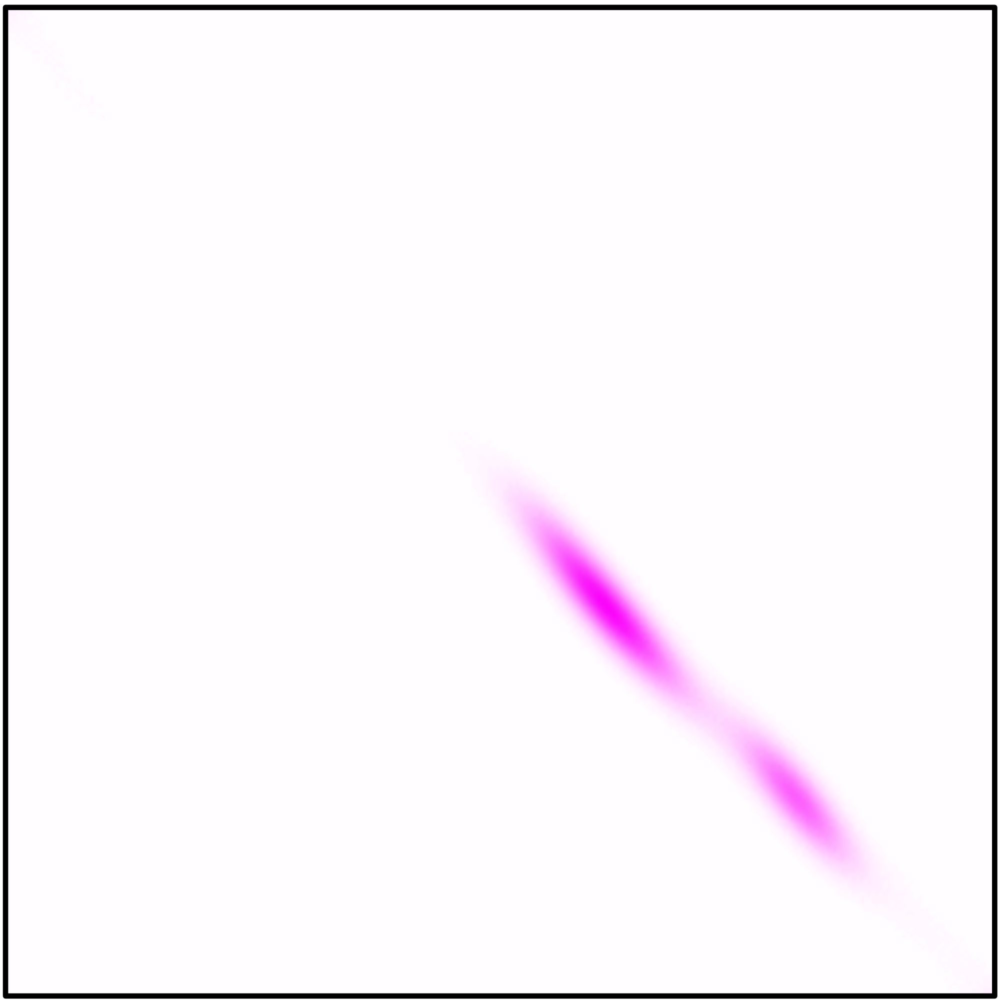
Kruithof, 1937

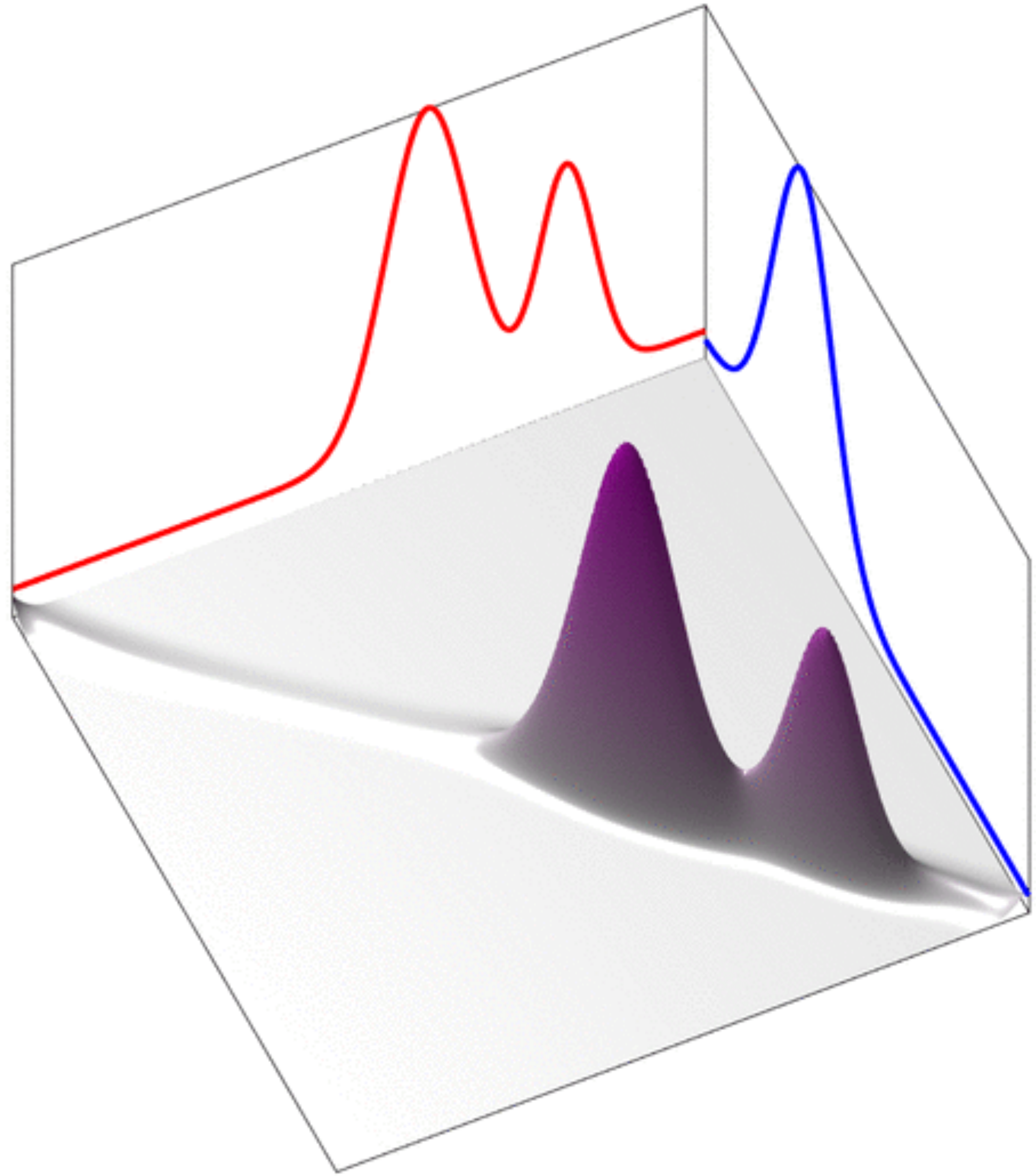
Deming and Stephan, 1940

Sinkhorn 1964

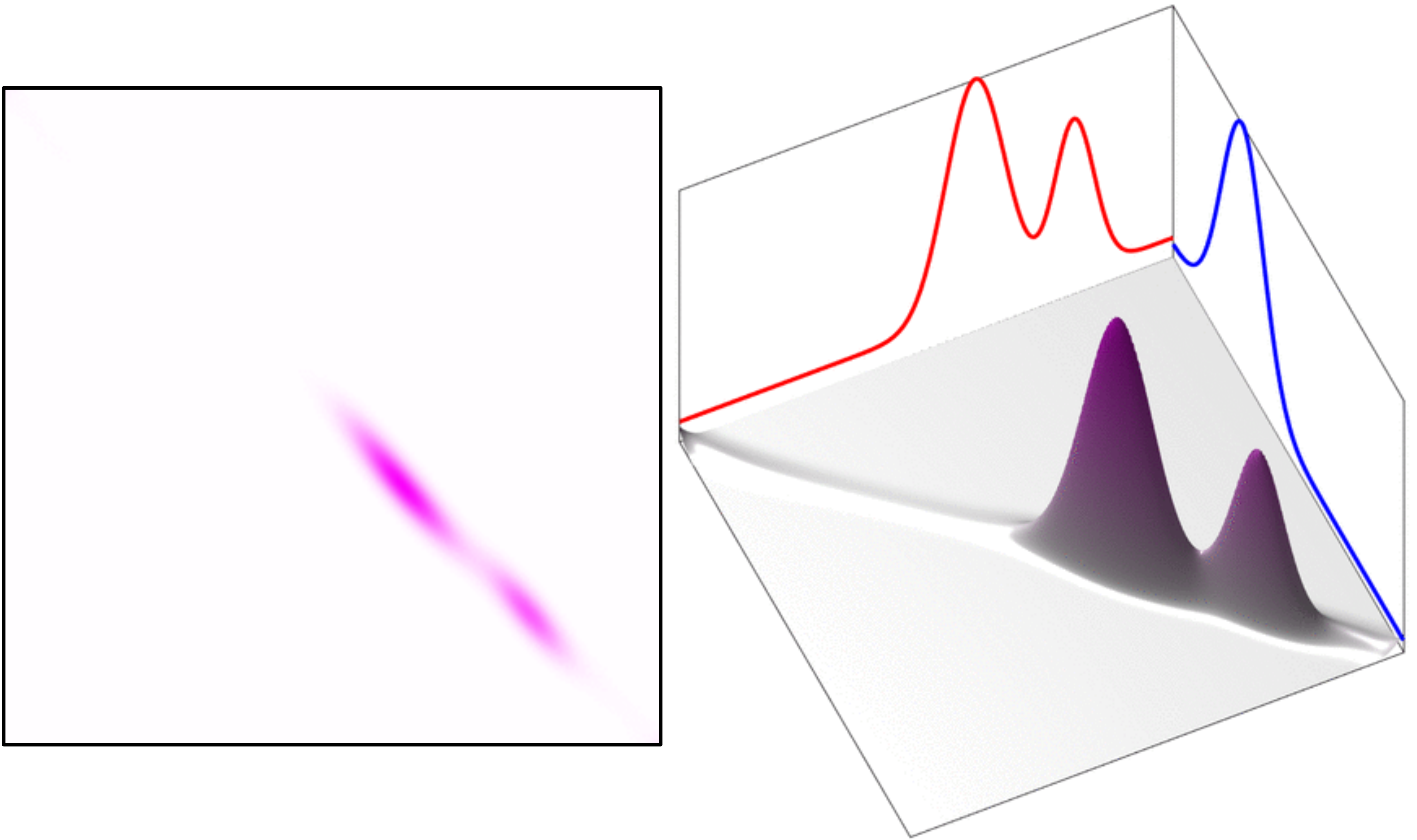
Special thanks to Tony Silveti-Falls for the picture of Sinkhorn

Sinkhorn Evolution



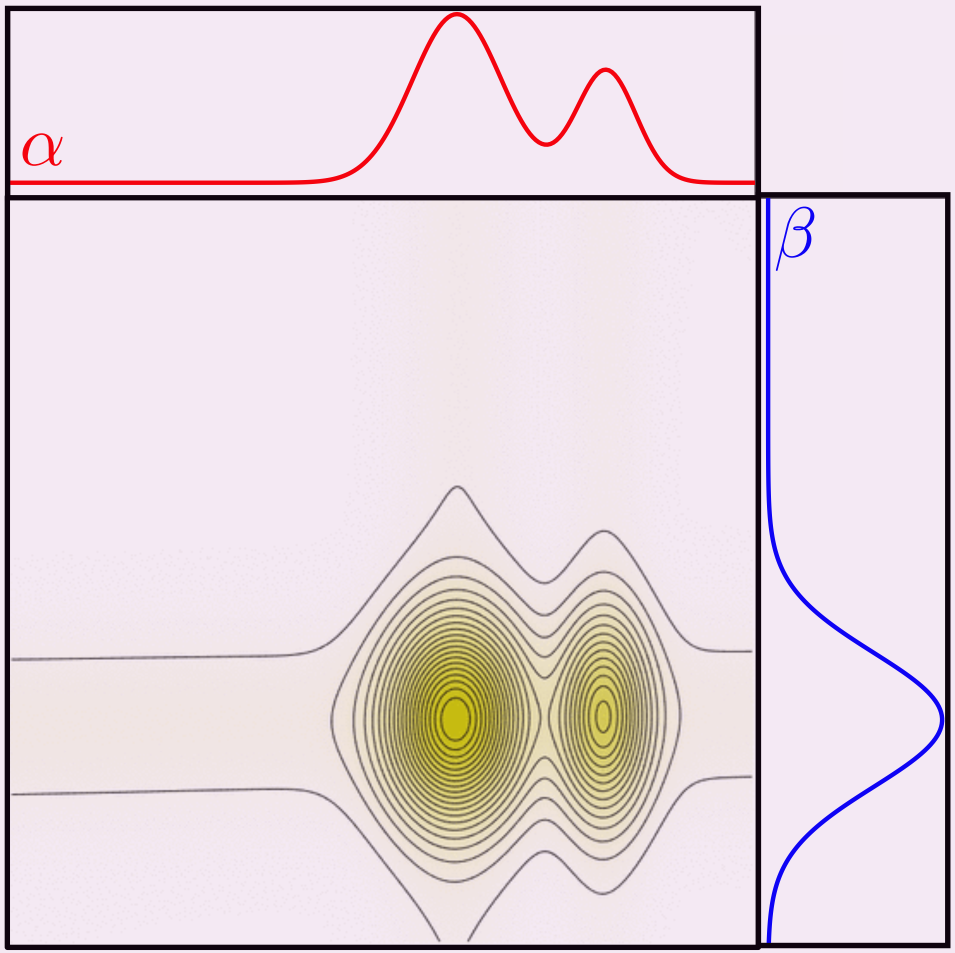


Sinkhorn Evolution



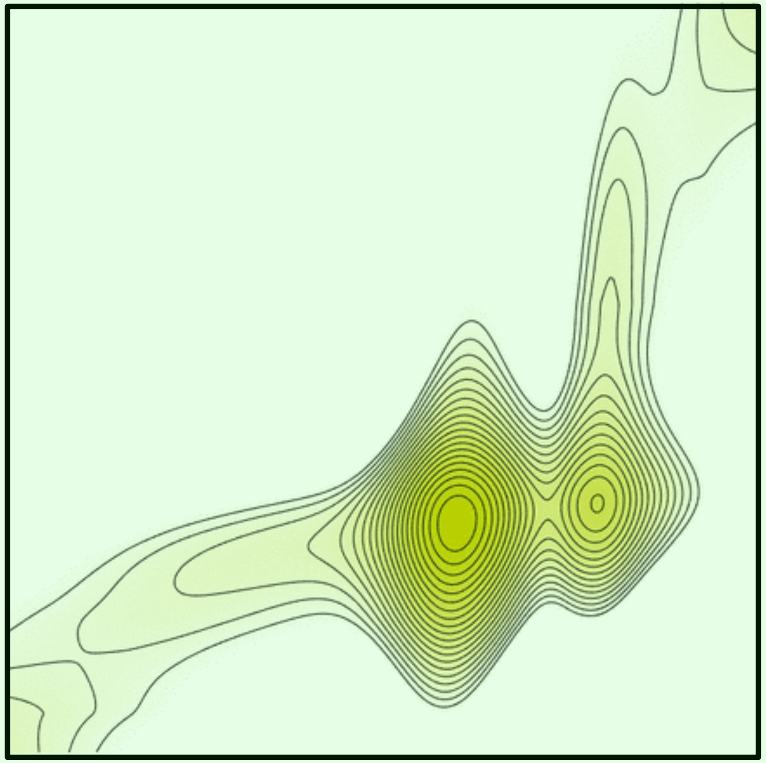
Other Regularizations

$$\min_{\pi} \left\{ \int_{\mathbb{R}^2} \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon R(\pi) \;;\; \pi_1 = \alpha, \pi_2 = \beta \right\}$$



$$R(\pi) = \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}\right) \mathrm{d}\pi(x,y)$$

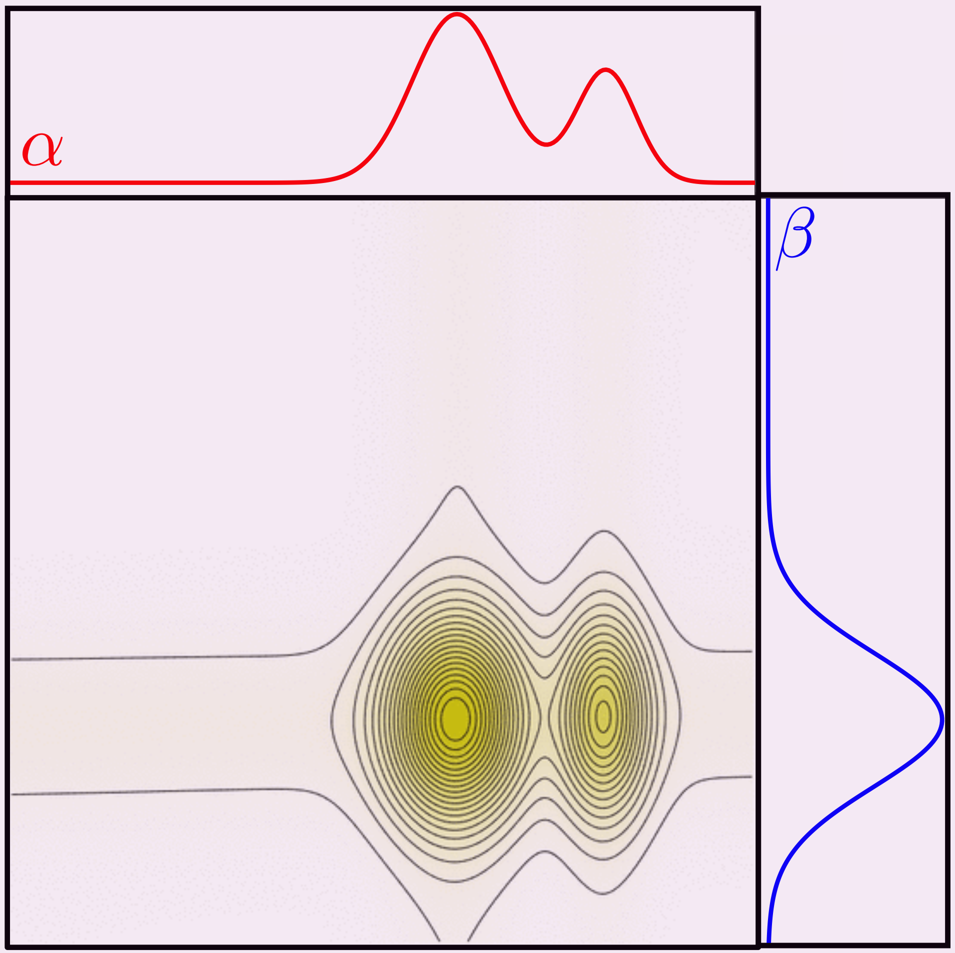
Dykstra's algorithm



$$R(\pi) = \int \left(\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}\right)^2 \mathrm{d}x\mathrm{d}y$$

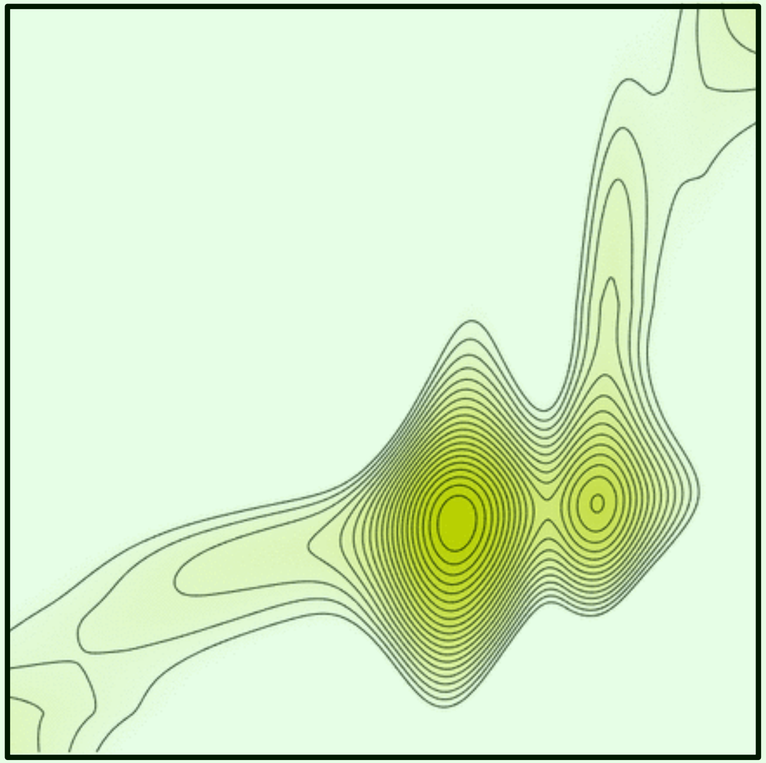
Other Regularizations

$$\min_{\pi} \left\{ \int_{\mathbb{R}^2} \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon R(\pi) \; ; \; \pi_1 = \alpha, \pi_2 = \beta \right\}$$



$$R(\pi) = \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}\right) \mathrm{d}\pi(x,y)$$

Dykstra's algorithm



$$R(\pi) = \int \left(\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}\right)^2 \mathrm{d}x\mathrm{d}y$$

Extension: Unbalanced OT

$$\mathrm{W}_p^{\tau,p}(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi} \int d^p \mathrm{d}\pi + \tau \mathrm{KL}(\pi_1|\alpha) + \tau \mathrm{KL}(\pi_2|\beta)$$

[Liereo, Mielke, Savaré 2015]

See also: [Chizat, Schmitzer, Peyré, Vialard 2015]
[Kondratyev, Monsaingeon, Vorotnikov 2015]

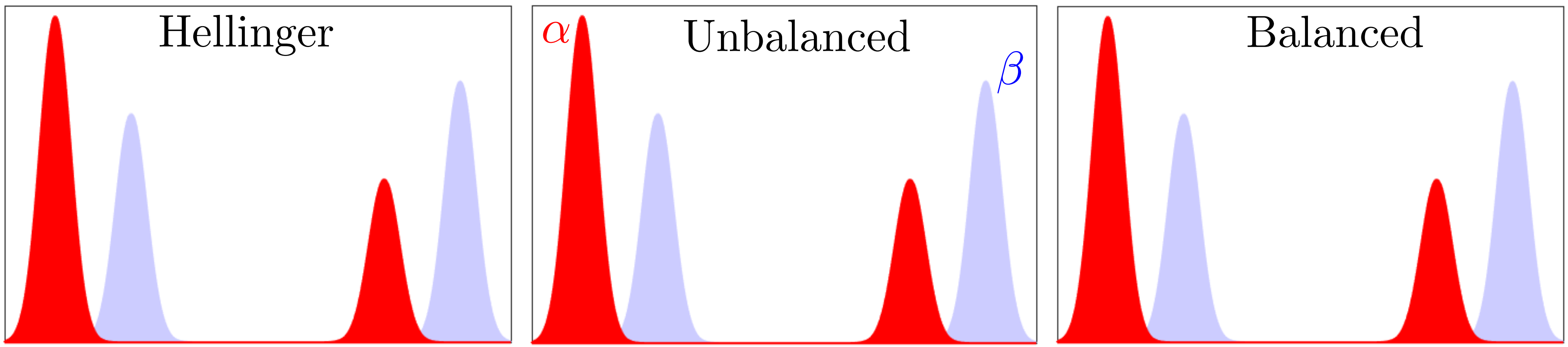
Extension: Unbalanced OT

$$\mathrm{W}_p^{\tau,p}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi} \int d^p \mathrm{d}\pi + \tau \mathrm{KL}(\pi_1|\alpha) + \tau \mathrm{KL}(\pi_2|\beta)$$

[Liereo, Mielke, Savaré 2015]

See also: [Chizat, Schmitzer, Peyré, Vialard 2015]
[Kondratyev, Monsaingeon, Vorotnikov 2015]

$$\int (\sqrt{\alpha} - \sqrt{\beta})^2 \xleftarrow{\tau \to 0} \mathrm{W}_p^{\tau,p}(\alpha,\beta) \xrightarrow{\tau \to +\infty} \mathrm{W}_p^p(\alpha,\beta)$$

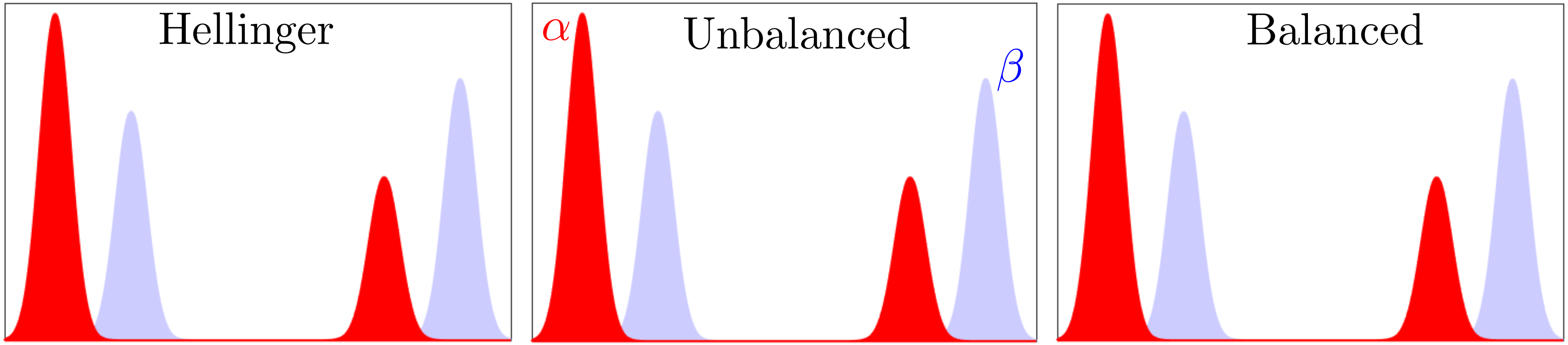


Extension: Unbalanced OT

$$\mathrm{W}_p^{\tau,p}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi} \int d^p \mathrm{d}\pi + \tau \mathrm{KL}(\pi_1|\alpha) + \tau \mathrm{KL}(\pi_2|\beta)$$

[Liereo, Mielke, Savaré 2015]   See also: [Chizat, Schmitzer, Peyré, Vialard 2015]
[Kondratyev, Monsaingeon, Vorotnikov 2015]

$$\int (\sqrt{\alpha} - \sqrt{\beta})^2 \xleftarrow{\tau \to 0} \mathrm{W}_p^{\tau,p}(\alpha,\beta) \xrightarrow{\tau \to +\infty} \mathrm{W}_p^p(\alpha,\beta)$$



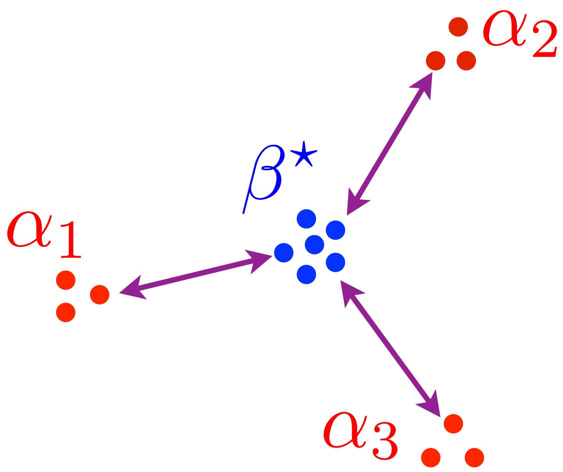
$$\mathrm{W}_{\varepsilon,p}^{\tau,p}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi} \int d^p \mathrm{d}\pi + \tau \mathrm{KL}(\pi_1|\alpha) + \tau \mathrm{KL}(\pi_2|\beta) + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta)$$

Sinkhorn's algorithm: $\mathbf{u} \leftarrow \left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}\right)^{1+\frac{\varepsilon}{\tau}} \longleftrightarrow \mathbf{v} \leftarrow \left(\frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}\right)^{1+\frac{\varepsilon}{\tau}}$

Extension: Wasserstein Barycenters

Barycenters of measures $(\alpha_s)_s$: $\quad \sum_s \lambda_s = 1$

$$\beta^\star \in \underset{\beta}{\operatorname{argmin}} \sum_s \lambda_s \mathrm{W}_p^p(\alpha_s, \beta)$$

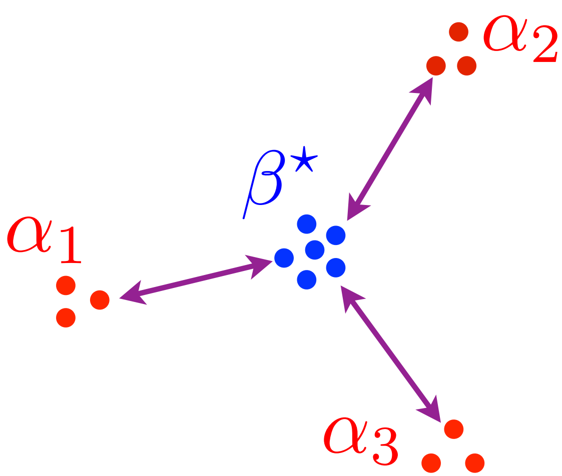


[Solomon et al, SIGGRAPH 2015]

Extension: Wasserstein Barycenters

Barycenters of measures $(\alpha_s)_s$: $\quad \sum_s \lambda_s = 1$

$$\beta^\star \in \underset{\beta}{\operatorname{argmin}} \sum_s \lambda_s \mathrm{W}_p^p(\alpha_s, \beta)$$



[Solomon et al, SIGGRAPH 2015]

Sinkhorn's algorithm:

$$\left(\mathbf{u}_s \leftarrow \frac{\mathbf{a}_s}{\mathbf{K}\mathbf{v}_s} \right)_s \quad \rightleftarrows \quad \left(\mathbf{v}_s \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}_s} \right)_s$$

$$\boxed{\mathbf{b} \leftarrow \prod_s (\mathbf{K}^\top \mathbf{u}_s)^{\lambda_s}}$$

Overview

- Entropic Regularization and Sinkhorn
- **Convergence Analysis**
- Sinkhorn Divergences
- Generative Model Fitting

Bregman Iterative Projections

$$\langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon \, \mathrm{KL}(\mathbf{P} | \mathbf{a} \otimes \mathbf{b}) = \varepsilon \, \mathrm{KL}(\mathbf{P} | \mathbf{K}) + \mathrm{cst} \quad \text{where} \quad \mathbf{K}_{i,j} = e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}} \mathbf{a}_i \mathbf{b}_j$$

Shrödinger problem: $\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \mathrm{KL}(\mathbf{P} | \mathbf{K})$

Constraints : $\mathcal{C}^1_{\mathbf{a}} \cup \mathcal{C}^2_{\mathbf{b}}$

$$\mathcal{C}^1_{\mathbf{a}} \overset{\text{def.}}{=} \{\mathbf{P} \ : \ \mathbf{P}\mathbb{1}_m = \mathbf{a}\}$$

$$\mathcal{C}^2_{\mathbf{b}} \overset{\text{def.}}{=} \left\{\mathbf{P} \ : \ \mathbf{P}^{\mathrm{T}}\mathbb{1}_m = \mathbf{b}\right\}$$

Bregman Iterative Projections

$$\langle \mathbf{P}, \mathbf{C}\rangle + \varepsilon\,\mathrm{KL}(\mathbf{P}|\mathbf{a}\otimes\mathbf{b}) = \varepsilon\,\mathrm{KL}(\mathbf{P}|\mathbf{K}) + \mathrm{cst} \quad \text{where} \quad \mathbf{K}_{i,j} = e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}\mathbf{a}_i\mathbf{b}_j$$

Shrödinger problem: $\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathrm{KL}(\mathbf{P}|\mathbf{K})$

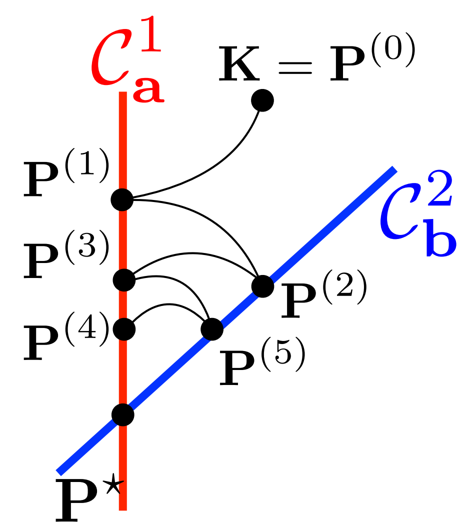
Constraints: $\mathcal{C}^1_{\mathbf{a}} \cup \mathcal{C}^2_{\mathbf{b}}$

$\mathcal{C}^1_{\mathbf{a}} \overset{\text{def.}}{=} \{\mathbf{P} \,:\, \mathbf{P}\mathbb{1}_m = \mathbf{a}\}$ $\quad$ $\mathcal{C}^2_{\mathbf{b}} \overset{\text{def.}}{=} \left\{\mathbf{P} \,:\, \mathbf{P}^{\mathrm{T}}\mathbb{1}_m = \mathbf{b}\right\}$

[Bregman, 1967]

Iterative projections: $\mathbf{P}^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\mathcal{C}^1_{\mathbf{a}}}(\mathbf{P}^{(\ell)})$ and $\mathbf{P}^{(\ell+2)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\mathcal{C}^2_{\mathbf{b}}}(\mathbf{P}^{(\ell+1)})$

Theorem: For affine $(\mathcal{C}^1_{\mathbf{a}}, \mathcal{C}^2_{\mathbf{b}})$, $\mathbf{P}^{(\ell)} \rightarrow \mathbf{P}^\star = \underset{\mathbf{P}\in\mathcal{C}^1_{\mathbf{a}}\cap\mathcal{C}^1_{\mathbf{b}}}{\mathrm{argmin}}\ \mathrm{KL}(\mathbf{P}|\mathbf{K})$



Bregman Iterative Projections

$$\langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon \, \mathrm{KL}(\mathbf{P} | \mathbf{a} \otimes \mathbf{b}) = \varepsilon \, \mathrm{KL}(\mathbf{P} | \mathbf{K}) + \mathrm{cst} \quad \text{where} \quad \mathbf{K}_{i,j} = e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}} \mathbf{a}_i \mathbf{b}_j$$

Shrödinger problem: $\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \mathrm{KL}(\mathbf{P} | \mathbf{K})$

Constraints : $\mathcal{C}^1_{\mathbf{a}} \cup \mathcal{C}^2_{\mathbf{b}}$

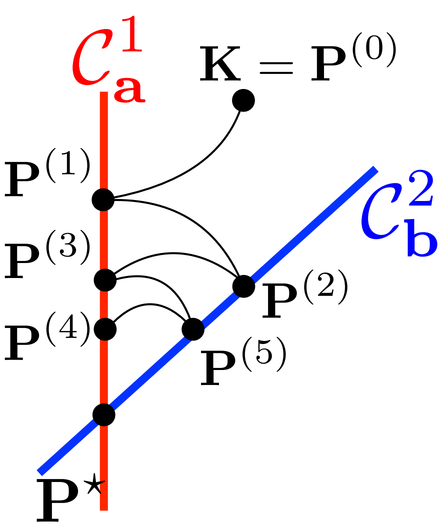
$$\mathcal{C}^1_{\mathbf{a}} \stackrel{\text{def.}}{=} \{ \mathbf{P} \ : \ \mathbf{P} \mathbb{1}_m = \mathbf{a} \}$$

$$\mathcal{C}^2_{\mathbf{b}} \stackrel{\text{def.}}{=} \left\{ \mathbf{P} \ : \ \mathbf{P}^{\mathrm{T}} \mathbb{1}_m = \mathbf{b} \right\}$$

[Bregman, 1967]

Iterative projections: $\mathbf{P}^{(\ell+1)} \stackrel{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\mathcal{C}^1_{\mathbf{a}}}(\mathbf{P}^{(\ell)})$ and $\mathbf{P}^{(\ell+2)} \stackrel{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\mathcal{C}^2_{\mathbf{b}}}(\mathbf{P}^{(\ell+1)})$

Theorem: For affine $(\mathcal{C}^1_{\mathbf{a}}, \mathcal{C}^2_{\mathbf{b}})$, $\mathbf{P}^{(\ell)} \rightarrow \mathbf{P}^\star = \underset{\mathbf{P} \in \mathcal{C}^1_{\mathbf{a}} \cap \mathcal{C}^1_{\mathbf{b}}}{\mathrm{argmin}} \ \mathrm{KL}(\mathbf{P} | \mathbf{K})$



Sinkhorn $\iff$ iterative projections.

$$\mathbf{P}^{(2\ell)} \stackrel{\text{def.}}{=} \mathrm{diag}(\mathbf{u}^{(\ell)}) \mathbf{K} \, \mathrm{diag}(\mathbf{v}^{(\ell)}), \qquad \mathbf{P}^{(2\ell+1)} \stackrel{\text{def.}}{=} \mathrm{diag}(\mathbf{u}^{(\ell+1)}) \mathbf{K} \, \mathrm{diag}(\mathbf{v}^{(\ell)})$$

Dual Analysis: Alternate Maximization

Dual problem:

$$\mathrm{W}_p^\varepsilon(\alpha, \beta)^p \overset{\text{def.}}{=} \sup_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \int f \mathrm{d}\alpha + \int g \mathrm{d}\beta + \varepsilon \int_{\mathcal{X}^2} (1 - e^{\frac{-d^p + f\oplus g}{\varepsilon}}) \mathrm{d}\alpha \otimes \mathrm{d}\beta$$

Primal-dual relations: $\mathrm{d}\pi^\star(x,y) = e^{\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$

Dual Analysis: Alternate Maximization

Dual problem:

$$\mathrm{W}_p^\varepsilon(\alpha,\beta)^p \overset{\text{def.}}{=} \sup_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \int f\mathrm{d}\alpha + \int g\mathrm{d}\beta + \varepsilon\int_{\mathcal{X}^2}(1-e^{\frac{-d^p+f\oplus g}{\varepsilon}})\mathrm{d}\alpha\otimes\mathrm{d}\beta$$

Primal-dual relations: $\mathrm{d}\pi^\star(x,y) = e^{\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$

Soft min: $\min_\alpha^\varepsilon(h) \overset{\text{def.}}{=} -\varepsilon\log\int_{\mathcal{X}} e^{-h/\varepsilon}\mathrm{d}\alpha \quad \xrightarrow{\varepsilon\to 0} \min_{\mathrm{supp}(\alpha)} h$

Soft c-transforms:

$$f^{c,\varepsilon}(y) \overset{\text{def.}}{=} \min_\alpha^\varepsilon(d^p(\cdot,y)-f)$$

$$g^{c,\varepsilon}(x) \overset{\text{def.}}{=} \min_\beta^\varepsilon(d^p(x,\cdot)-g)$$

Dual Analysis: Alternate Maximization

Dual problem:

$$W_p^\varepsilon(\alpha,\beta)^p \overset{\text{def.}}{=} \sup_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \int f\,\mathrm{d}\alpha + \int g\,\mathrm{d}\beta + \varepsilon \int_{\mathcal{X}^2} \left(1 - e^{\frac{-d^p + f\oplus g}{\varepsilon}}\right)\mathrm{d}\alpha\otimes\mathrm{d}\beta$$

Primal-dual relations: $\mathrm{d}\pi^\star(x,y) = e^{\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$

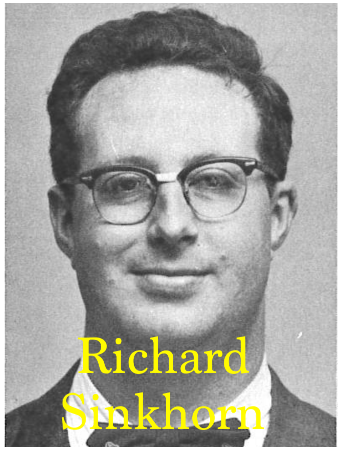
Soft min: $\min_\alpha{}^\varepsilon(h) \overset{\text{def.}}{=} -\varepsilon\log\int_{\mathcal{X}} e^{-h/\varepsilon}\mathrm{d}\alpha \quad \xrightarrow{\varepsilon\to 0} \quad \min_{\mathrm{supp}(\alpha)} h$

Soft c-transforms:

$$f^{c,\varepsilon}(y) \overset{\text{def.}}{=} \min_\alpha{}^\varepsilon(d^p(\cdot,y) - f)$$

$$g^{c,\varepsilon}(x) \overset{\text{def.}}{=} \min_\beta{}^\varepsilon(d^p(x,\cdot) - g)$$

Sinkhorn's algorithm: $\boxed{f_{k+1} \overset{\text{def.}}{=} (g_k)^{c,\varepsilon}}$ $\rightleftarrows$ $\boxed{g_{k+1} \overset{\text{def.}}{=} (f_{k+1})^{c,\varepsilon}}$



Proposition: $f_0 = 0 \leqslant f_1 \leqslant f_2 \leqslant \ldots$ If c is bounded, $f_k \to f^\star$.

[Robert Fortet 1938]

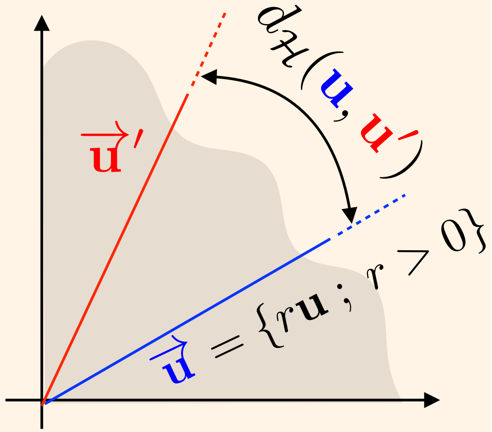
Hilbert Projective Metric

Hilbert's projective metric: $\forall\, (\mathbf{u}, \mathbf{u}') \in (\mathbb{R}^n_{+,*})^2$

$$d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') \stackrel{\text{def.}}{=} \|\log(\mathbf{u}) - \log(\mathbf{u}')\|_V$$

$$\|f\|_V \stackrel{\text{def.}}{=} \max(f) - \min(f)$$

$d_{\mathcal{H}}$ is a distance on the set of rays $\overrightarrow{\mathbf{u}}$.



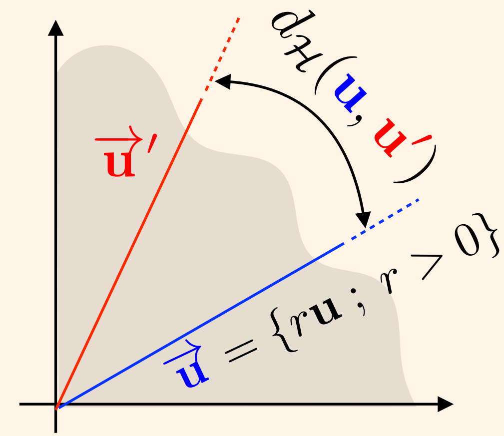
Hilbert Projective Metric

Hilbert's projective metric: $\quad \forall\, (\mathbf{u}, \mathbf{u}') \in (\mathbb{R}^n_{+,*})^2$

$$d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') \stackrel{\text{def.}}{=} \| \log(\mathbf{u}) - \log(\mathbf{u}') \|_V$$

$$\|f\|_V \stackrel{\text{def.}}{=} \max(f) - \min(f)$$

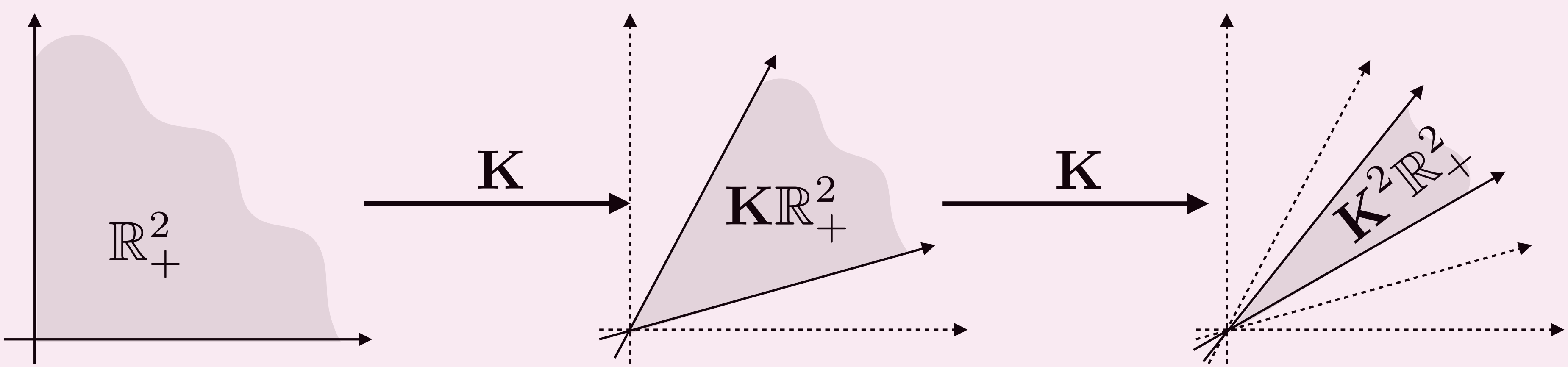
$d_{\mathcal{H}}$ is a distance on the set of rays $\overrightarrow{\mathbf{u}}$.



Birkhoff's contraction theorem:

Theorem 1.1. Let $\mathbf{K} \in \mathbb{R}^{n \times m}_{+,*}$, then for $(\mathbf{v}, \mathbf{v}') \in (\mathbb{R}^m_{+,*})^2$

$$d_{\mathcal{H}}(\mathbf{K}\mathbf{v}, \mathbf{K}\mathbf{v}') \leq \lambda(\mathbf{K}) d_{\mathcal{H}}(\mathbf{v}, \mathbf{v}') \text{ where } \begin{cases} \lambda(\mathbf{K}) \stackrel{\text{def.}}{=} \frac{\sqrt{\eta(\mathbf{K})}-1}{\sqrt{\eta(\mathbf{K})}+1} < 1 \\ \eta(\mathbf{K}) \stackrel{\text{def.}}{=} \max\limits_{i,j,k,\ell} \frac{\mathbf{K}_{i,k}\mathbf{K}_{j,\ell}}{\mathbf{K}_{j,k}\mathbf{K}_{i,\ell}}. \end{cases}$$



Perron Frobenius

Simplex: $\Sigma_k = \{p \in \mathbb{R}_+^k \;;\; \sum_i p_i = 1\}$

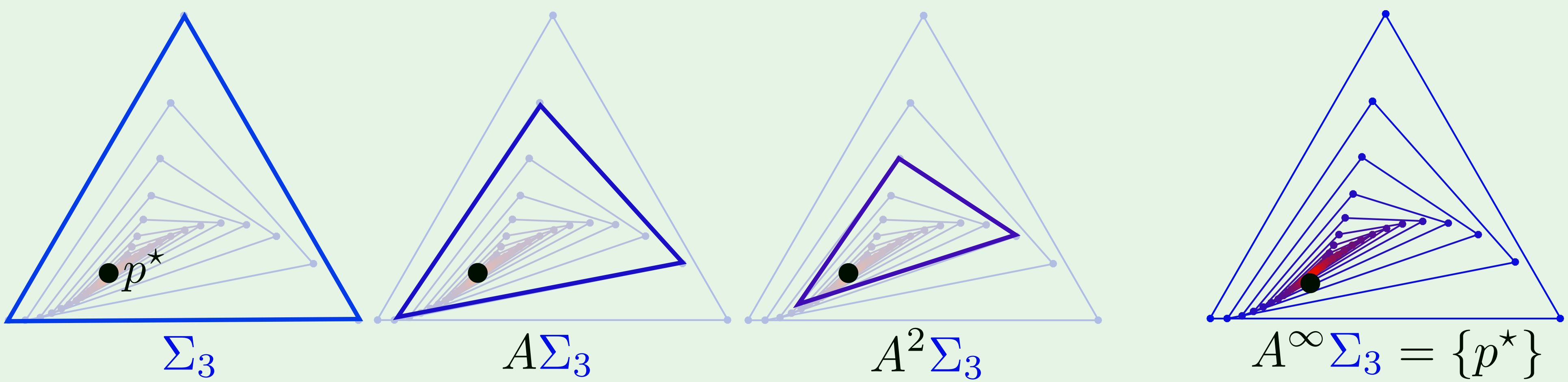
Stochastic matrix: $A \in \mathbb{R}_+^n,\; A^\top \mathbb{1}_k = \mathbb{1}_k$

$A : \Sigma_k \to \Sigma_k$

Theorem: [Perron-Frobenius]

If $A > 0,\; \exists! p^\star,\; Ap^\star = p^\star$.

$\exists \rho \in [0,1[,\; \|A^k p - p^\star\| \leqslant \rho^k$



Sinkhorn under Hilbert's Metric

Sinkhorn iterations: $\mathbf{u}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}}$ and $\mathbf{v}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}^{(\ell+1)}}$

[Franklin and Lorenz, 1989]

Theorem: One has $(\mathbf{u}^{(\ell)}, \mathbf{v}^{(\ell)}) \to (\mathbf{u}^\star, \mathbf{v}^\star)$

$$d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) = O(\lambda(\mathbf{K})^{2\ell}), \quad d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star) = O(\lambda(\mathbf{K})^{2\ell}).$$

$$d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) \leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell)}\mathbb{1}_m, \mathbf{a})}{1-\lambda(\mathbf{K})^2} \qquad d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star) \leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell),\top}\mathbb{1}_n, \mathbf{b})}{1-\lambda(\mathbf{K})^2}$$

$$\|\log(\mathbf{P}^{(\ell)}) - \log(\mathbf{P}^\star)\|_\infty \leq d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) + d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star)$$

Hilbert Metric Analysis

Dual cost: $\mathrm{W}^{(k)}(\alpha, \beta) \overset{\text{def.}}{=} \int f_k \mathrm{d}\alpha + \int g_k \mathrm{d}\beta$

Theorem: $|\mathrm{W}^{(k)}(\alpha, \beta) - \mathrm{W}_p^{\varepsilon}(\alpha, \beta)^p| \leqslant C(1 - e^{-\frac{\|d\|_\infty^p}{\varepsilon}})^k$

Fast in term of k ... slow in term of ε
$\longrightarrow$ useless to approximate W_p

Hilbert Metric Analysis

Dual cost: $\mathrm{W}^{(k)}(\alpha, \beta) \stackrel{\text{def.}}{=} \int f_k \mathrm{d}\alpha + \int g_k \mathrm{d}\beta$

Theorem: $|\mathrm{W}^{(k)}(\alpha, \beta) - \mathrm{W}_p^\varepsilon(\alpha, \beta)^p| \leqslant C(1 - e^{-\frac{\|d\|_\infty^p}{\varepsilon}})^k$

Fast in term of k ... slow in term of ε

$\longrightarrow$ useless to approximate W_p

Proof:

Variation semi-norm: $\|h\|_V \triangleq \sup(f) - \inf(f)$

Birkhoff's contraction theorem:

$f \mapsto f^{c,\varepsilon}$ is contractant for $\|\cdot\|_V$

$\Longrightarrow \quad \|f_k - f^\star\|_V = O(1 - e^{-\frac{\|d^p\|_\infty}{\varepsilon}})^k$

Hilbert vs Mirror Analysis

$$\textit{Theorem:} \qquad |\mathrm{W}^{(k)}(\alpha, \beta) - \mathrm{W}_p^{\varepsilon}(\alpha, \beta)^p| \leqslant \begin{cases} C(1 - e^{-\frac{\|d\|_\infty^p}{\varepsilon}})^k \\ \dfrac{\|d\|_\infty^{2p}}{\varepsilon k} \end{cases}$$

Hilbert vs Mirror Analysis

Theorem: $$|\mathrm{W}^{(k)}(\alpha,\beta) - \mathrm{W}_p^{\varepsilon}(\alpha,\beta)^p| \leqslant \begin{cases} C(1-e^{-\frac{\|d\|_\infty^p}{\varepsilon}})^k \\ \frac{\|d\|_\infty^{2p}}{\varepsilon k} \end{cases}$$

[Altschuler et al 2017]

$$\Delta_k \overset{\text{def.}}{=} \mathrm{W}_{\varepsilon,p}^p(\alpha,\beta) - \mathrm{W}^{(k)}(\alpha,\beta)$$

$$\begin{aligned} \Delta_k - \Delta_{k+1} &= \varepsilon \mathrm{KL}(\alpha|(\pi_k)_1) + \varepsilon \mathrm{KL}(\beta|(\pi_k)_2) \\ &\geqslant \varepsilon\|\alpha - (\pi_k)_1\|_1^2 + \varepsilon\|\beta - (\pi_k)_2\|_1^2 \quad \text{(Pinsker)} \\ &\geqslant \frac{\varepsilon}{\|d\|_\infty^{2p}} \Delta_k^2 \qquad (\|f_k\|_V, \|g_k\|_V \leqslant \|d\|_\infty^p) \end{aligned}$$

Hilbert vs Mirror Analysis

Theorem: $$|\mathrm{W}^{(k)}(\alpha,\beta) - \mathrm{W}_p^{\varepsilon}(\alpha,\beta)^p| \leqslant \begin{cases} C\left(1 - e^{-\frac{\|d\|_\infty^p}{\varepsilon}}\right)^k \\ \dfrac{\|d\|_\infty^{2p}}{\varepsilon k} \end{cases}$$

[Altschuler et al 2017]

$$\Delta_k \stackrel{\text{def.}}{=} \mathrm{W}_{\varepsilon,p}^p(\alpha,\beta) - \mathrm{W}^{(k)}(\alpha,\beta)$$

$$\begin{aligned} \Delta_k - \Delta_{k+1} &= \varepsilon \mathrm{KL}(\alpha|(\pi_k)_1) + \varepsilon \mathrm{KL}(\beta|(\pi_k)_2) \\ &\geqslant \varepsilon\|\alpha - (\pi_k)_1\|_1^2 + \varepsilon\|\beta - (\pi_k)_2\|_1^2 \quad \text{(Pinsker)} \\ &\geqslant \frac{\varepsilon}{\|d\|_\infty^{2p}}\Delta_k^2 \qquad (\|f_k\|_V, \|g_k\|_V \leqslant \|d\|_\infty^p) \end{aligned}$$

for $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$

Proposition: $$|\mathrm{W}_p^{\varepsilon,p} - \mathrm{W}_p^p| \leqslant \varepsilon \log(n)$$

$$\varepsilon = \frac{\delta}{\log(n)}$$

$|\mathrm{W}^{(k)} - \mathrm{W}_p^p| \leqslant \delta$ in $n^2 \log(n)\|d\|_\infty^{2p}/\varepsilon^2$ operations

Overview

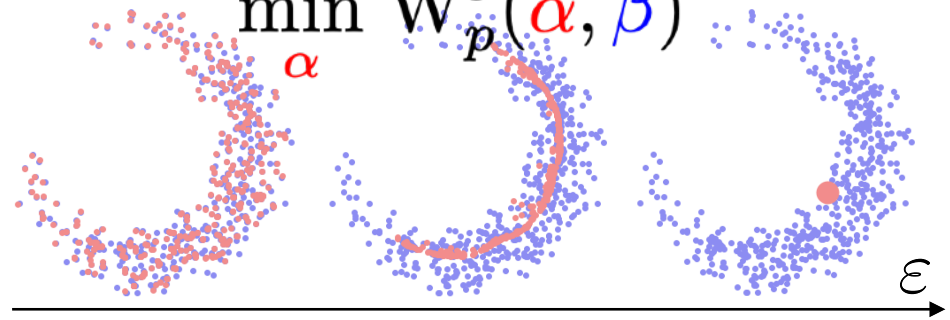
- Entropic Regularization and Sinkhorn
- Convergence Analysis
- Sinkhorn Divergences
- Generative Model Fitting

Kernel norms and MMDs

$$\mathrm{W}_p^\varepsilon(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi_1 = \alpha, \pi_2 = \beta} \int_{\mathcal{X}^2} d(x, y)^p \mathrm{d}\pi(x, y) + \varepsilon \, \mathrm{KL}(\pi | \alpha \otimes \beta)$$

Problem: $\mathrm{W}_p^\varepsilon(\alpha, \alpha) \neq 0$

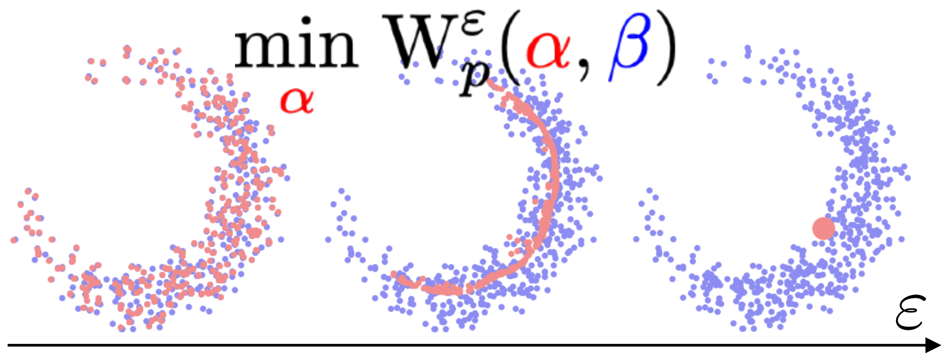
$$\min_{\alpha} \mathrm{W}_p^\varepsilon(\alpha, \beta)$$



Kernel norms and MMDs

$$\mathrm{W}_p^\varepsilon(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha, \pi_2=\beta} \int_{\mathcal{X}^2} d(x, y)^p \mathrm{d}\pi(x, y) + \varepsilon \, \mathrm{KL}(\pi | \alpha \otimes \beta)$$

Problem: $\mathrm{W}_p^\varepsilon(\alpha, \alpha) \neq 0$



Prop.: $\pi^{(\varepsilon)} \xrightarrow{\varepsilon \to +\infty} \alpha \otimes \beta$

$\mathrm{W}_p^\varepsilon(\alpha, \beta)^p \xrightarrow{\varepsilon \to +\infty} -\langle \alpha, \beta \rangle_k$

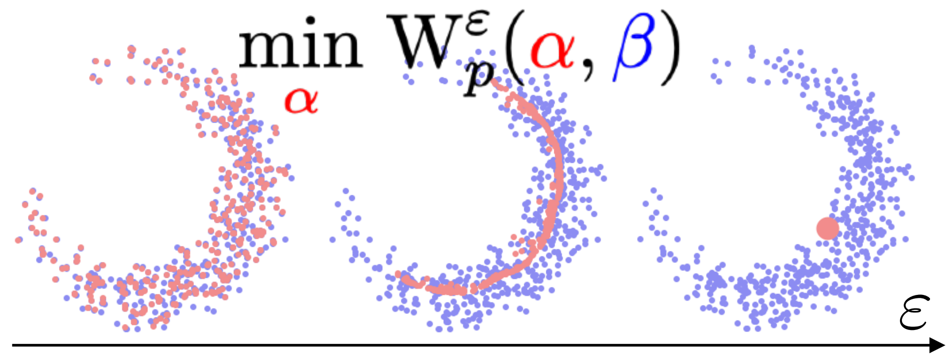
for $k(x, y) = -d(x, y)^p$ and

$\langle \alpha, \beta \rangle_k \overset{\text{def.}}{=} \int k(x, y) \mathrm{d}\alpha(x) \mathrm{d}\beta(y)$

Kernel norms and MMDs

$$\mathrm{W}_p^\varepsilon(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha, \pi_2=\beta} \int_{\mathcal{X}^2} d(x,y)^p \mathrm{d}\pi(x,y) + \varepsilon\, \mathrm{KL}(\pi|\alpha \otimes \beta)$$

Problem: $\mathrm{W}_p^\varepsilon(\alpha, \alpha) \neq 0$



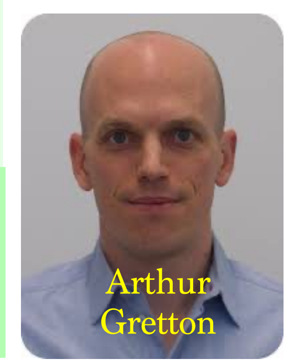
Prop.: $\pi^{(\varepsilon)} \xrightarrow{\varepsilon \to +\infty} \alpha \otimes \beta$

$\mathrm{W}_p^\varepsilon(\alpha, \beta)^p \xrightarrow{\varepsilon \to +\infty} -\langle \alpha, \beta \rangle_k$

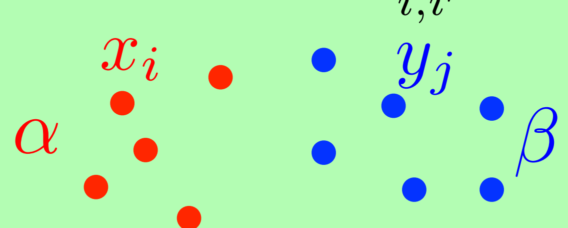
for $k(x,y) = -d(x,y)^p$ and

$$\langle \alpha, \beta \rangle_k \overset{\text{def.}}{=} \int k(x,y) \mathrm{d}\alpha(x) \mathrm{d}\beta(y)$$

Kernel norms (MMD): $\|\alpha - \beta\|_k^2 \overset{\text{def.}}{=} \langle \alpha - \beta, \alpha - \beta \rangle_k$



$$\|\alpha - \beta\|_k^2 = \frac{1}{n^2} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{m^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j)$$



k must be:
- conditionally positive
- universal

Sinkhorn Divergences

Sinkhorn Divergences:

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

[Ramdas, García Trillos, Cuturi, 2017]

Sinkhorn Divergences

Sinkhorn Divergences:

$$\overline{\mathrm{W}}_p^\varepsilon(\alpha,\beta)^p \stackrel{\text{def.}}{=} \mathrm{W}_p^\varepsilon(\alpha,\beta)^p - \tfrac{1}{2}\mathrm{W}_p^\varepsilon(\alpha,\alpha)^p - \tfrac{1}{2}\mathrm{W}_p^\varepsilon(\beta,\beta)^p$$

[Ramdas, García Trillos, Cuturi, 2017]

Theorem: $$\mathrm{W}_p(\alpha,\beta)^p \xleftarrow[\text{[Léonard 2012]}]{\varepsilon \to 0} \overline{\mathrm{W}}_p^\varepsilon(\alpha,\beta)^p \xrightarrow[\text{[Ramdas, García Trillos, Cuturi, 2017]}]{\varepsilon \to +\infty} \frac{1}{2}\|\alpha-\beta\|^2_{-d^p}$$

Sinkhorn Divergences

Sinkhorn Divergences:

$$\overline{\mathrm{W}}_p^{\varepsilon}(\alpha,\beta)^p \stackrel{\text{def.}}{=} \mathrm{W}_p^{\varepsilon}(\alpha,\beta)^p - \tfrac{1}{2}\mathrm{W}_p^{\varepsilon}(\alpha,\alpha)^p - \tfrac{1}{2}\mathrm{W}_p^{\varepsilon}(\beta,\beta)^p$$

[Ramdas, García Trillos, Cuturi, 2017]

Theorem: $\mathrm{W}_p(\alpha,\beta)^p \xleftarrow{\varepsilon \to 0} \overline{\mathrm{W}}_p^{\varepsilon}(\alpha,\beta)^p \xrightarrow{\varepsilon \to +\infty} \frac{1}{2}\|\alpha-\beta\|^2_{-d^p}$

[Léonard 2012]

[Ramdas, García Trillos, Cuturi, 2017]

Key problem: when is $k(x,y) = -d(x,y)^p$ a universal conditionaly positive kernel?

Proposition: $\|\cdot\|_{-\|\cdot\|^p}$ is a norm for $0 < p < 2$.

For $p=1$: $\dot{H}^{-\frac{d+1}{2}}(\mathbb{R}^d)$ Sobolev norm;

For $p=2$: $\|\xi\|^2_{-\|\cdot\|^2} = |\int x \mathrm{d}\xi(x)|^2$

Sinkhorn Divergences Positivity

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

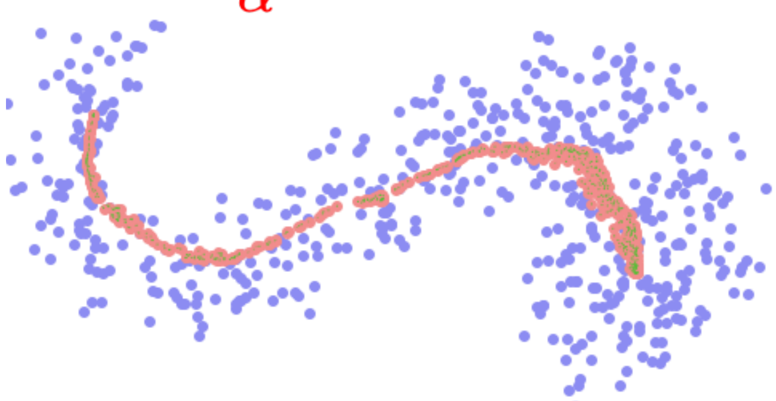
Theorem: [Feydy, Séjourné, P, Vialard, Trouvé, Amari 2018]

If $e^{-\frac{d^p}{\varepsilon}}$ is positive:

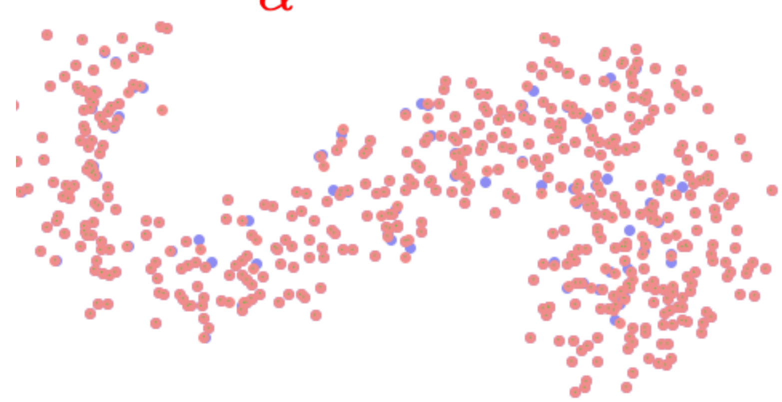
$\overline{\mathrm{W}}_p^\varepsilon \geqslant 0$ and $\overline{\mathrm{W}}_p^\varepsilon(\cdot,\beta)^p$ is convex.

$$\overline{\mathrm{W}}_p^\varepsilon(\alpha_n,\beta) \to 0 \iff \alpha_n \overset{\text{weak}*}{\rightharpoonup} \beta$$

$\min_{\alpha} \mathrm{W}_p^\varepsilon(\alpha,\beta)$



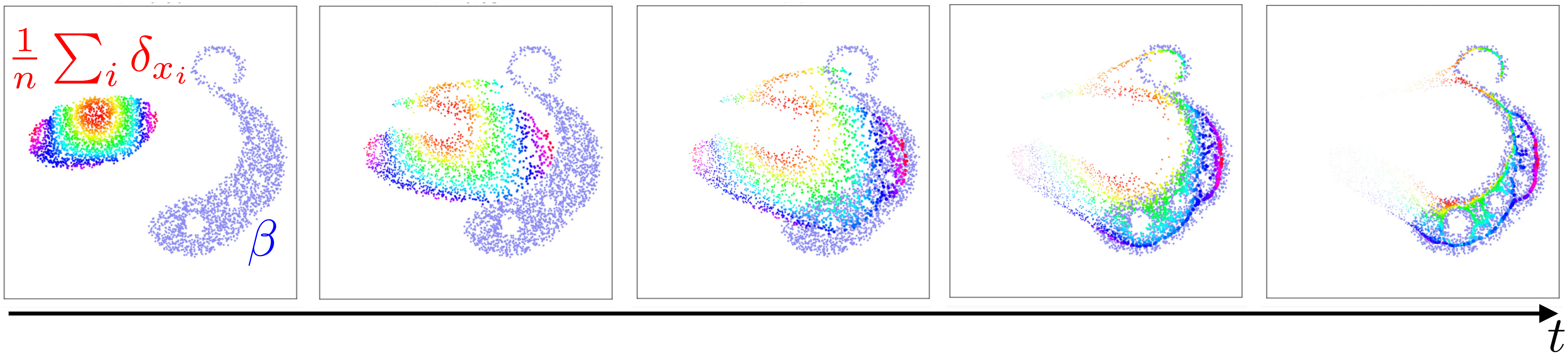
$\min_{\alpha} \overline{\mathrm{W}}_p^\varepsilon(\alpha,\beta)$



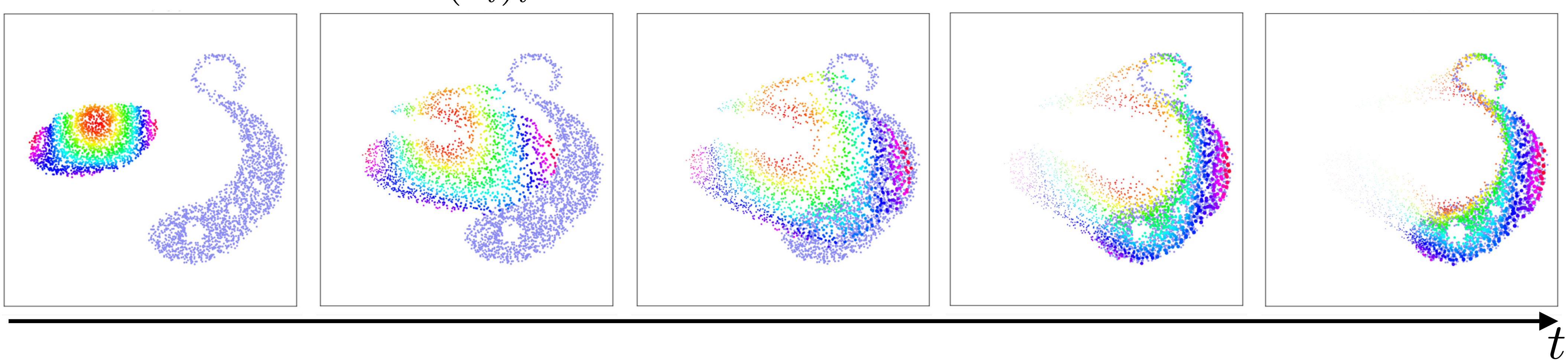
Wasserstein Gradient Flows

$$\min_{x=(x_i)_i} \mathcal{E}(x) \triangleq W_2^\varepsilon\left(\frac{1}{n}\sum_i \delta_{x_i}, \beta\right)$$

$$\text{Wasserstein flow:} \quad \frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\nabla\mathcal{E}(x(t))$$

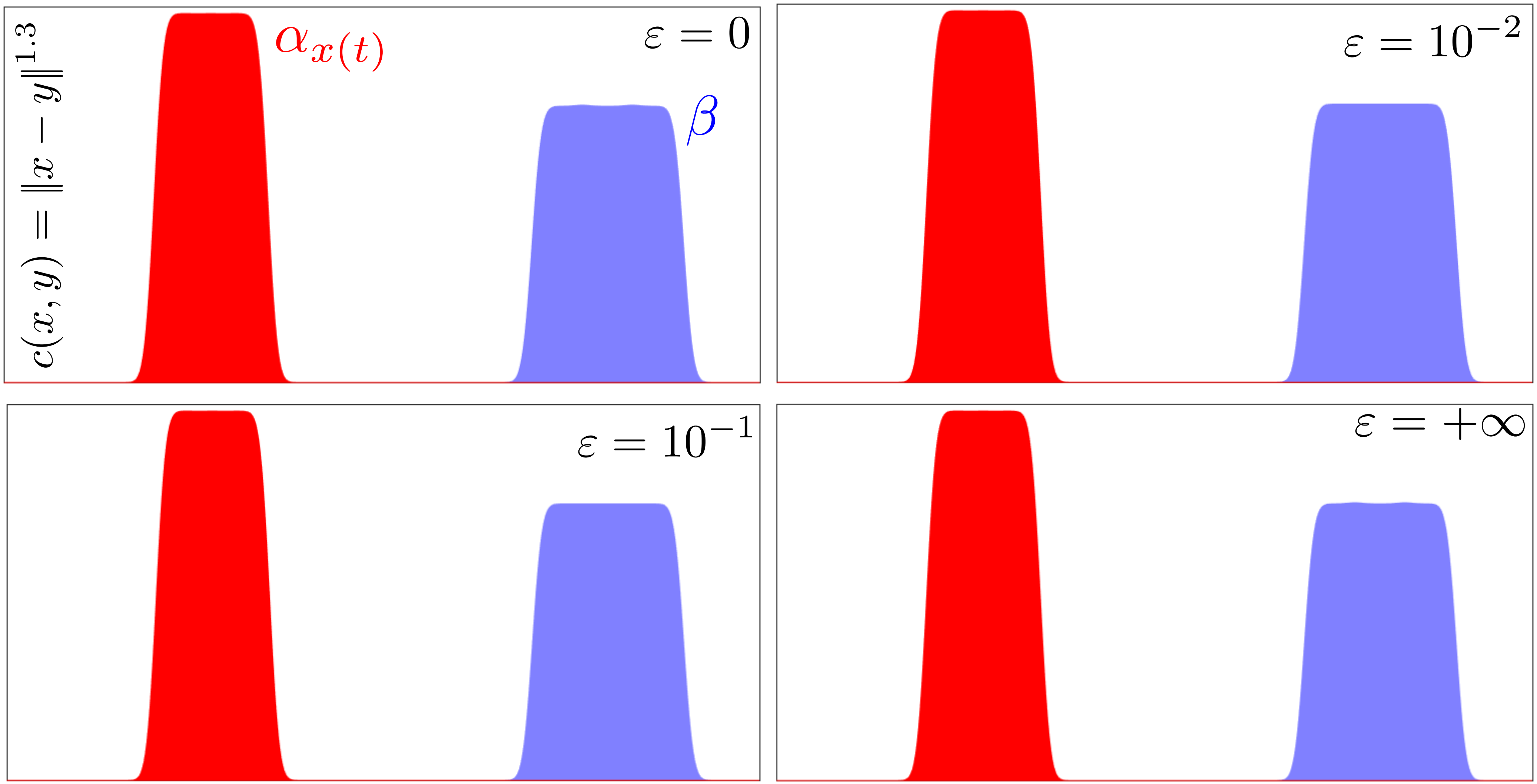


$$\min_{x=(x_i)_i} f(x) \triangleq \bar{W}_2^\varepsilon\left(\frac{1}{n}\sum_i \delta_{x_i}, \beta\right)$$



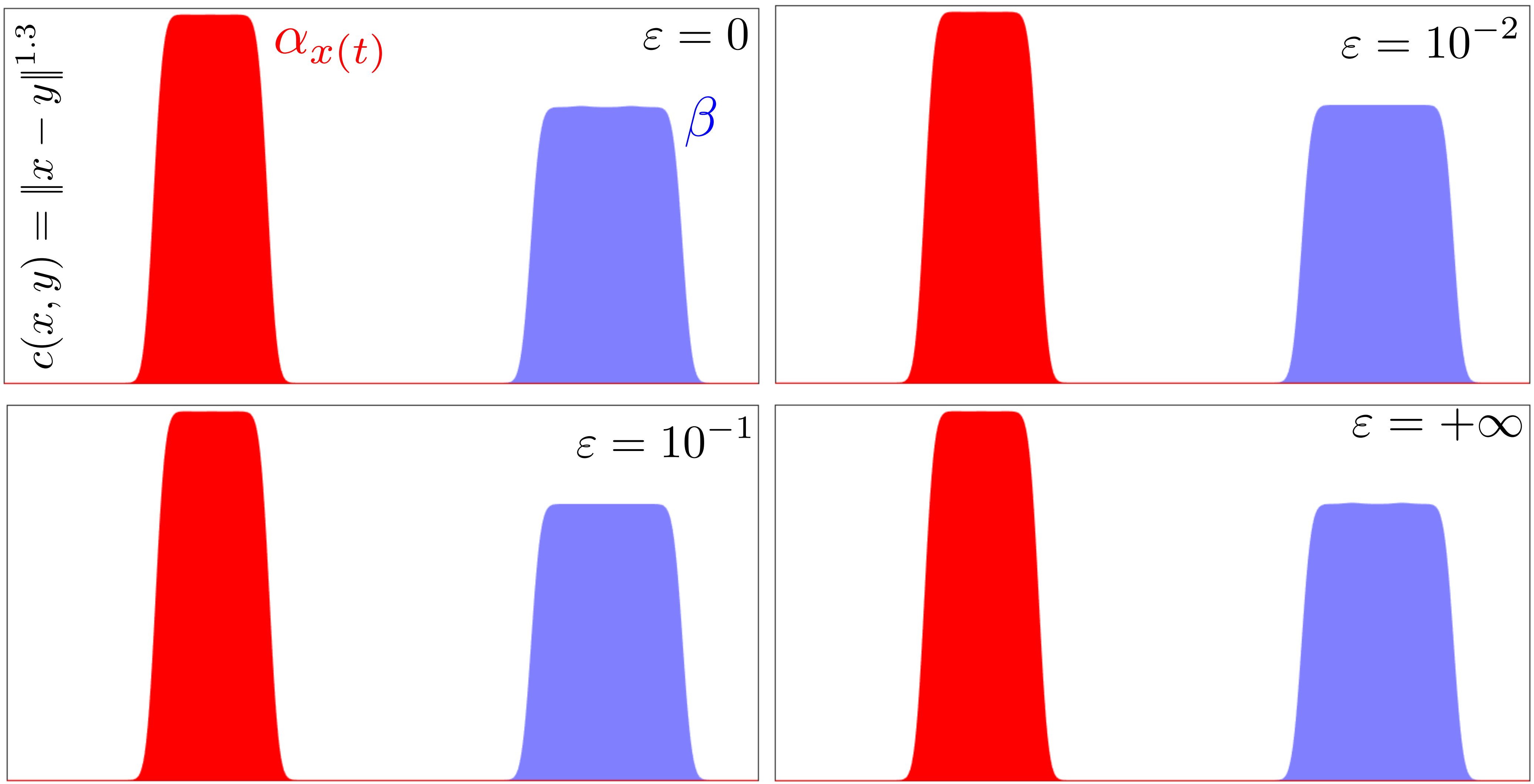
Wasserstein Gradient Flows

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\nabla \mathcal{E}(x(t)) \qquad \mathcal{E}(x) \overset{\text{def.}}{=} \overline{\mathrm{W}}^{1.3}_{\varepsilon,1.3}(\alpha_x, \beta) \qquad \alpha_x \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

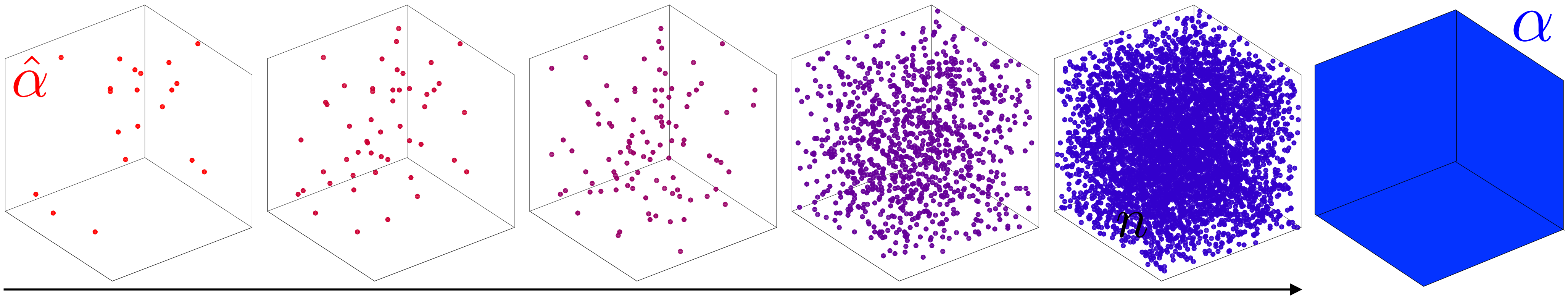


Wasserstein Gradient Flows

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\nabla\mathcal{E}(x(t)) \qquad \mathcal{E}(x) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}^{1.3}_{\varepsilon,1.3}(\alpha_x, \beta) \qquad \alpha_x \stackrel{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

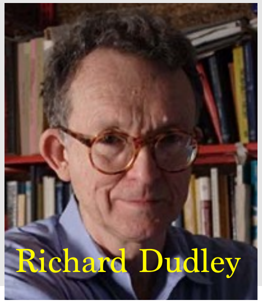


Sample Complexity

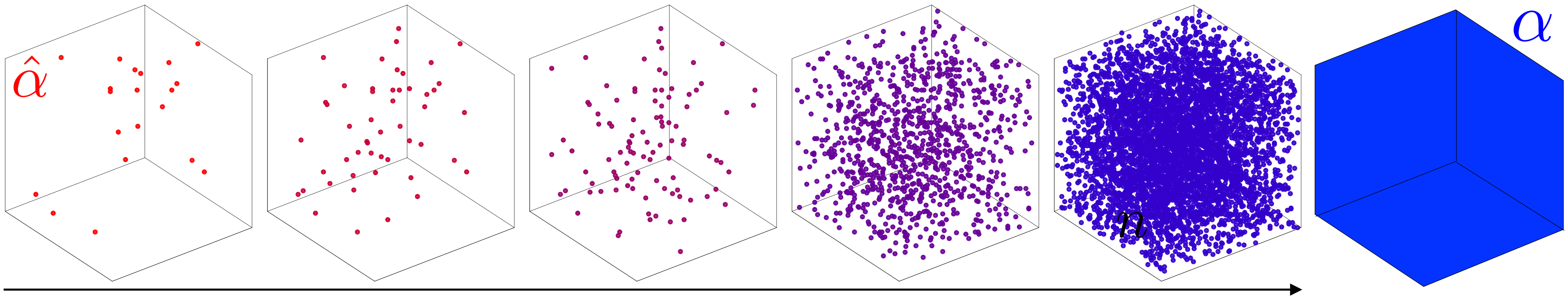


Theorem:

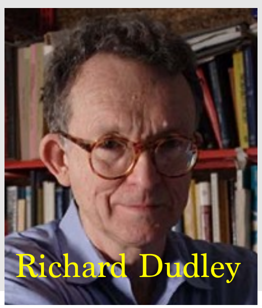
$$\mathbb{E}(|\mathrm{W}_p(\hat{\alpha}, \hat{\beta}) - \mathrm{W}_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$$



Sample Complexity



Theorem:



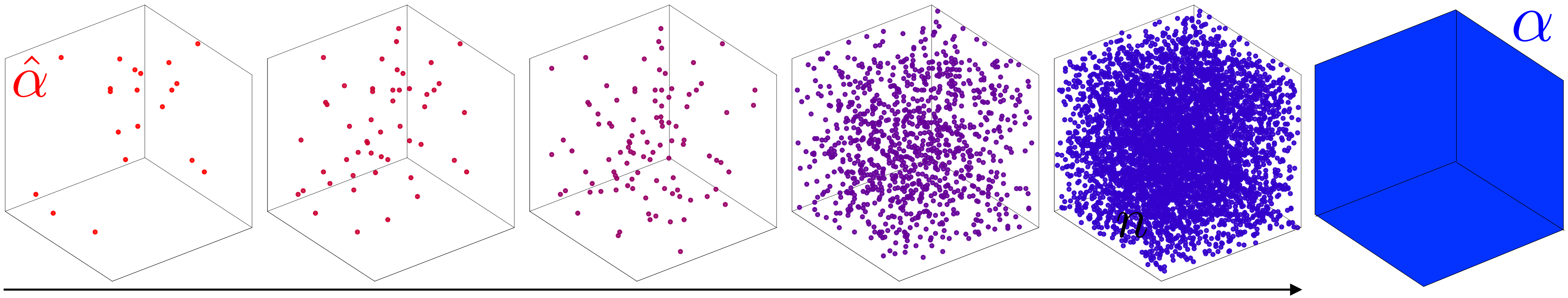
$$\mathbb{E}(|\mathrm{W}_p(\hat{\alpha}, \hat{\beta}) - \mathrm{W}_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$$

$$\mathbb{E}(|\mathrm{W}_2(\hat{\alpha}, \hat{\beta}) - \mathrm{W}_2(\alpha, \beta)|) = O(n^{-\frac{2}{d}})$$

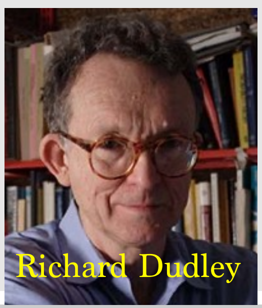
if $\alpha \neq \beta$

[Chizat, Roussillon, Léger, Vialard, P. 2020]

Sample Complexity



Theorem:



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

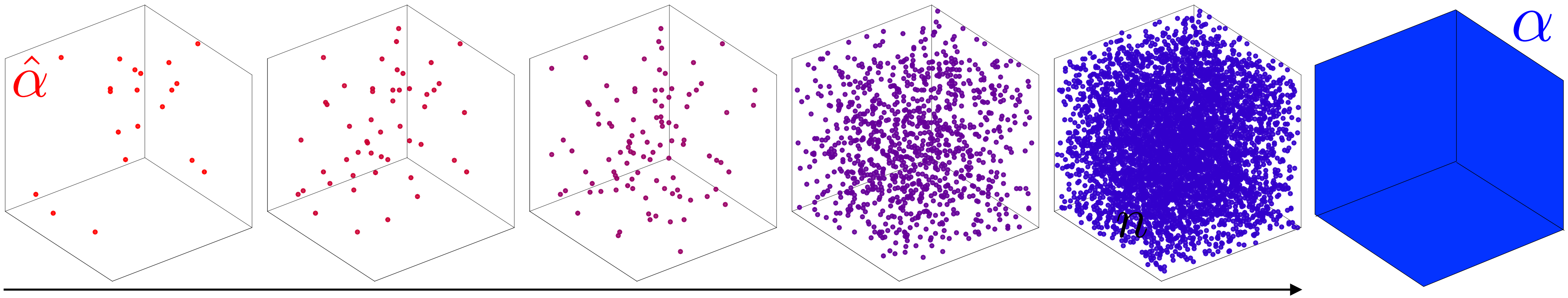
$$\mathbb{E}(|\mathrm{W}_2(\hat{\alpha},\hat{\beta}) - \mathrm{W}_2(\alpha,\beta)|) = O(n^{-\frac{2}{d}})$$

if $\alpha \neq \beta$

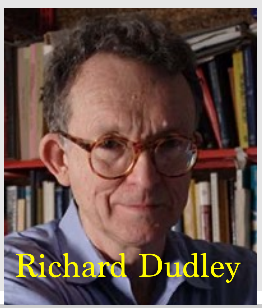
[Chizat, Roussillon, Léger, Vialard, P. 2020]

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

Sample Complexity



Theorem:



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

if $\alpha \neq \beta$

$$\mathbb{E}(|\mathrm{W}_2(\hat{\alpha}, \hat{\beta}) - \mathrm{W}_2(\alpha, \beta)|) = O(n^{-\frac{2}{d}})$$

[Chizat, Roussillon, Léger, Vialard, P. 2020]

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

$$\mathbb{E}(|\overline{\mathrm{W}}_p^{\varepsilon}(\hat{\alpha}, \hat{\beta}) - \overline{\mathrm{W}}_p^{\varepsilon}(\alpha, \beta)|) = O(\varepsilon^{-\frac{d}{2}} n^{-\frac{1}{2}})$$

[Genevay, Bach, P, Cuturi, 2019]

Sinkhorn Divergence Estimator

Theorem: [S. Pal, 2019]

$$\mathrm{W}_2^\varepsilon(\alpha, \beta)^2 - \mathrm{W}_2(\alpha, \beta)^2 = d\varepsilon \log \sqrt{2\pi\varepsilon} + \varepsilon(\mathrm{H}(\alpha) + \mathrm{H}(\beta)) + o(\varepsilon)$$

Sinkhorn Divergence Estimator

Theorem: [S. Pal, 2019]

$$\mathrm{W}_2^\varepsilon(\alpha,\beta)^2 - \mathrm{W}_2(\alpha,\beta)^2 = d\varepsilon \log\sqrt{2\pi\varepsilon} + \varepsilon(\mathrm{H}(\alpha)+\mathrm{H}(\beta)) + o(\varepsilon)$$

Displacement interpolation: $\rho_t : \alpha \to \beta$

Fisher information: $I(\alpha,\beta) \stackrel{\text{def.}}{=} \int_0^1 \int_{\mathbb{R}^d} \|\nabla \log(\rho_t(x))\|^2 \rho_t(x) \mathrm{d}x\mathrm{d}t$

Theorem: [G. Conforti and L. Tamanini 2019]

$$\overline{\mathrm{W}}_2^\varepsilon(\alpha,\beta)^2 - \mathrm{W}_2(\alpha,\beta)^2 = \varepsilon^2\left(I(\alpha,\beta) - \frac{I(\alpha,\alpha)}{2} - \frac{I(\beta,\beta)}{2}\right) + o(\varepsilon^2)$$

Sinkhorn Divergence Estimator

Theorem: [S. Pal, 2019]

$$\mathrm{W}_2^\varepsilon(\alpha,\beta)^2 - \mathrm{W}_2(\alpha,\beta)^2 = d\varepsilon \log\sqrt{2\pi\varepsilon} + \varepsilon(\mathrm{H}(\alpha)+\mathrm{H}(\beta)) + o(\varepsilon)$$

Displacement interpolation: $\rho_t : \alpha \to \beta$

Fisher information: $I(\alpha,\beta) \overset{\text{def.}}{=} \int_0^1 \int_{\mathbb{R}^d} \|\nabla \log(\rho_t(x))\|^2 \rho_t(x) \mathrm{d}x\mathrm{d}t$

Theorem: [G. Conforti and L. Tamanini 2019]

$$\overline{\mathrm{W}}_2^\varepsilon(\alpha,\beta)^2 - \mathrm{W}_2(\alpha,\beta)^2 = \varepsilon^2\left(I(\alpha,\beta) - \frac{I(\alpha,\alpha)}{2} - \frac{I(\beta,\beta)}{2}\right) + o(\varepsilon^2)$$

$$\mathbb{E}(|\overline{\mathrm{W}}_2^\varepsilon(\hat\alpha,\hat\beta)^2 - \overline{\mathrm{W}}_2^\varepsilon(\alpha,\beta)^2|) = O(\varepsilon^{-\frac{d}{2}} n^{-\frac{1}{2}})$$

$$\varepsilon = n^{-\frac{1}{d}}$$

$$\alpha \neq \beta$$

$$\mathbb{E}(|\overline{\mathrm{W}}_2^\varepsilon(\hat\alpha,\hat\beta) - \mathrm{W}_2(\alpha,\beta)|) = O(n^{-\frac{2}{d}})$$

Leveraging smoothness in high dimension?

$\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}, \frac{\mathrm{d}\beta}{\mathrm{d}y}\right)$ with d derivatives

$\longrightarrow \mathbb{E}|\hat{W} - W| \leqslant C(d) n^{-1/2}$

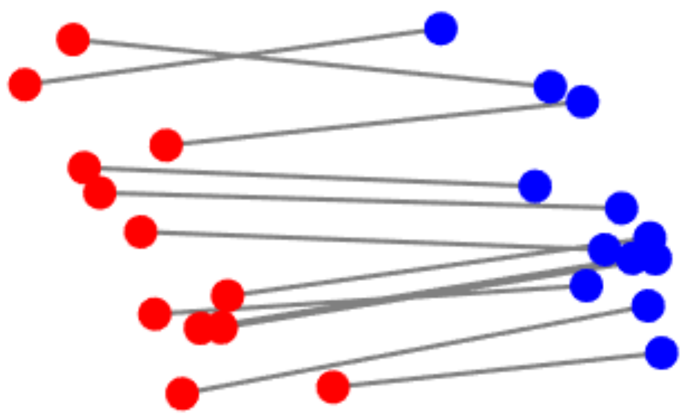
$\longrightarrow$ Complexity $O(n^2)$

Leveraging smoothness in high dimension?

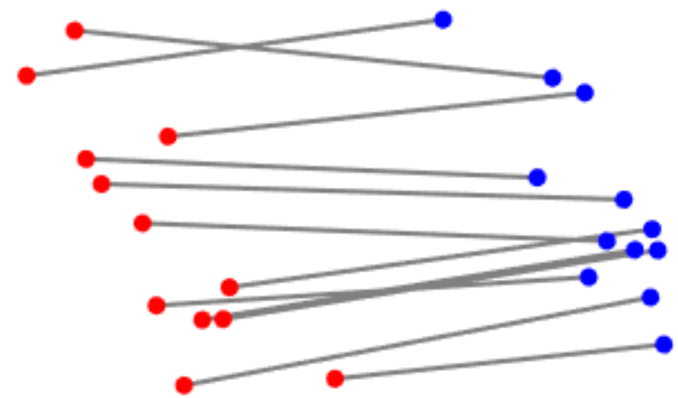
$\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}, \frac{\mathrm{d}\beta}{\mathrm{d}y}\right)$ with d derivatives

- $\mathbb{E}|\hat{W} - W| \leqslant C(d) n^{-1/2}$
- Complexity $O(n^2)$

$\overline{\mathrm{W}}_2^{\varepsilon}(\hat{\alpha}, \hat{\beta})$ is a bad proxy for $\mathrm{W}_2(G_\varepsilon \star \hat{\alpha}, G_\varepsilon \star \hat{\beta})$ ...



$\overline{\mathrm{W}}_2^{\varepsilon}(\hat{\alpha}, \hat{\beta})$

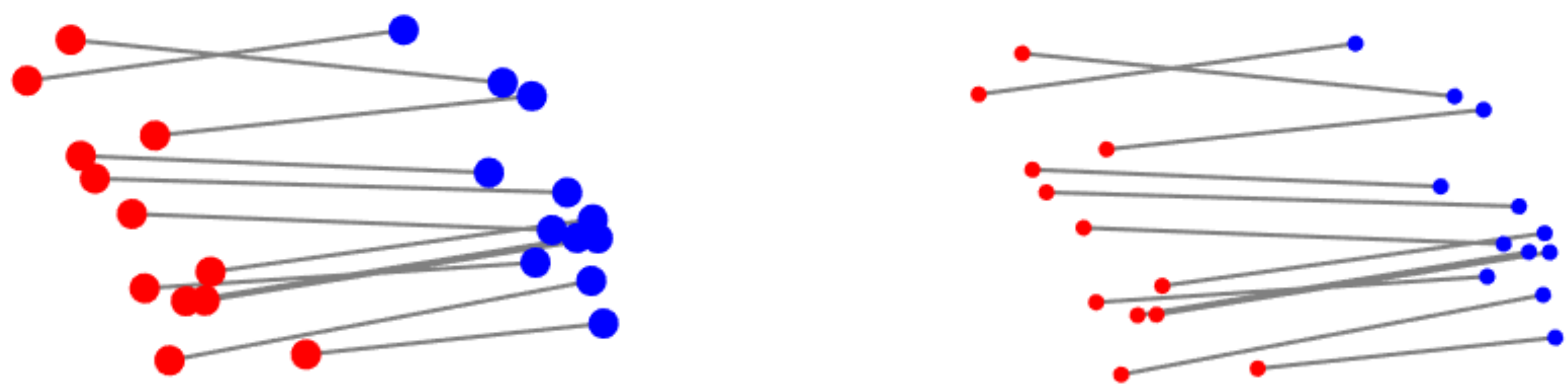


$\mathrm{W}_2(G_\varepsilon \star \hat{\alpha}, G_\varepsilon \star \hat{\beta})$

Leveraging smoothness in high dimension?

$(\frac{d\alpha}{dx}, \frac{d\beta}{dy})$ with d derivatives $\to$ $\mathbb{E}|\hat{W} - W| \leqslant C(d) n^{-1/2}$; Complexity $O(n^2)$

$\overline{\mathrm{W}}_2^{\varepsilon}(\hat{\alpha}, \hat{\beta})$ is a bad proxy for $\mathrm{W}_2(G_\varepsilon \star \hat{\alpha}, G_\varepsilon \star \hat{\beta})$ ...



$\overline{\mathrm{W}}_2^{\varepsilon}(\hat{\alpha}, \hat{\beta})$ $\qquad$ $\mathrm{W}_2(G_\varepsilon \star \hat{\alpha}, G_\varepsilon \star \hat{\beta})$

Recent breakthrough: [Vacher, Muzellec, Rudi, Bach, Vialard, 2021]
$\to$ SDP programming.

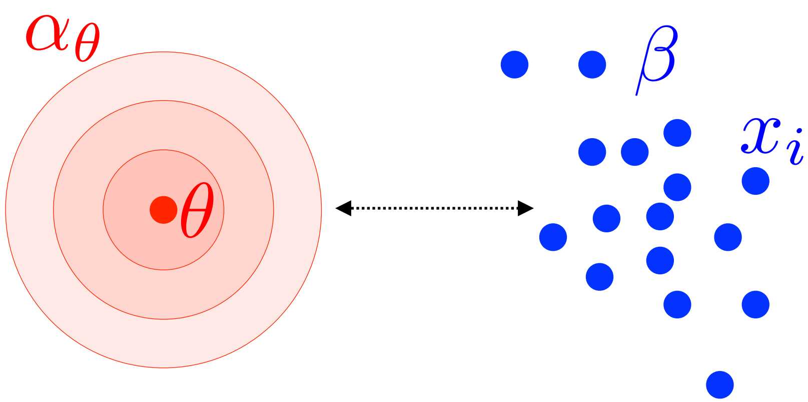
Overview

- Entropic Regularization and Sinkhorn
- Convergence Analysis
- Sinkhorn Divergences
- **Generative Model Fitting**

Density Fitting and Generative Models

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

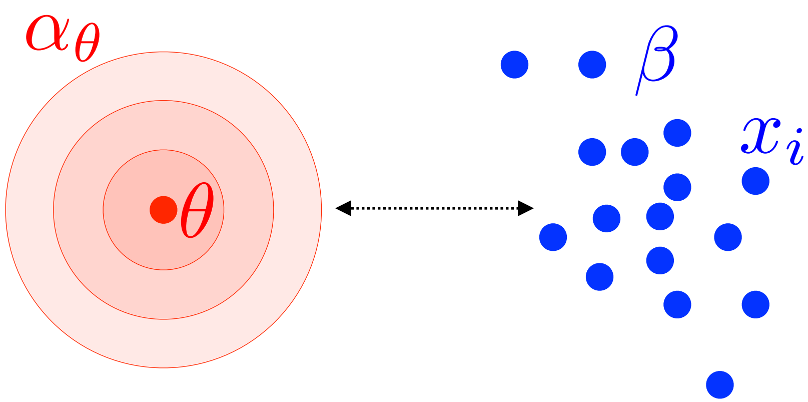
Parametric model: $\theta \mapsto \alpha_\theta$



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

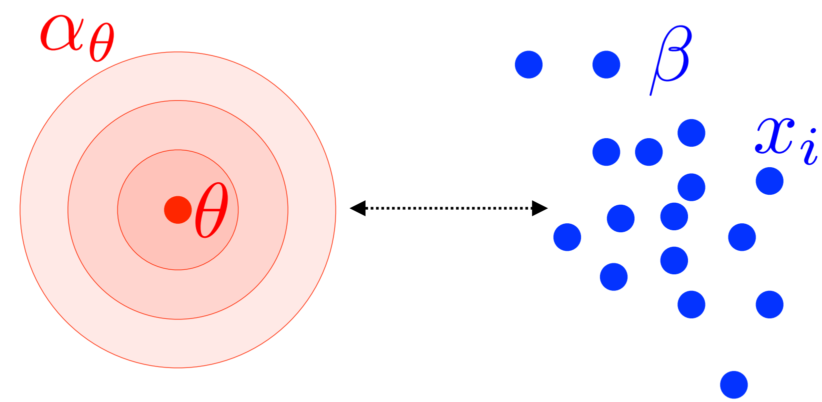
$$\min_\theta \; -\sum_i \log(\rho_\theta(x_i)) \overset{n \to +\infty}{\longrightarrow} \mathrm{KL}(\beta|\alpha_\theta)$$

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

$$\min_\theta \; -\sum_i \log(\rho_\theta(x_i)) \overset{n\to+\infty}{\longrightarrow} \mathrm{KL}(\beta|\alpha_\theta)$$

Maximum likelihood (MLE)

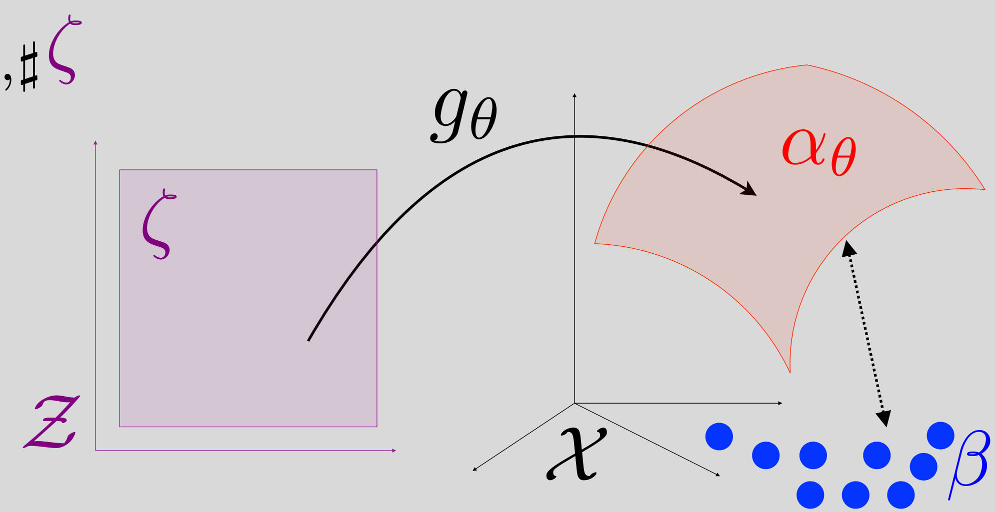
Generative model fit: $\alpha_\theta = g_{\theta,\sharp}\zeta$

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

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

$$\min_\theta \overline{\mathrm{W}}_{\varepsilon,p}^{p}(\alpha_\theta, \beta)$$

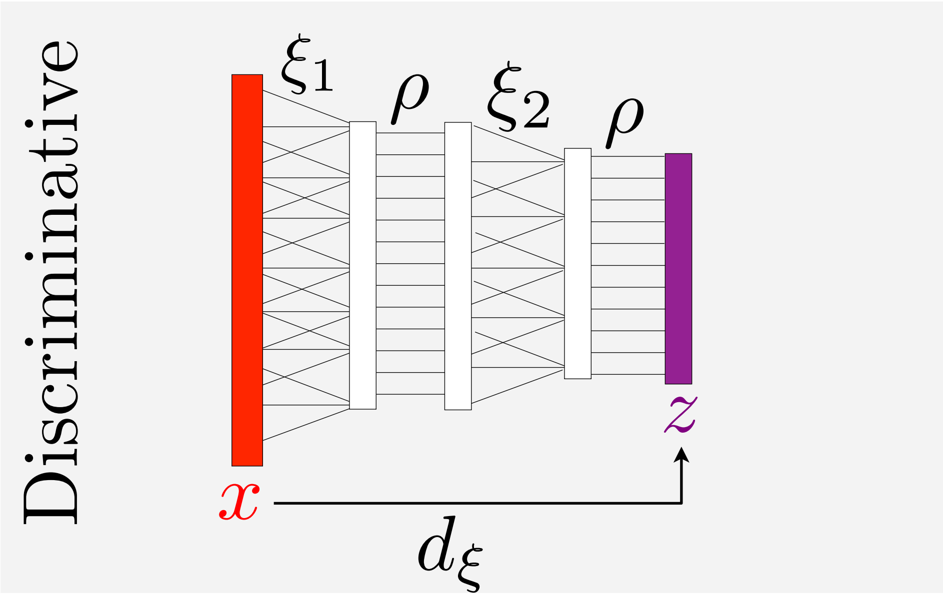


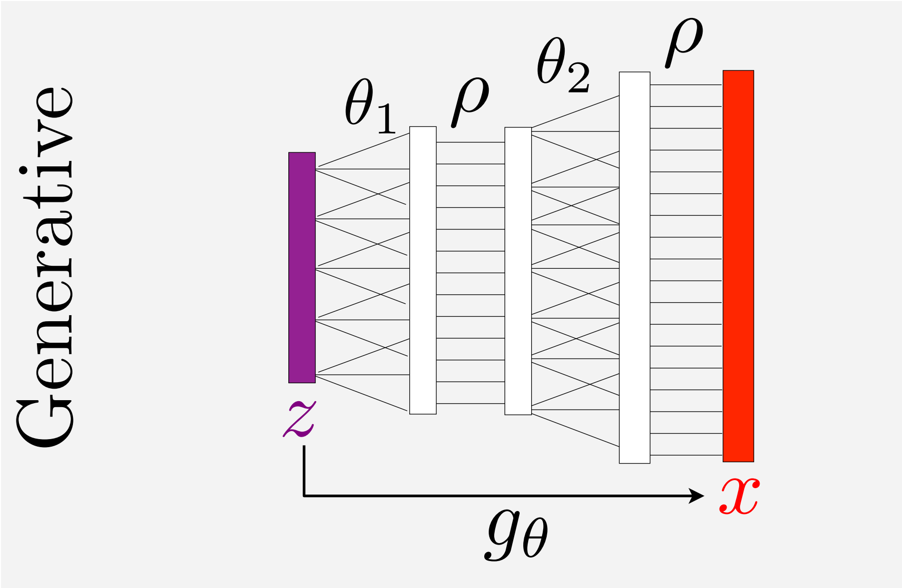
Deep Discriminative vs Generative Models

Deep networks:

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

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



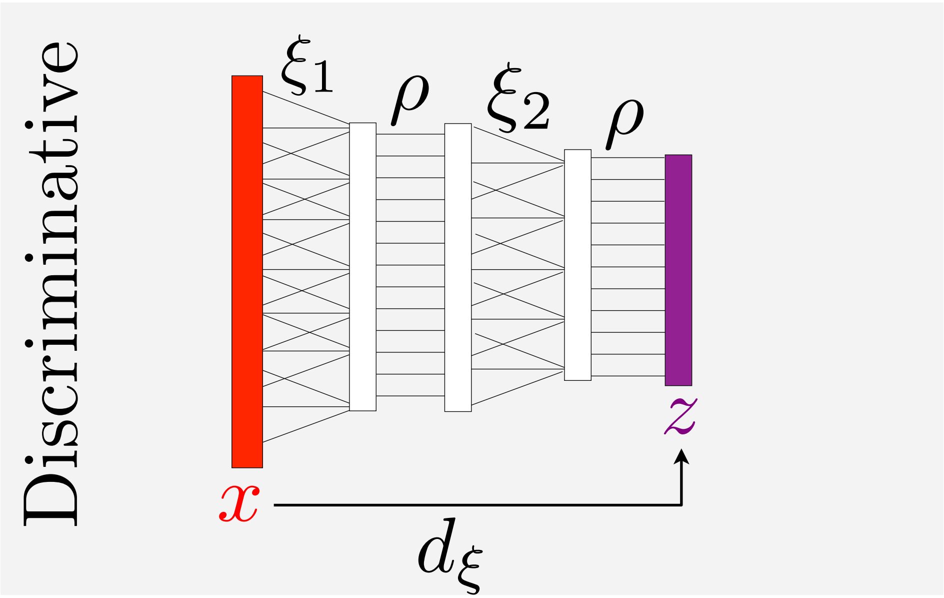


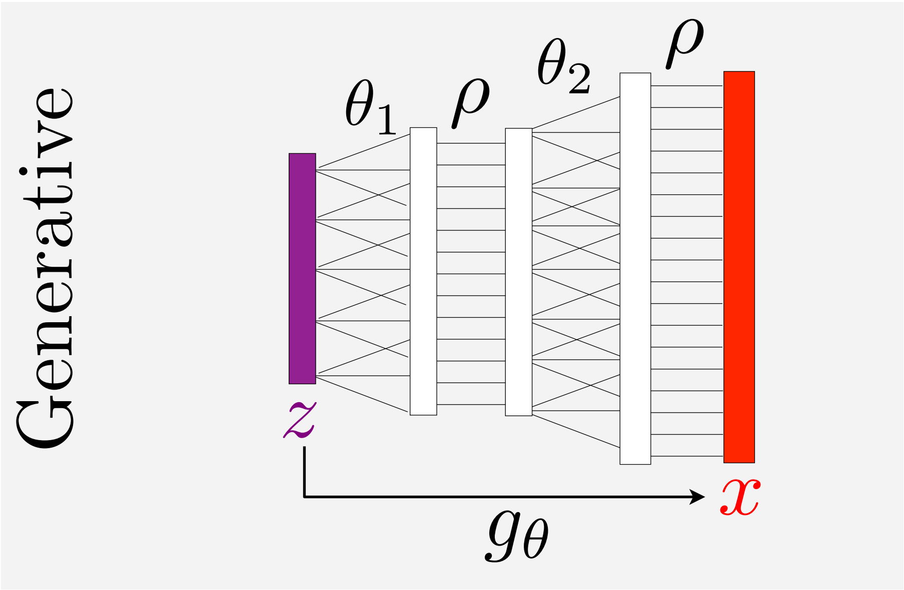
Deep Discriminative vs Generative Models

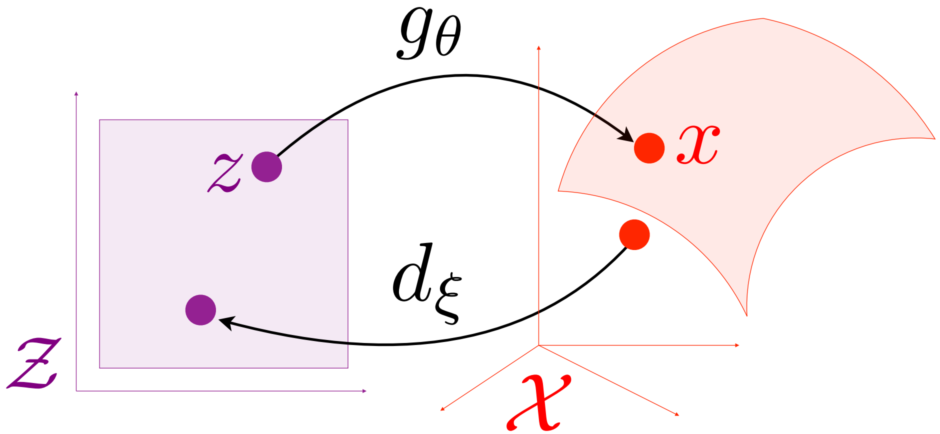
Deep networks:

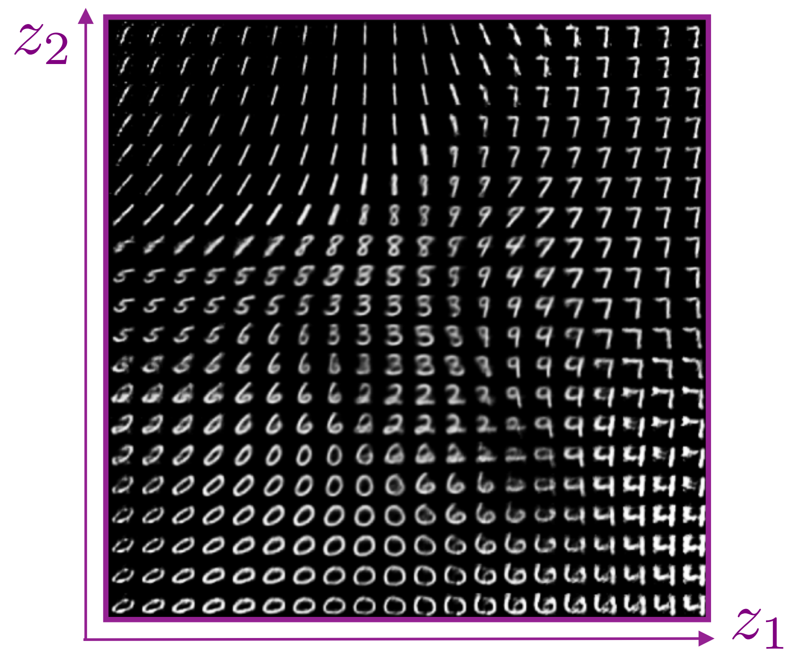
$$d_\xi(x) = \rho(\xi_K(\ldots\rho(\xi_2(\rho(\xi_1(x)\ldots)$$

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

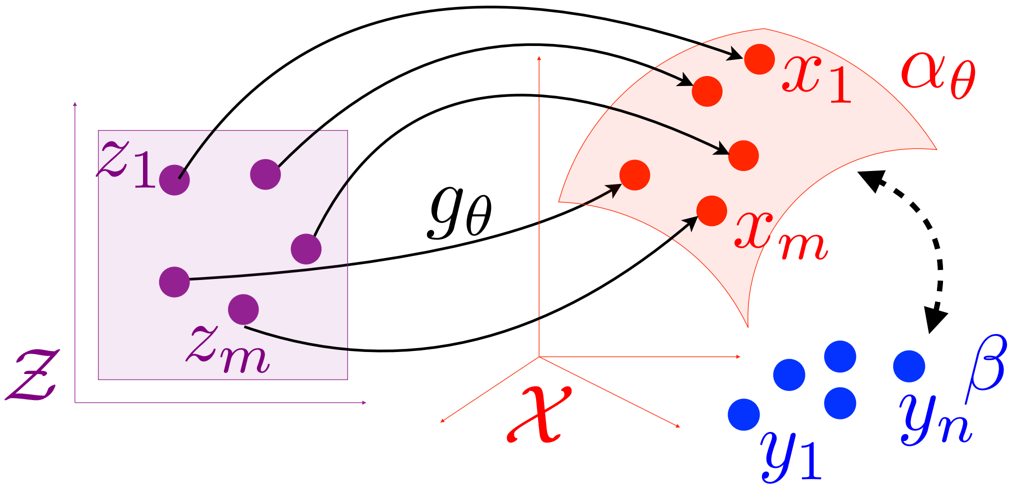








Training Architecture



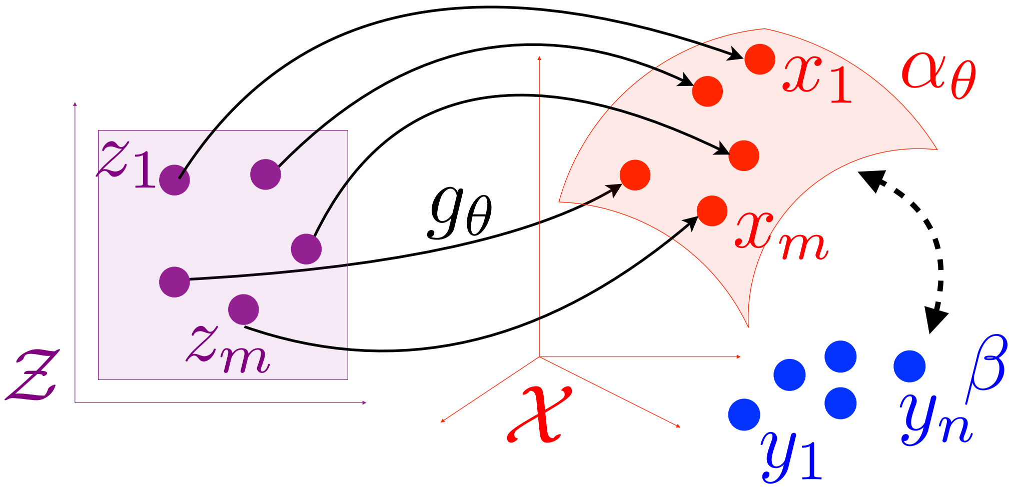
$$\min_\theta \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

$$\theta \leftarrow \theta - \tau \nabla \hat{\mathcal{E}}(\theta)$$

$$\hat{\mathcal{E}}(\theta) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}\left(\tfrac{1}{m}\sum_i \delta_{g_\theta(z_i)}, \beta\right)$$

Training Architecture

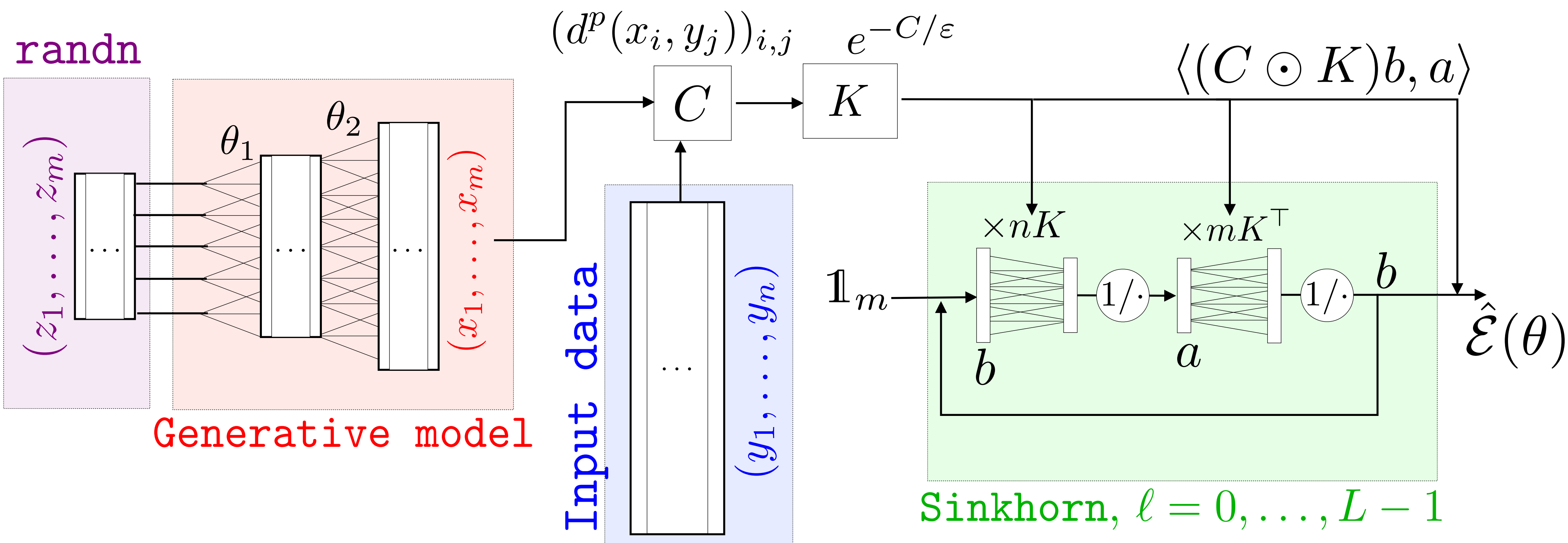


$$\min_\theta \mathcal{E}(\theta) \overset{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

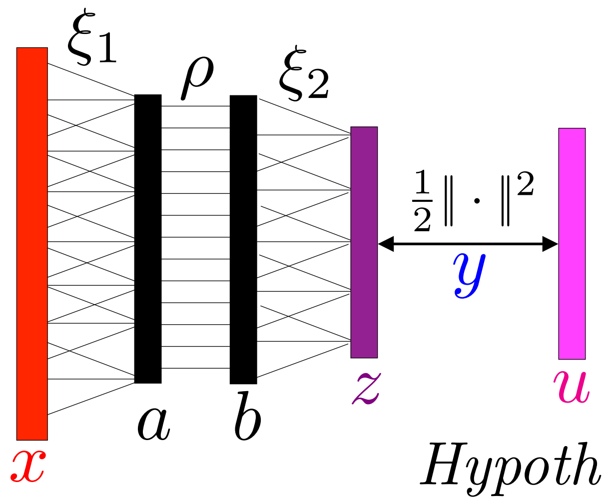
$$\theta \leftarrow \theta - \tau \nabla \hat{\mathcal{E}}(\theta)$$

$$\hat{\mathcal{E}}(\theta) \overset{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}\left(\tfrac{1}{m}\textstyle\sum_i \delta_{g_\theta(z_i)}, \beta\right)$$



Automatic Differentiation

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



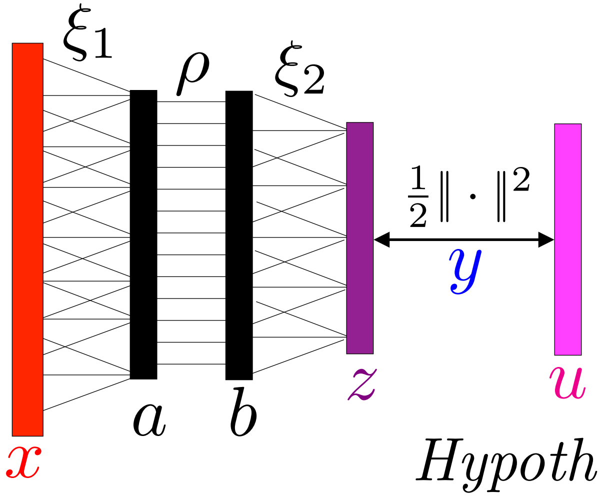

```
function y = E(x)
  a = xi1*x
  b = rho(a)
  z = xi2*b
  y = 1/2*norm(z-u)^2
```

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \dots)$ 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: $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ computable in K operations.




```
function y = E(x)
  a = xi1*x
  b = rho(a)
  z = xi2*b
  y = 1/2*norm(z-u)^2
```

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

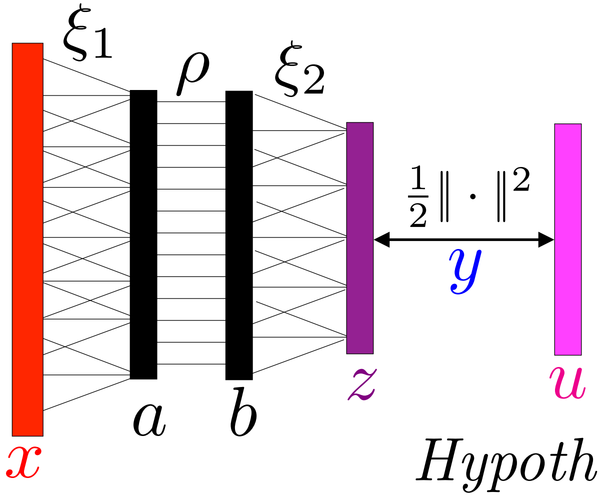
Question: What is the complexity of computing $\nabla\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^n$?

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

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

Automatic Differentiation

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




```
function y = E(x)
  a = xi1*x
  b = rho(a)
  z = xi2*b
  y = 1/2*norm(z-u)^2
```

```
function dx = nablaE(x)
  dz = z-u
  db = xi2'*dz
  da = diag(dphi(a)) * db
  dx = xi1'*da
```

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

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

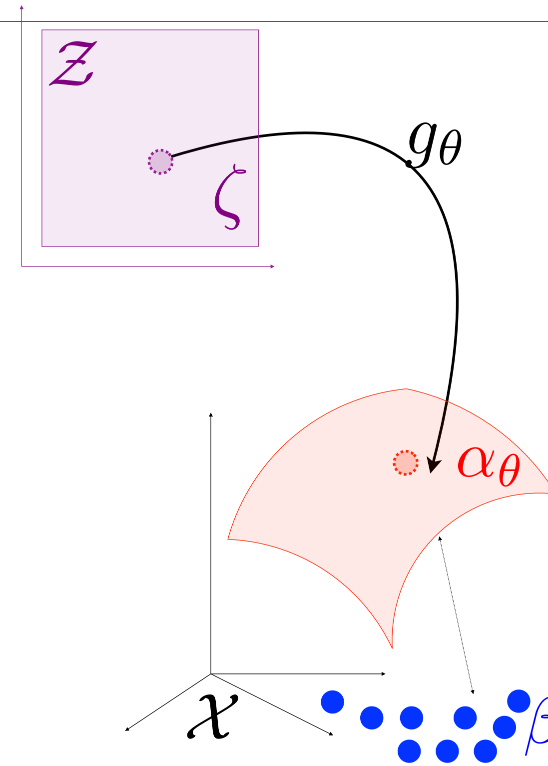
Theorem: there is an algorithm to compute $\nabla \mathcal{E}$ in $O(K)$ operations. [Seppo Linnainmaa, 1970]

This algorithm is reverse mode automatic differentiation

Examples of Images Generation

Inputs β

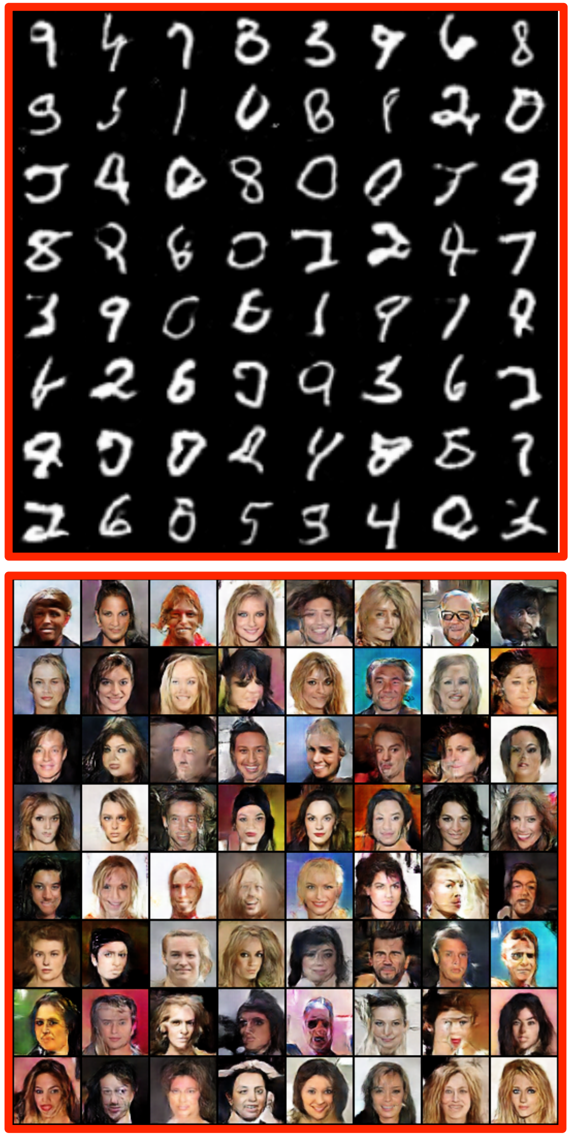
Generated α_θ

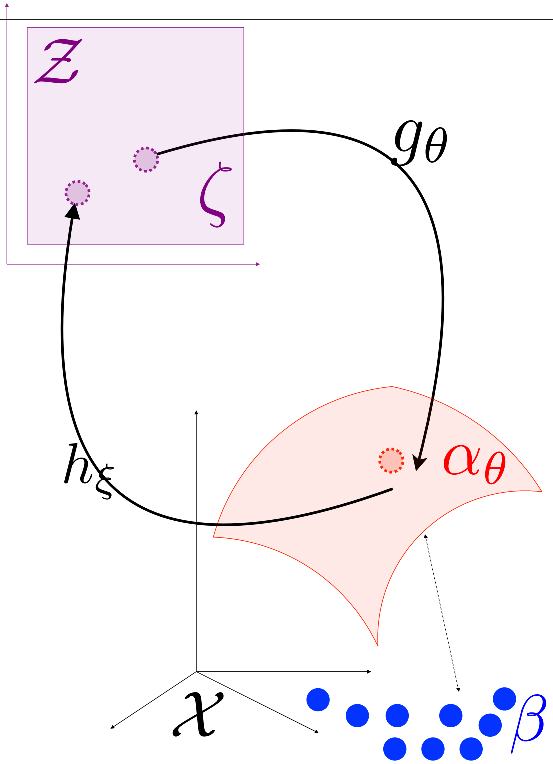


Examples of Images Generation

Inputs β

Generated α_θ



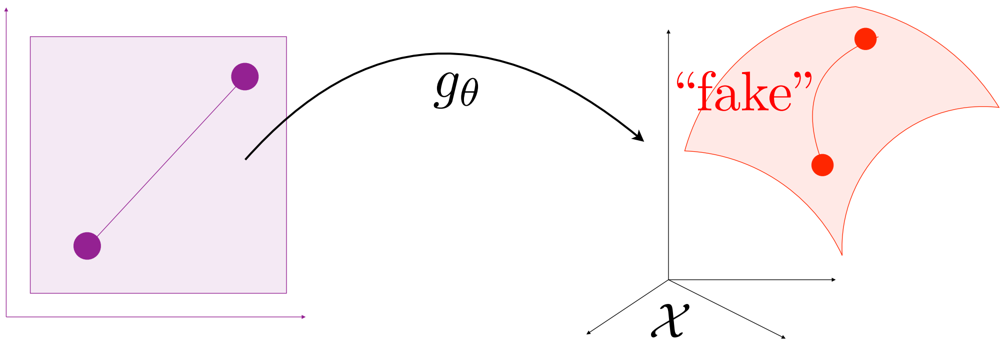


→ Need to learn the metric $d(x,y) = \|h_\xi(x) - h_\xi(y)\|$ (GANs)

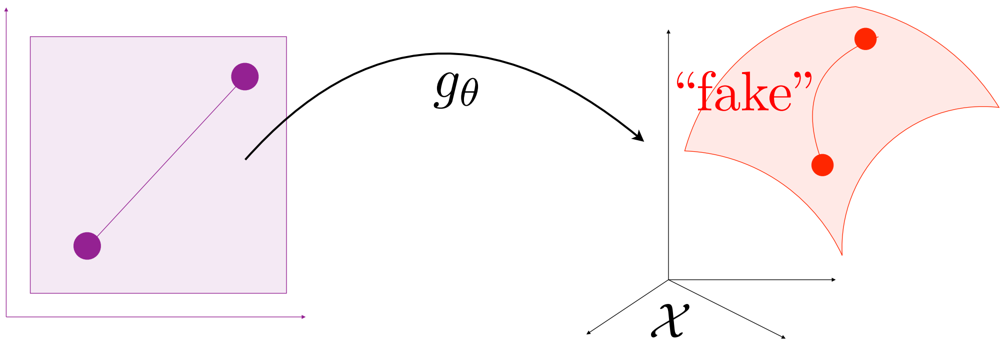
→ Influence of ε?

→ Performance evaluation of generative models is an open problem.

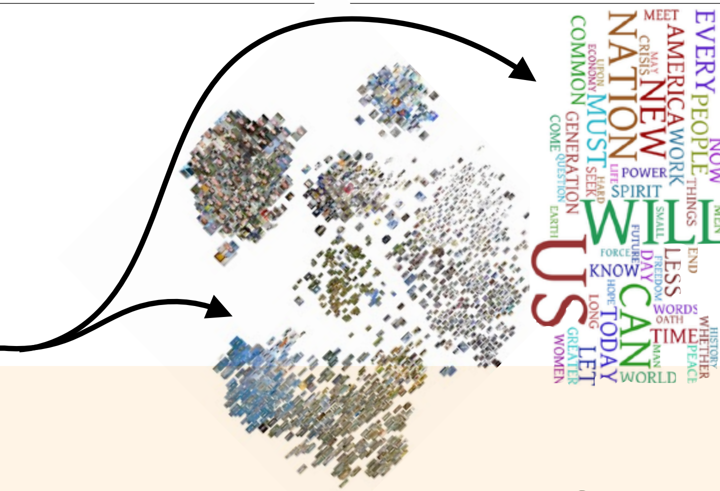
Progressive Growing of GANs for Improved Quality, Stability, and Variation
Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, ICLR 2018



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



Open Problems

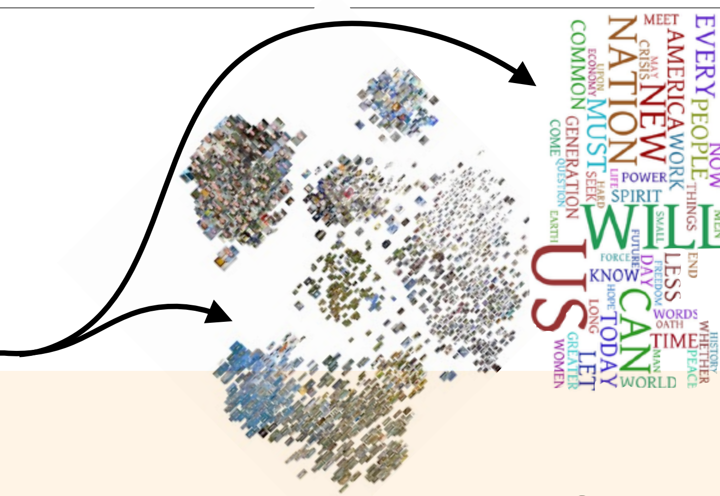


Toward high-dimensional OT:

$\rightarrow$ Scalable geometrical loss functions in high dimension?

$\rightarrow$ Performance quality measures for unsupervised learning?

Open Problems



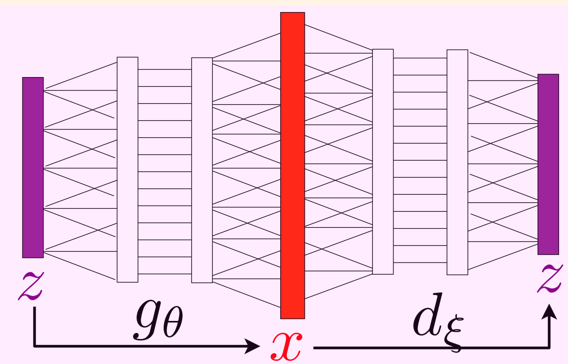
Toward high-dimensional OT:

→ Scalable geometrical loss functions in high dimension?

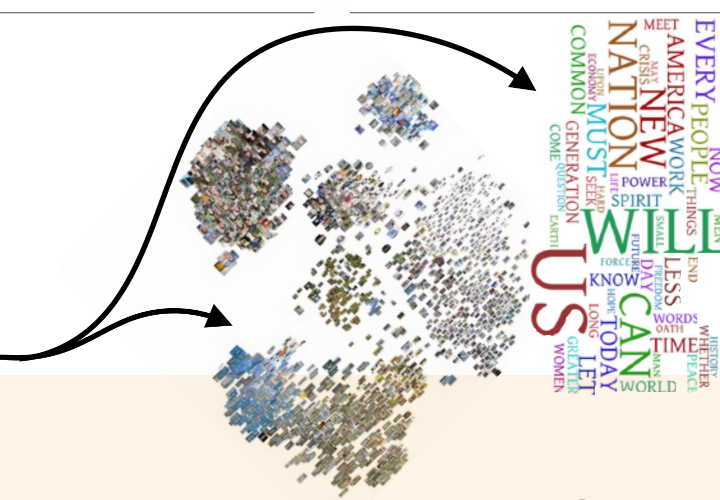
→ Performance quality measures for unsupervised learning?

Metric learning for OT:

→ Adversarial training to leverage multi scale priors?



Open Problems



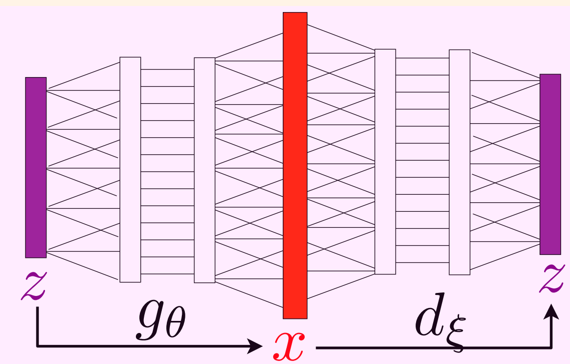
Toward high-dimensional OT:

→ Scalable geometrical loss functions in high dimension?

→ Performance quality measures for unsupervised learning?

Metric learning for OT:

→ Adversarial training to leverage multi scale priors?



Beyond comparing measures:

→ Learning for surfaces, graphs, metric spaces?

→ Using Gromov-Wasserstein geometry?