Numerical Optimal Transport

http://optimaltransport.github.io

Entropic Regularization

Gabriel Peyré

www.numerical-tours.com

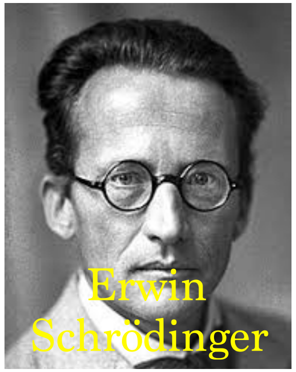
Overview

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

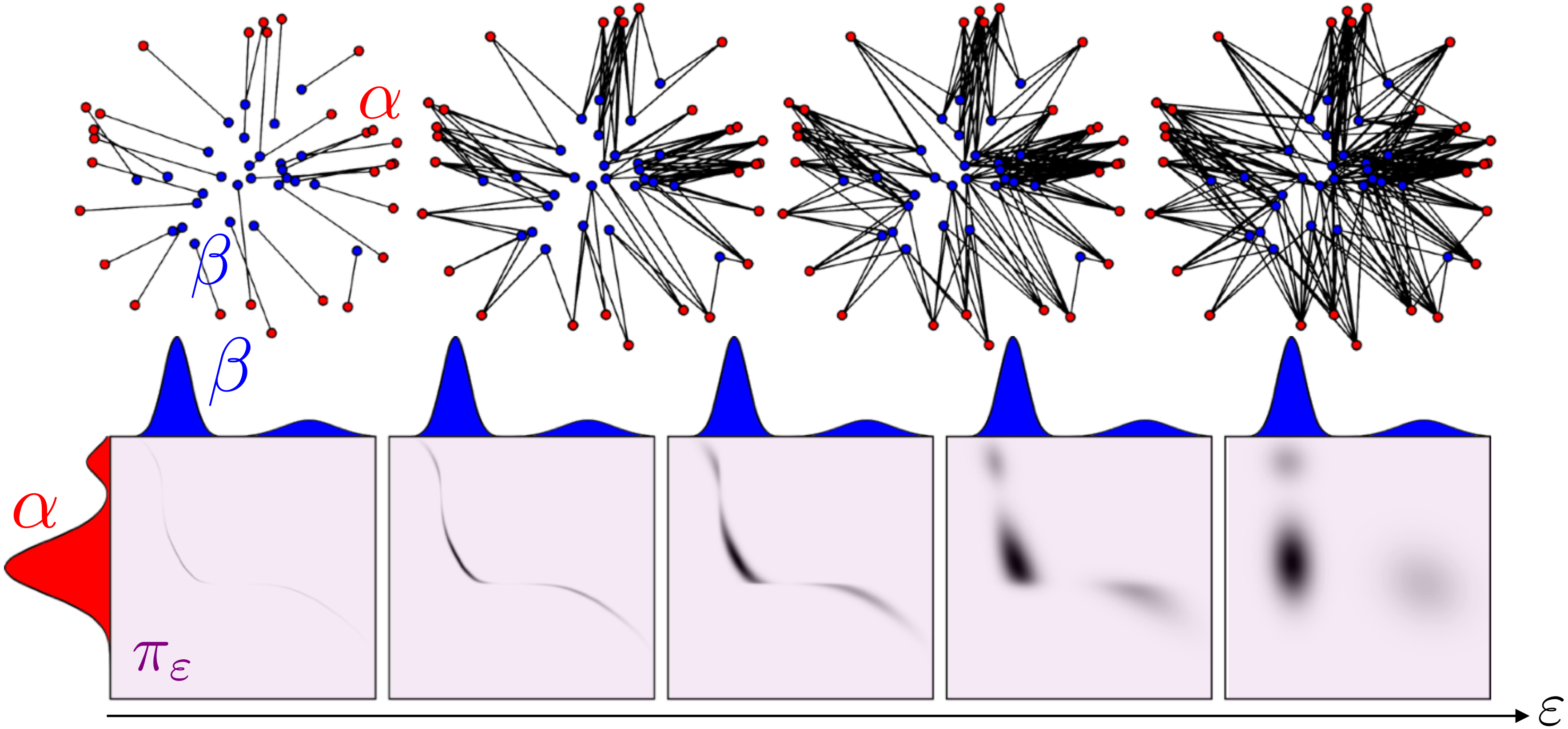
Entropic Regularization

Schrödinger's problem: [1931]

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

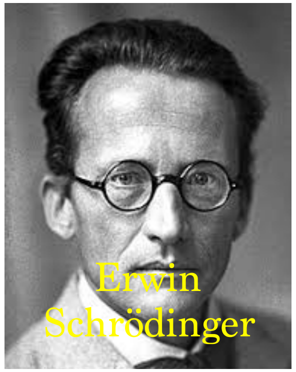


$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$



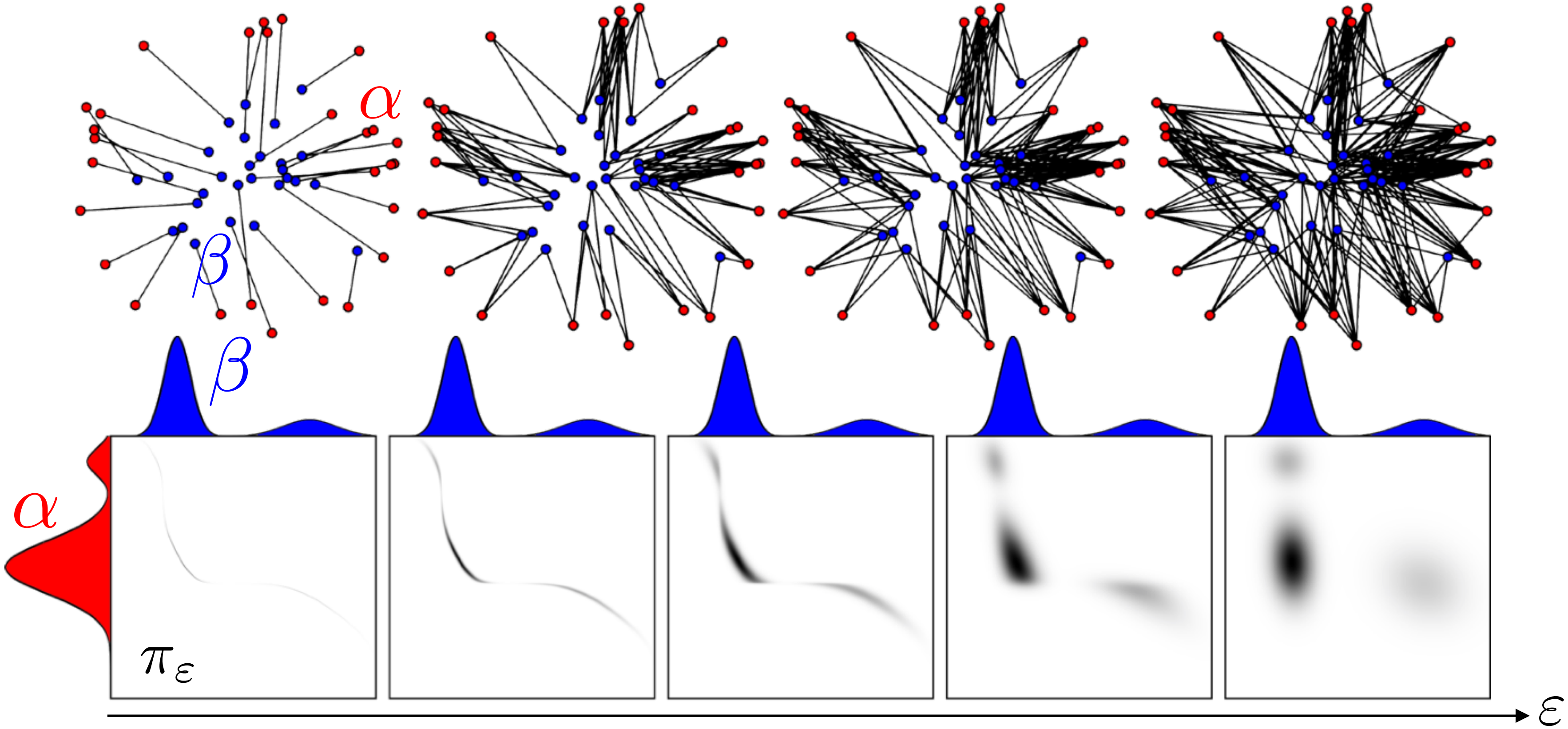
Entropic Regularization

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]

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



Probabilistic Interpretation

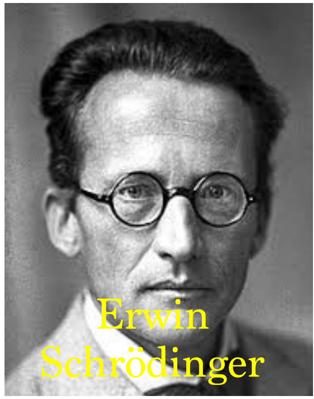
Relative-entropy: $\mathrm{KL}(\pi|\alpha \otimes \beta) \overset{\text{def.}}{=} \int_{\mathcal{X}^2} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y)\right)\mathrm{d}\pi(x,y)$

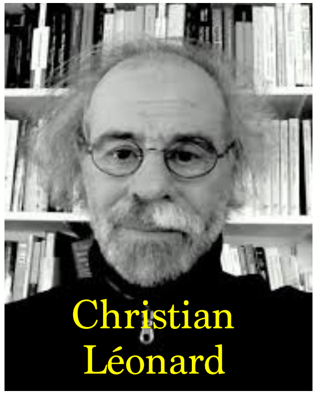
Schrödinger's problem: [1931]

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

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

Mutual information



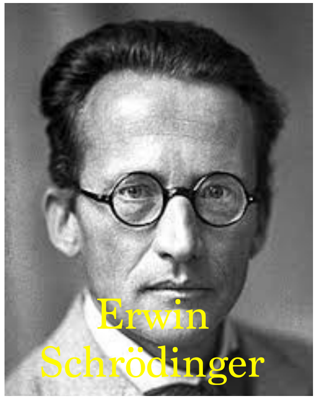


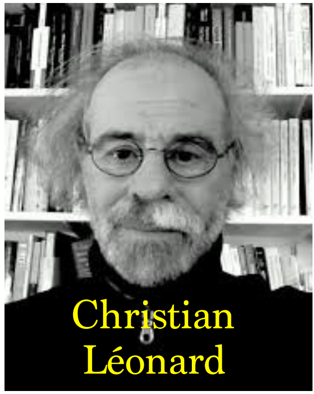
Probabilistic Interpretation

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]

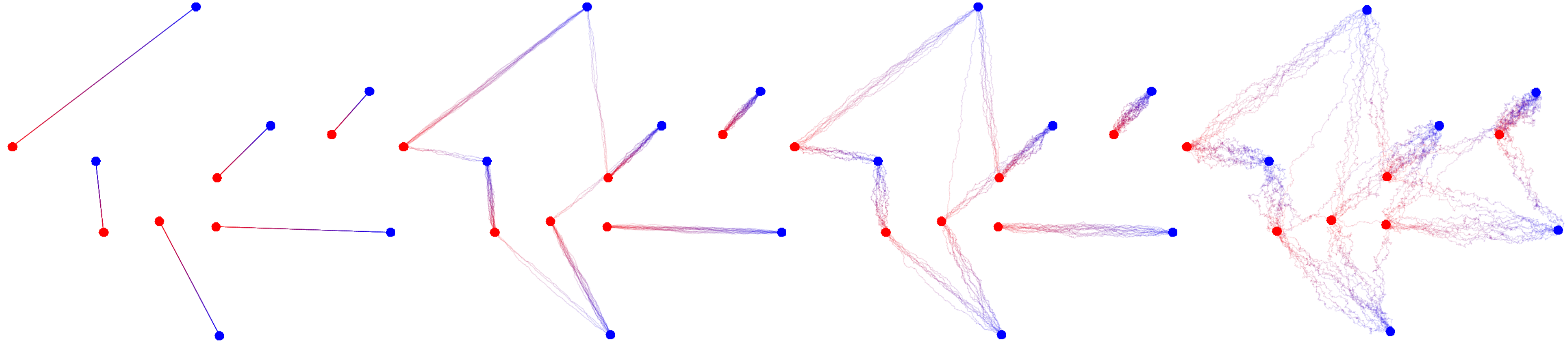
$$W^p_{\varepsilon,p}(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon\mathrm{KL}(\pi|\alpha\otimes\beta)$$





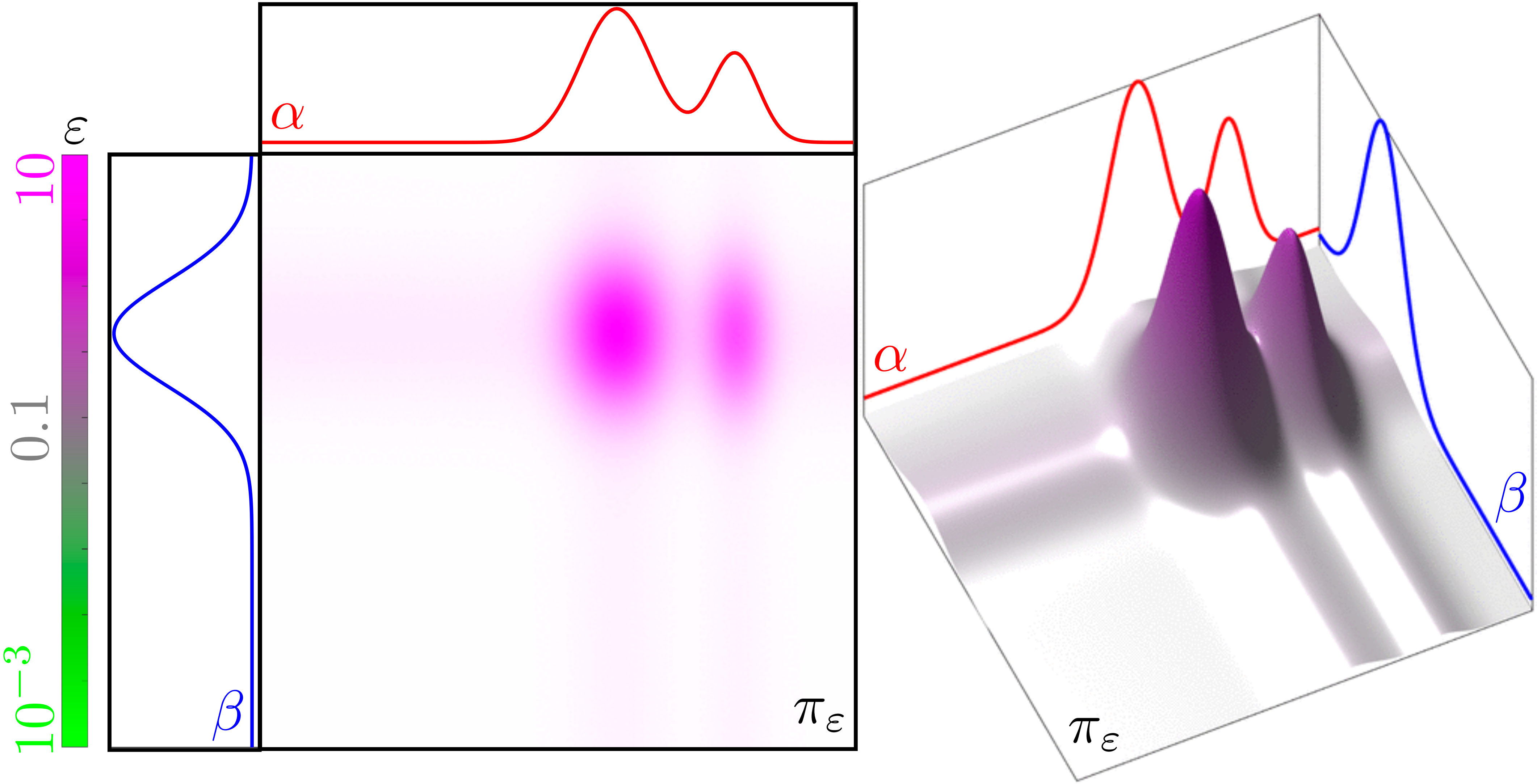
$$\min_{(X,Y)} \{\mathbb{E}(c(X,Y)) + \varepsilon \mathrm{I}(X,Y) \; ; \; X\sim\alpha, Y\sim\beta\}$$

Mutual information



$\varepsilon = 0$ $\varepsilon = .05$ $\varepsilon = 0.2$ $\varepsilon = 1$

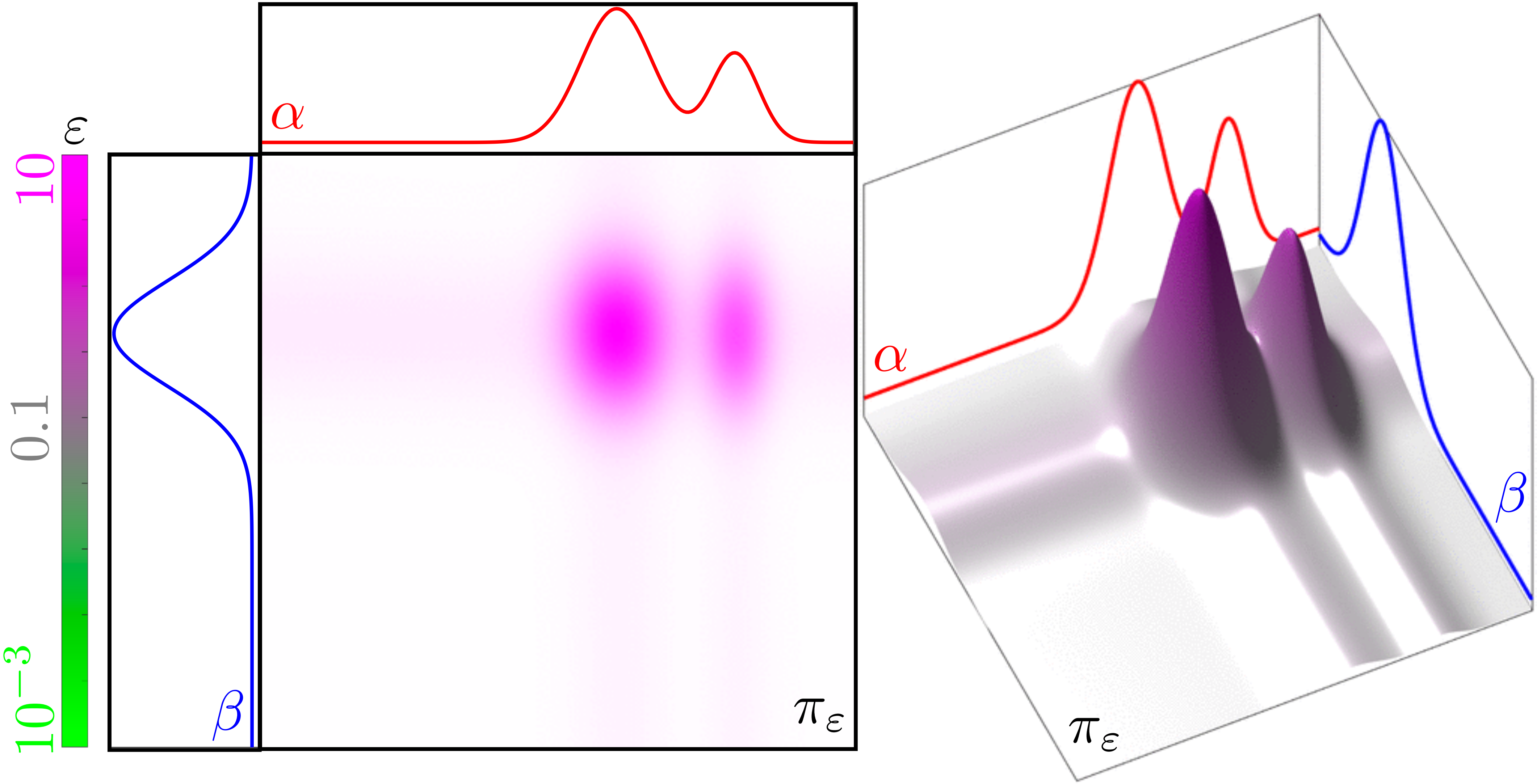
Impact of Regularization



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

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

Impact of Regularization



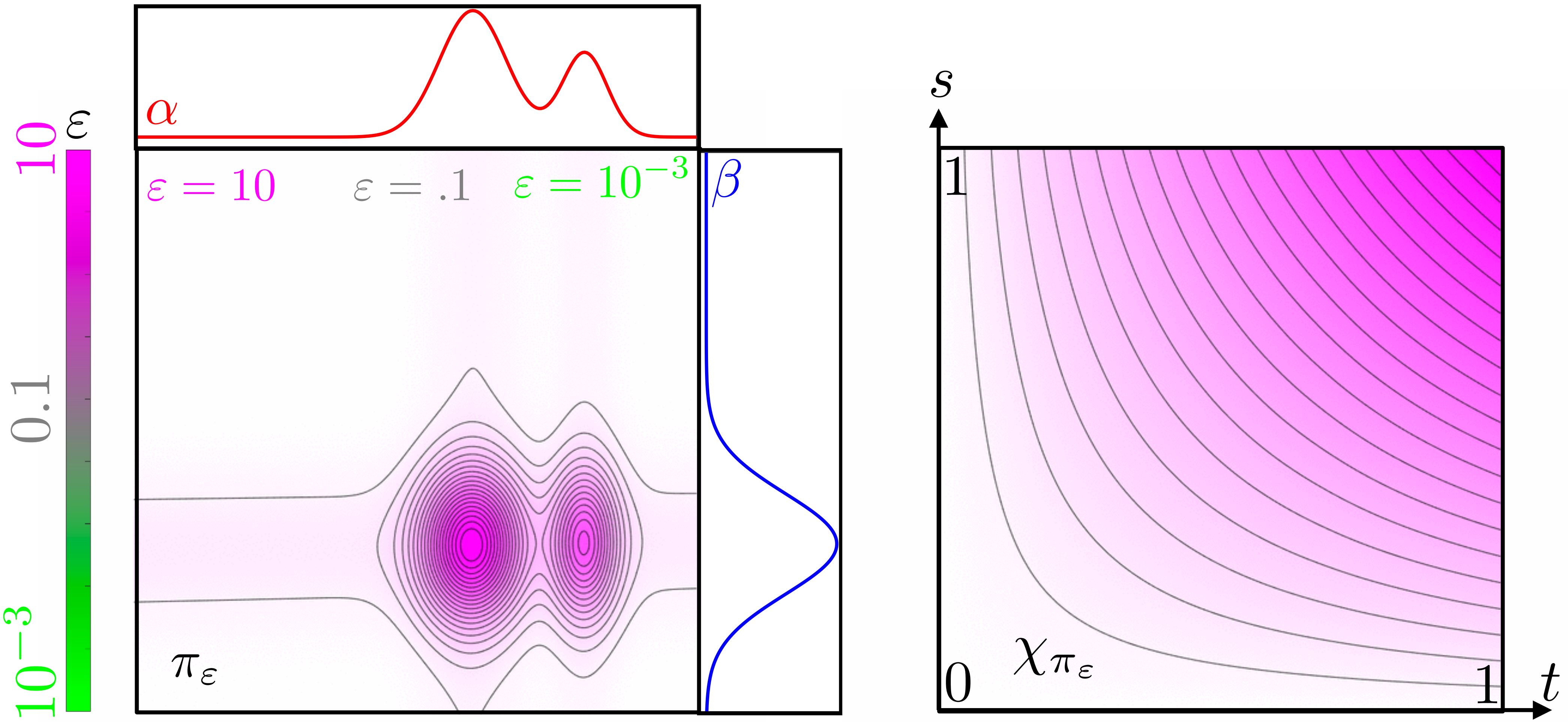
$$\pi_\varepsilon = \operatorname{argmin}_\pi \left\{ \int_{\mathbb{R}^2} \left(\|x-y\|^2 + \varepsilon \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}(x,y) \right) \right) \mathrm{d}\pi(x,y) + \;;\; \pi_1 = \alpha, \pi_2 = \beta \right\}$$

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

Impact of Regularization

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

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



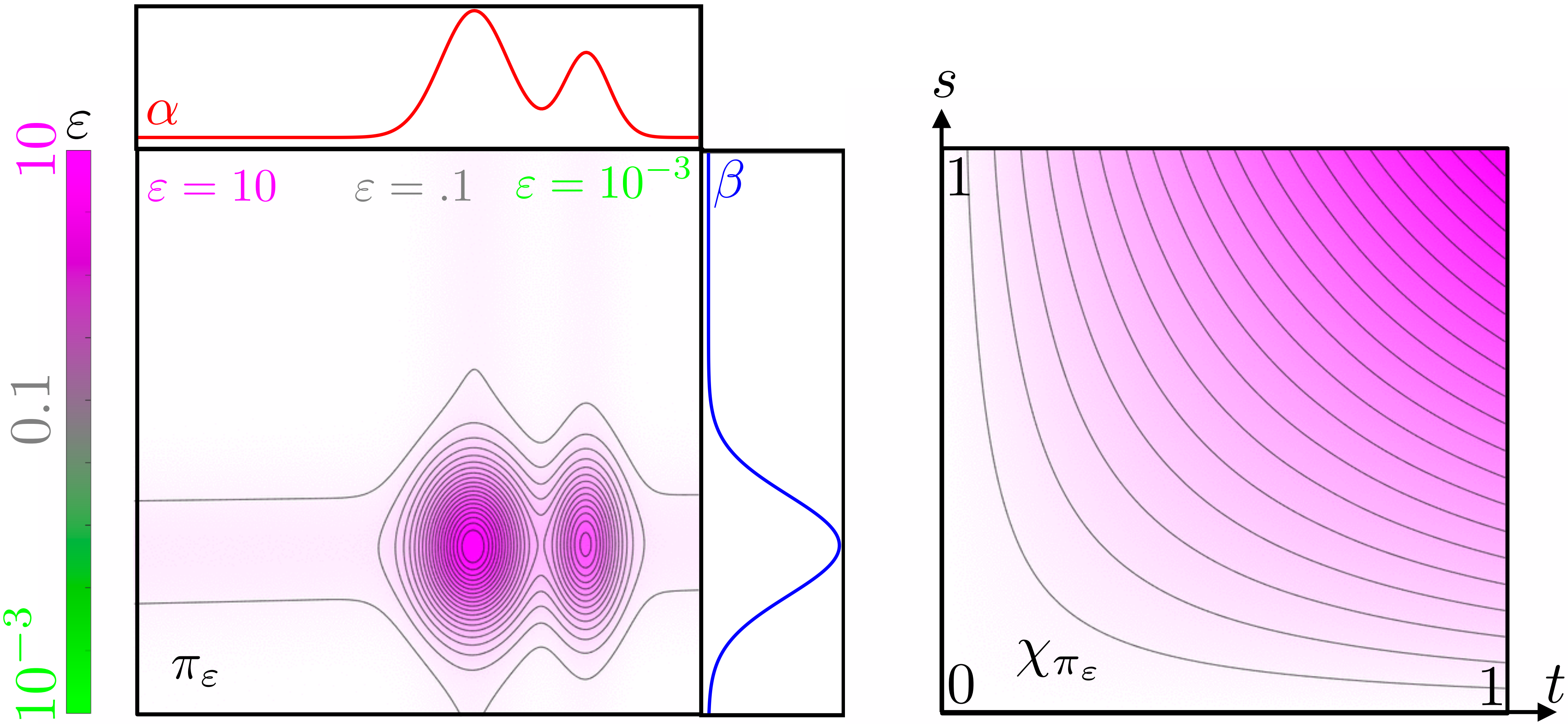
Theorem: $\chi_{\pi_\varepsilon}(s,t) \xrightarrow{\varepsilon \to 0} \min(s,t)$ (dependence)

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

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j$ $\qquad \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

Sinkhorn's Algorithm

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j \qquad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

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

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j$ $\qquad \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

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: $\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$ $\qquad \mathbf{v} \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$

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

Sinkhorn's Algorithm

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j$ $\quad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{d(x_i,y_j)^p}{\varepsilon}}$

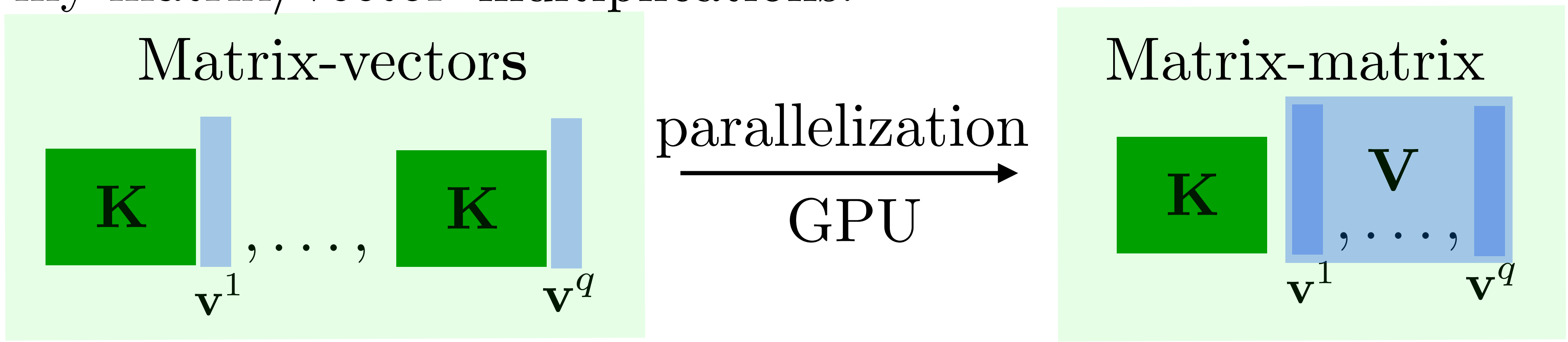
Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

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

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

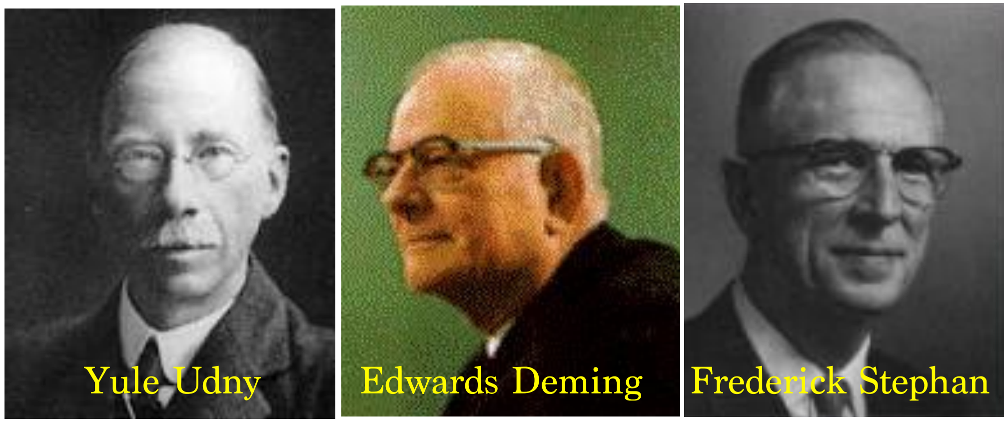
Theorem: [Sinkhorn 1964] $(\mathbf{u}, \mathbf{v})$ converges.

Only matrix/vector multiplications.



→ Convolution on regular grids, separable kernels.

Sinkhorn, IPFP, RAS, ...



Many names:

- Sinkhorn algorithm
- DAD scaling
- Iterative proportional fitting
- Biproportional fitting
- RAS algorithm
- Matrix scaling

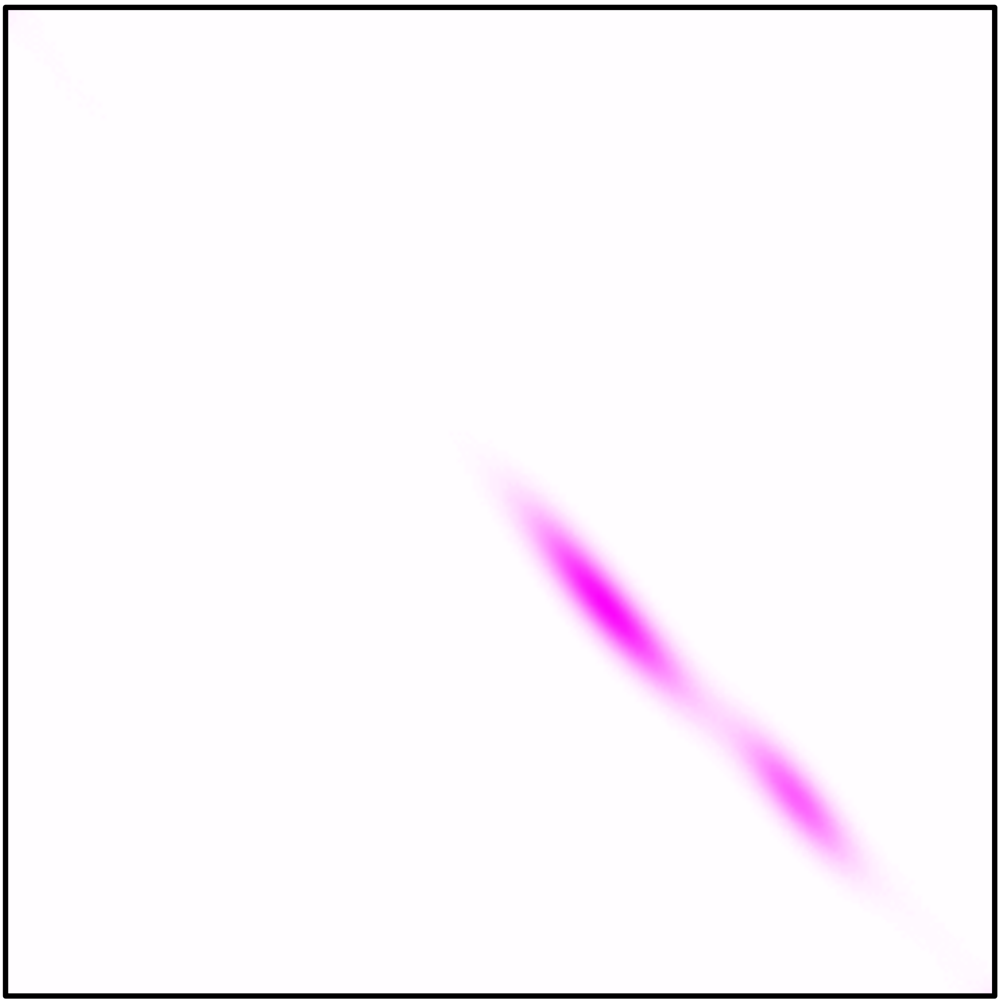
Udny 1912

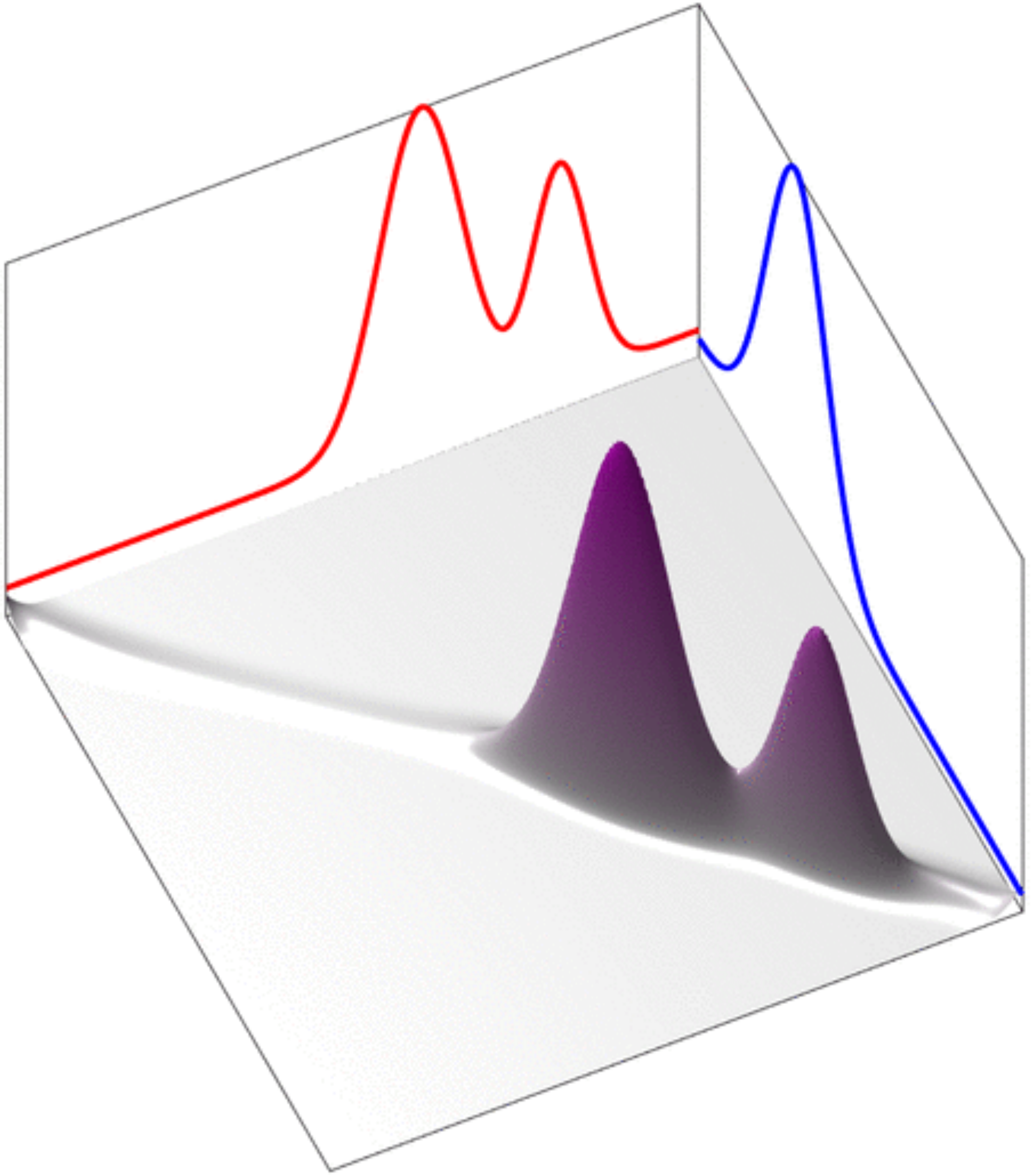
Kruithof, 1937

Deming and Stephan in 1940

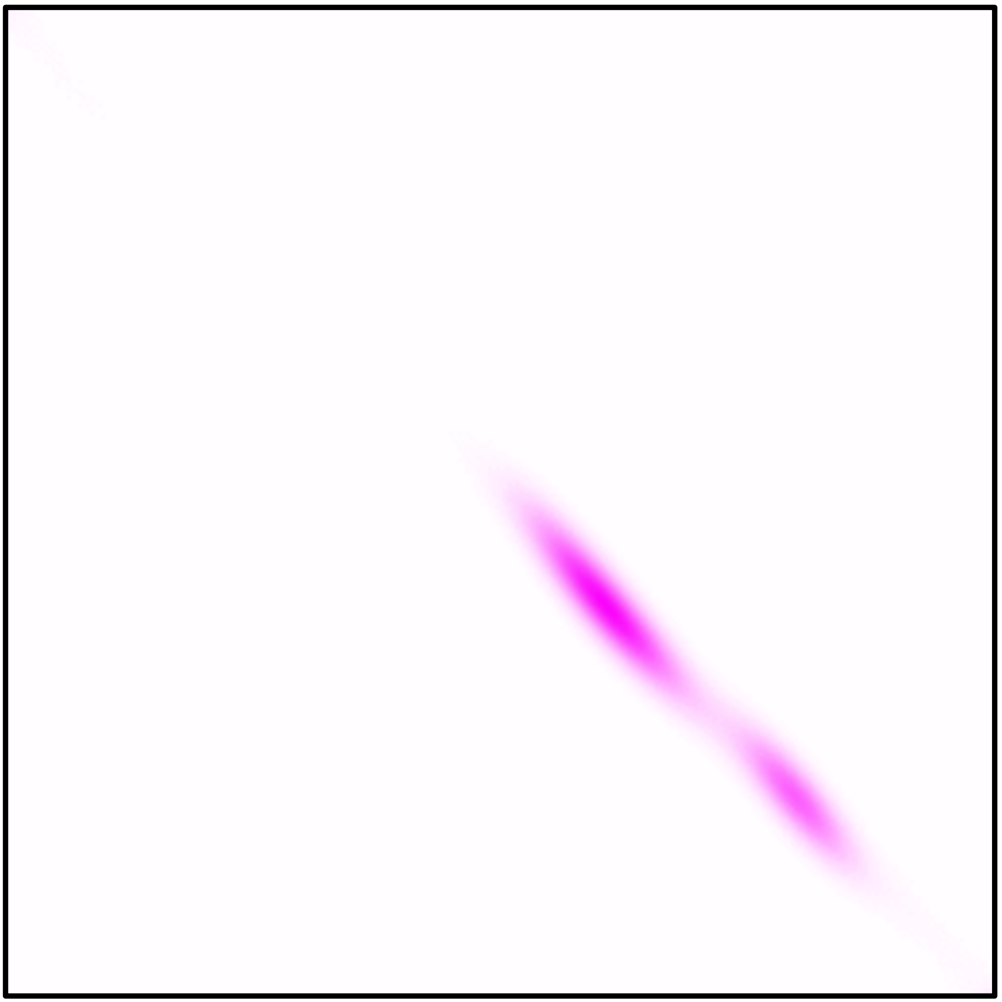
Sinkhorn 1964

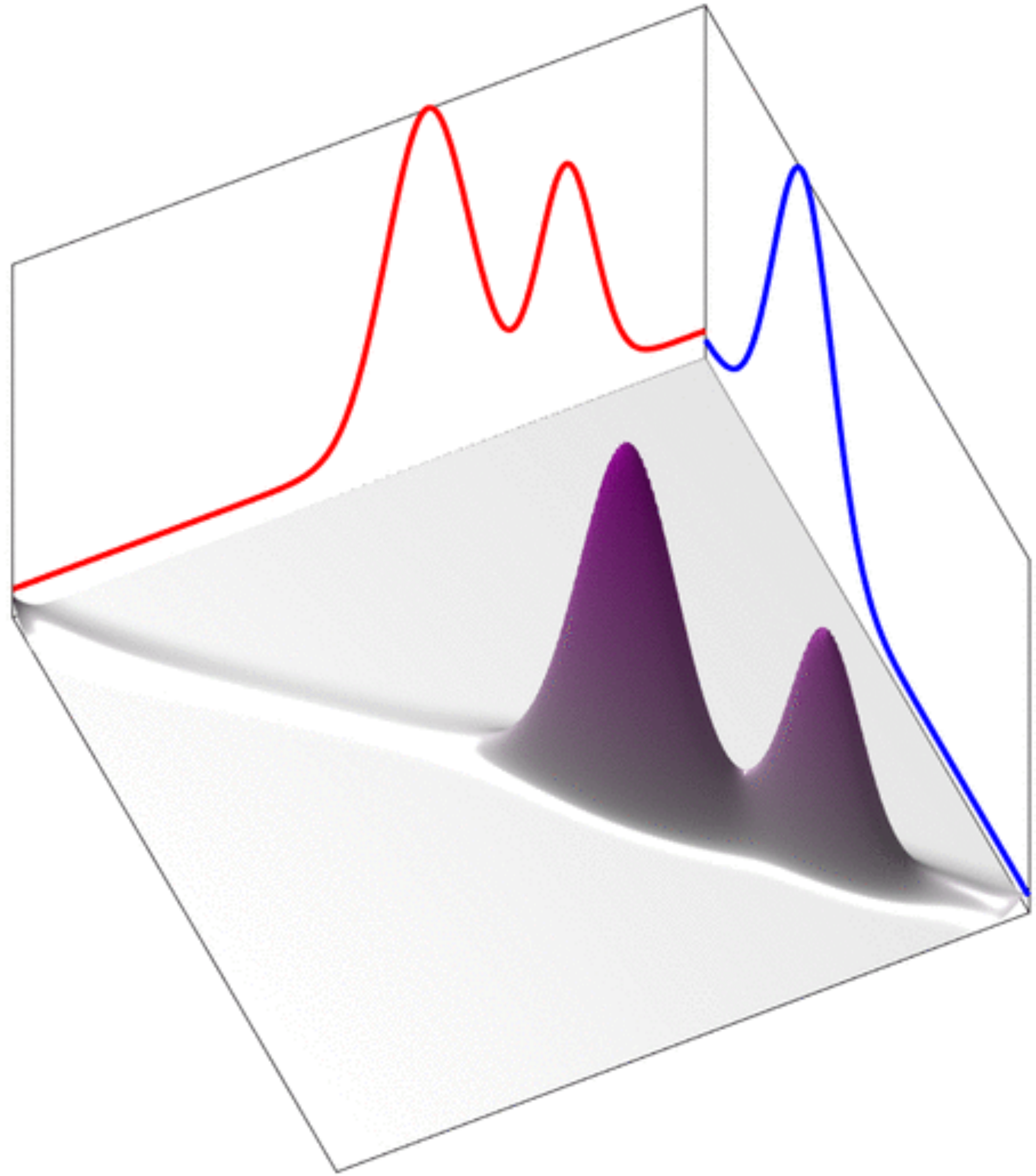
Sinkhorn Evolution





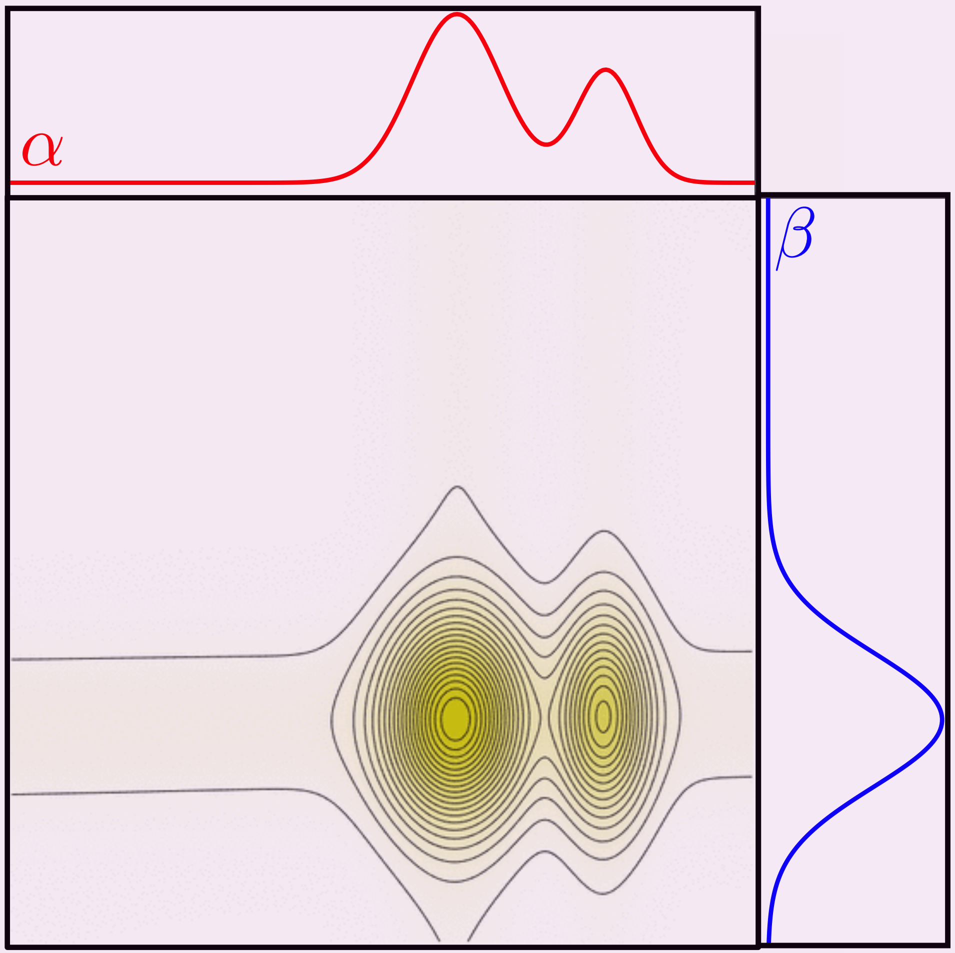
Sinkhorn Evolution





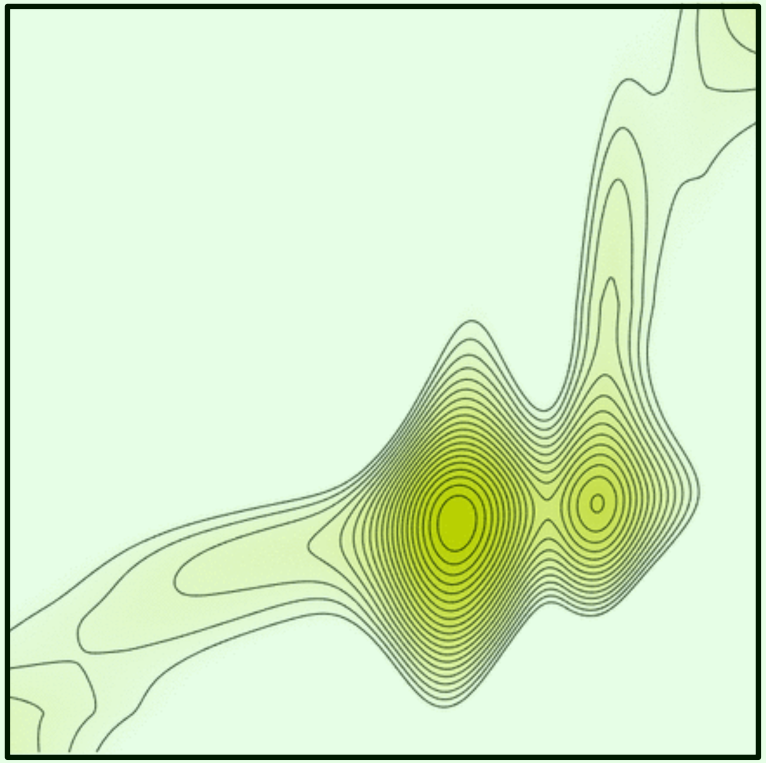
Other Regularizations

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



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

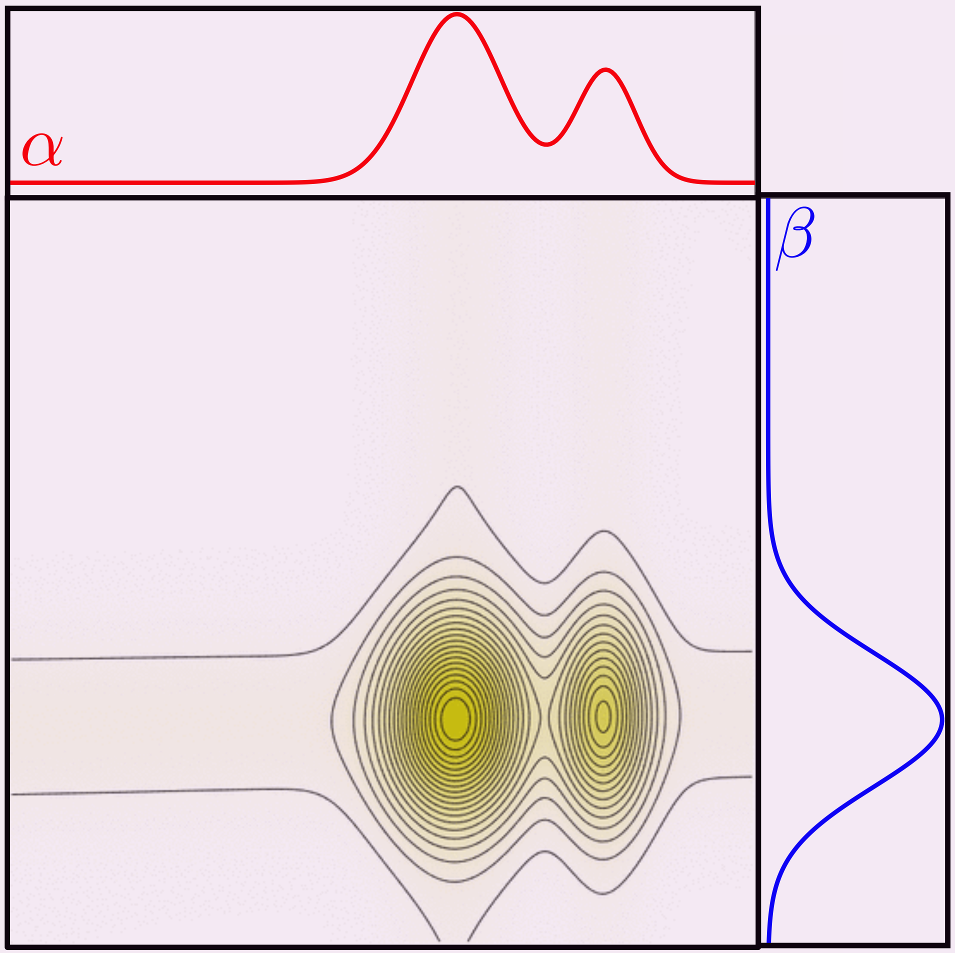
Dykstra's algorithm



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

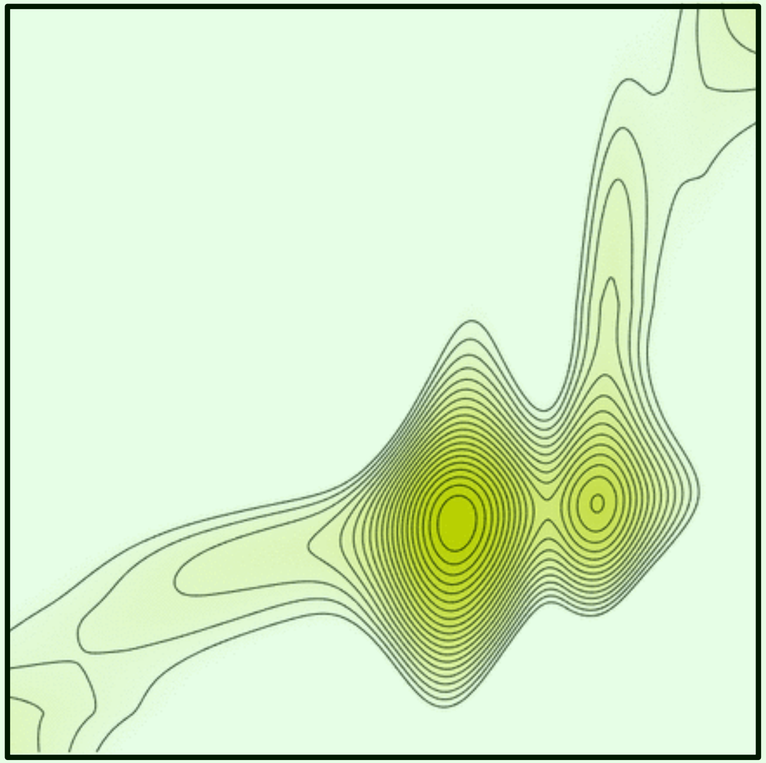
Other Regularizations

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



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

Dykstra's algorithm



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

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]

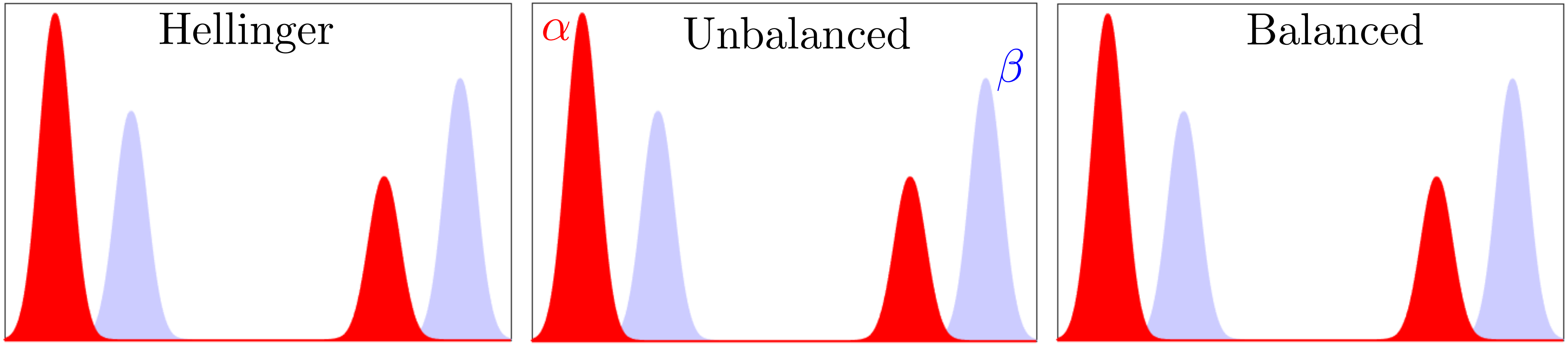
Unbalanced OT

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

[Liereo, Mielke, Savaré 2015]

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

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



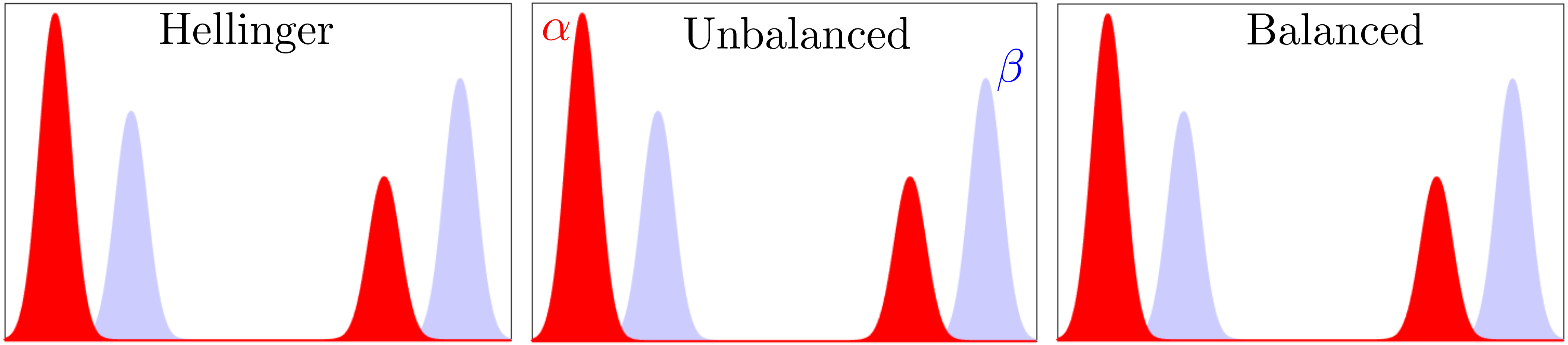
Unbalanced OT

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

[Liereo, Mielke, Savaré 2015]

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

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



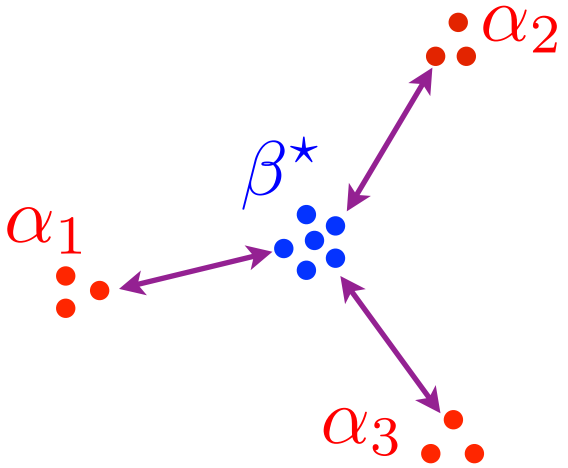
$$\mathrm{W}_{\varepsilon,p}^{\tau,p}(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\pi} \int d^p \mathrm{d}\pi + \tau \mathrm{KL}(\pi_1|\alpha) + \tau \mathrm{KL}(\pi_2|\beta) + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta)$$

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

Wasserstein Barycenters

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

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

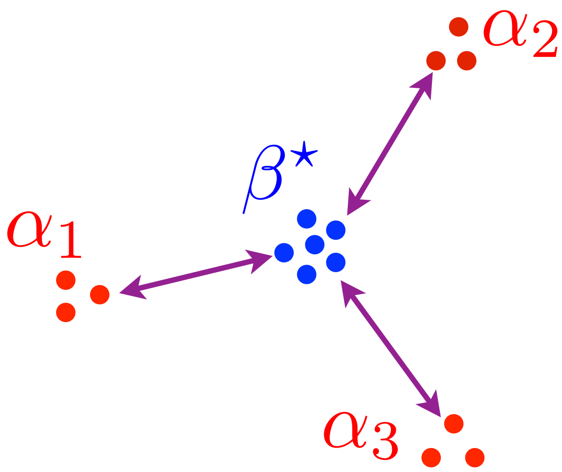


[Solomon et al, SIGGRAPH 2015]

Wasserstein Barycenters

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

$$\beta^\star \in \underset{\beta}{\text{argmin}} \sum_s \lambda_s W_p^p(\alpha_s, \beta)$$



[Solomon et al, SIGGRAPH 2015]

Sinkhorn's algorithm:

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

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

Overview

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

KL divergence:

$$\mathbf{KL}(\mathbf{P}|\mathbf{K}) \overset{\text{def.}}{=} \sum_{i,j} \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}}\right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j}$$

$$\mathrm{KL}(\mathbf{P}|\mathbf{K}) = D_\varphi(\mathbf{P}|\mathbf{K}) \quad \text{for} \quad \varphi(\mathbf{P}) = \sum_{i,j} \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}\rangle + \varepsilon\,\mathrm{KL}(\mathbf{P}|\mathbf{a}\otimes\mathbf{b}) \Leftrightarrow \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \varepsilon\,\mathrm{KL}(\mathbf{P}|\mathbf{K}) \qquad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}$$

[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: $\mathbf{P}^{(\ell)} \to \mathbf{P}^\star = \underset{\mathbf{P}\in\mathcal{C}^1_{\mathbf{a}}\cap\mathcal{C}^1_{\mathbf{b}}}{\mathrm{argmin}}\ \mathrm{KL}(\mathbf{P}|\mathbf{K})$

For affine $(\mathcal{C}^1_{\mathbf{a}}, \mathcal{C}^2_{\mathbf{b}})$,

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

$$\mathbf{U}(\mathbf{a},\mathbf{b}) = \mathcal{C}^1_{\mathbf{a}} \cup \mathcal{C}^2_{\mathbf{b}}$$

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

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

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

$$\mathbf{U}(\mathbf{a},\mathbf{b}) = \mathcal{C}^1_{\mathbf{a}} \cup \mathcal{C}^2_{\mathbf{b}}$$

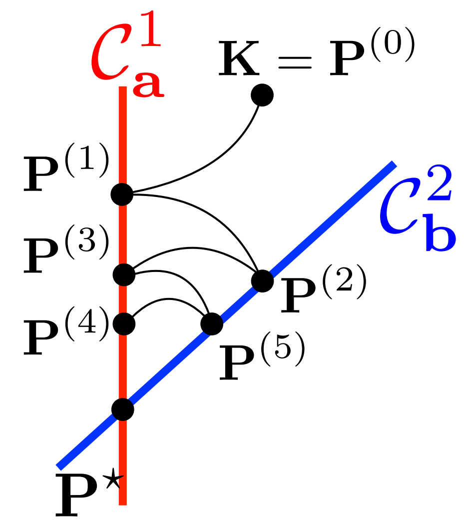
$$\mathcal{C}^1_{\mathbf{a}} \overset{\text{def.}}{=} \{\mathbf{P} \;:\; \mathbf{P}\mathbb{1}_m = \mathbf{a}\}$$

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

[Bregman, 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

[Bregman, 1967]

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

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

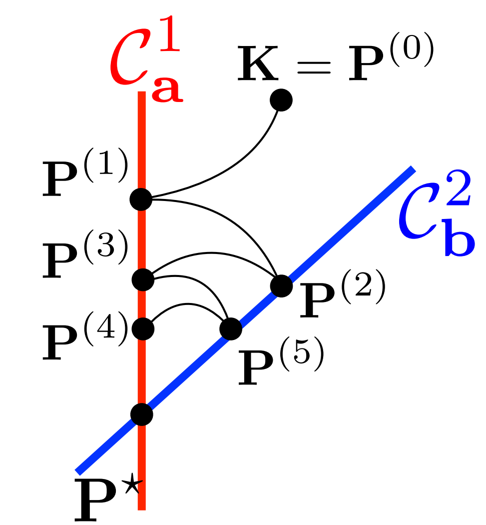
$$\mathbf{U}(\mathbf{a},\mathbf{b}) = \mathcal{C}^1_{\mathbf{a}} \cup \mathcal{C}^2_{\mathbf{b}}$$

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

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

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



Sinkhorn $\iff$ iterative projections.

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

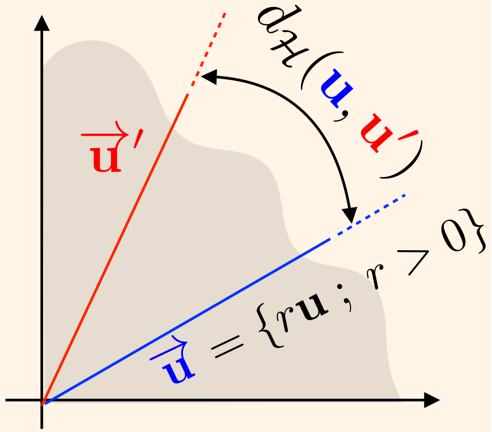
Hilbert Projective Metric

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

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

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

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



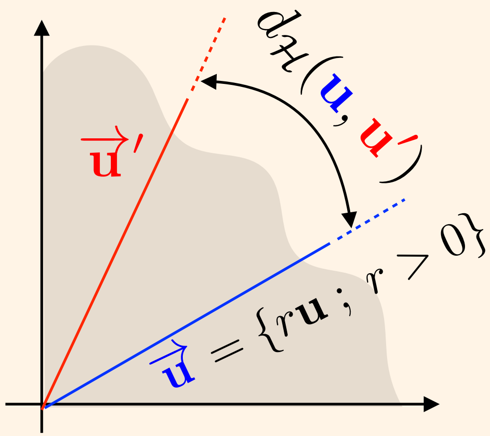
Hilbert Projective Metric

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

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

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

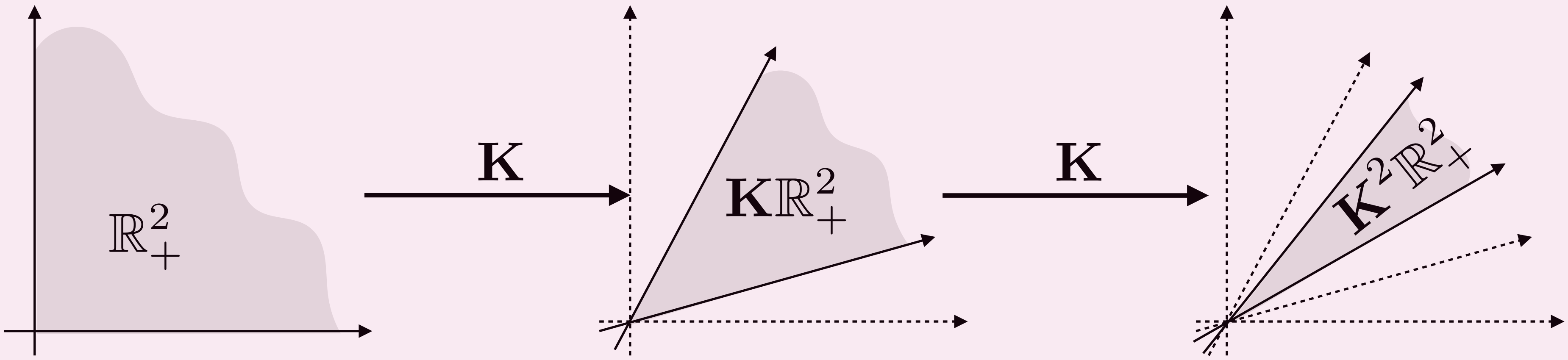
$d_{\mathcal{H}}$ is a distance on the set of rays $\overrightarrow{\mathbf{u}}$.



Birkhoff's contraction theorem:

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

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



Perron Frobenius

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

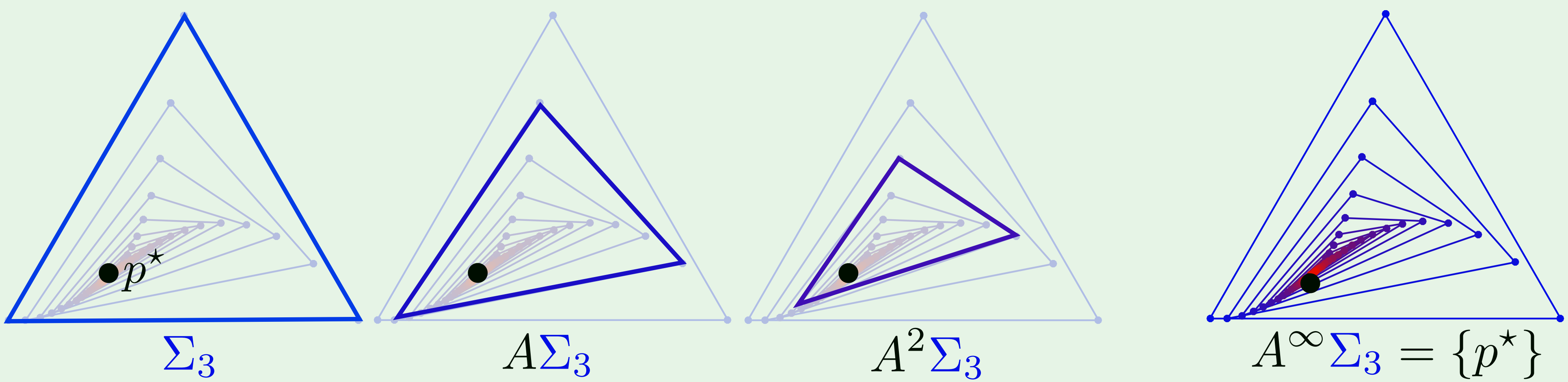
Stochastic matrix: $A \in \mathbb{R}_+^n, \; A^\top \mathbb{1}_k = \mathbb{1}_k$

$A : \Sigma_k \to \Sigma_k$

Theorem: [Perron-Frobenius]

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

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



Sinkhorn under Hilbert's Metric

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

[Franklin and Lorenz, 1989]

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

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

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

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

Local Analysis of Sinkhorn

Sinkhorn fixed point: $\mathbf{f}^{(\ell+1)} = \Phi(\mathbf{f}^{(\ell)})$

$$\Phi \overset{\text{def.}}{=} \Phi_2 \odot \Phi_1 \quad \text{where} \quad \begin{cases} \Phi_1(\mathbf{f}) = \varepsilon \log \mathbf{K}^{\mathrm{T}}(e^{\mathbf{f}/\varepsilon}) - \log(\mathbf{b}), \\ \Phi_2(\mathbf{g}) = \varepsilon \log \mathbf{K}(e^{\mathbf{g}/\varepsilon}) - \log(\mathbf{a}). \end{cases}$$

Proposition: $\partial\Phi(\mathbf{f}) = \mathrm{diag}(\mathbf{a})^{-1} \odot \mathbf{P} \odot \mathrm{diag}(\mathbf{b})^{-1} \odot \mathbf{P}^{\mathrm{T}}.$

For ℓ large enough, $\|\mathbf{f}^{(\ell)} - \mathbf{f}\| = O((1-\kappa)^{\ell})$

Local Analysis of Sinkhorn

Sinkhorn fixed point: $\mathbf{f}^{(\ell+1)} = \Phi(\mathbf{f}^{(\ell)})$

$$\Phi \overset{\text{def.}}{=} \Phi_2 \odot \Phi_1 \quad \text{where} \quad \begin{cases} \Phi_1(\mathbf{f}) = \varepsilon \log \mathbf{K}^{\mathrm{T}}(e^{\mathbf{f}/\varepsilon}) - \log(\mathbf{b}), \\ \Phi_2(\mathbf{g}) = \varepsilon \log \mathbf{K}(e^{\mathbf{g}/\varepsilon}) - \log(\mathbf{a}). \end{cases}$$

Proposition: $\partial\Phi(\mathbf{f}) = \mathrm{diag}(\mathbf{a})^{-1} \odot \mathbf{P} \odot \mathrm{diag}(\mathbf{b})^{-1} \odot \mathbf{P}^{\mathrm{T}}.$

For ℓ large enough, $\|\mathbf{f}^{(\ell)} - \mathbf{f}\| = O((1-\kappa)^\ell)$

Global rate: $\kappa \sim e^{-\frac{1}{\varepsilon}}$
[Franklin and Lorenz, 1989]

⟷

Local rate: $\kappa \sim \varepsilon$
[Robert Berman 2017]

Local Analysis of Sinkhorn

Sinkhorn fixed point: $\mathbf{f}^{(\ell+1)} = \Phi(\mathbf{f}^{(\ell)})$

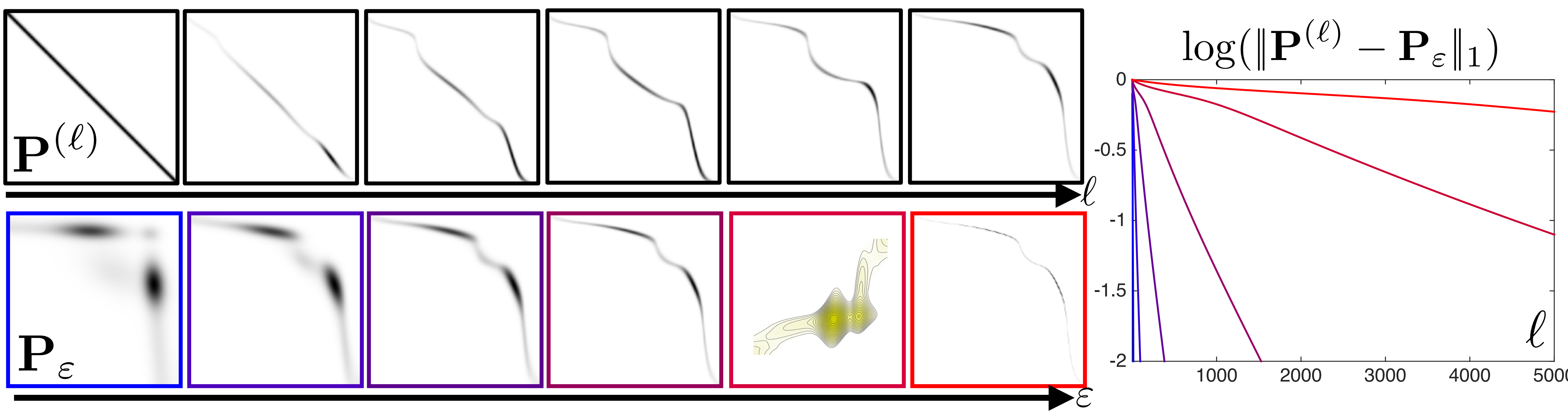
$$\Phi \overset{\text{def.}}{=} \Phi_2 \odot \Phi_1 \quad \text{where} \quad \begin{cases} \Phi_1(\mathbf{f}) = \varepsilon \log \mathbf{K}^{\mathrm{T}}(e^{\mathbf{f}/\varepsilon}) - \log(\mathbf{b}), \\ \Phi_2(\mathbf{g}) = \varepsilon \log \mathbf{K}(e^{\mathbf{g}/\varepsilon}) - \log(\mathbf{a}). \end{cases}$$

Proposition: $\partial\Phi(\mathbf{f}) = \mathrm{diag}(\mathbf{a})^{-1} \odot \mathbf{P} \odot \mathrm{diag}(\mathbf{b})^{-1} \odot \mathbf{P}^{\mathrm{T}}.$

For ℓ large enough, $\|\mathbf{f}^{(\ell)} - \mathbf{f}\| = O((1-\kappa)^{\ell})$

Global rate: $\kappa \sim e^{-\frac{1}{\varepsilon}}$
[Franklin and Lorenz, 1989]

Local rate: $\kappa \sim \varepsilon$
[Robert Berman 2017]



Overview

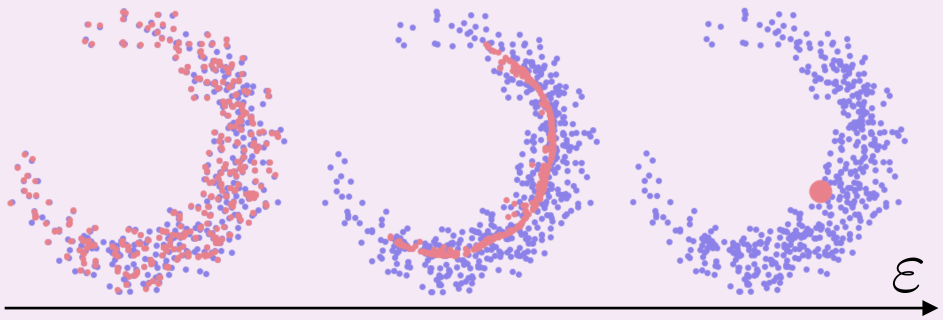
- Entropic Regularization and Sinkhorn
- Convergence Analysis
- **Sinkhorn Divergences**
- Generative Model Fitting

Sinkhorn Divergences

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

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

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

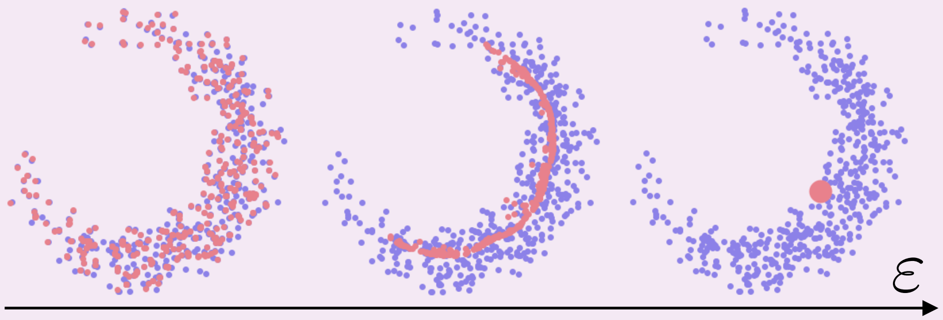


Sinkhorn Divergences

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

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

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



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

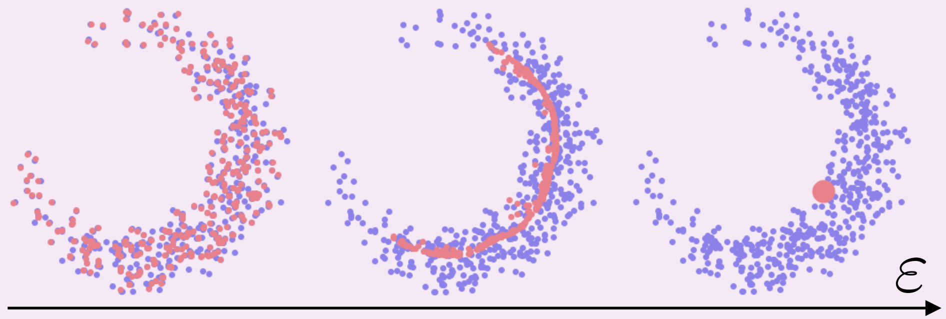
[Ramdas, García Trillos, Cuturi, 2017]

Sinkhorn Divergences

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

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

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



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

[Ramdas, García Trillos, Cuturi, 2017]

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

[Léonard 2012]
[Carlier et al 2017]

[Ramdas, García Trillos, Cuturi, 2017]

Kernel norms (MMD): $\|\xi\|^2_{-d^p} \overset{\text{def.}}{=} -\int_{\mathcal{X}^2} d(x,y)^p \mathrm{d}\xi(x)\mathrm{d}\xi(y)$

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

Sinkhorn Divergences

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

concave   concave

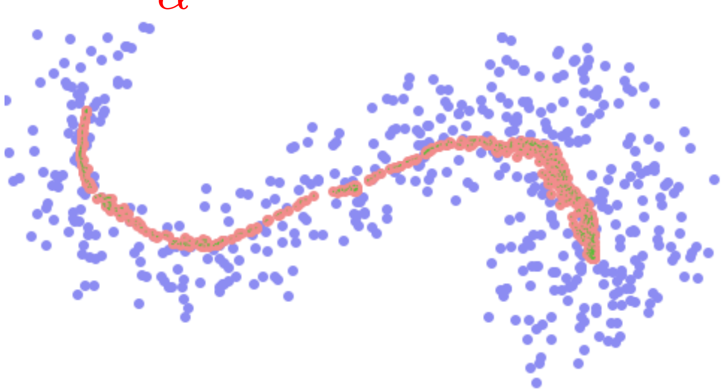
Theorem: [Feydy, Séjourné, P, Vialard, Trouvé, Amari 2018]

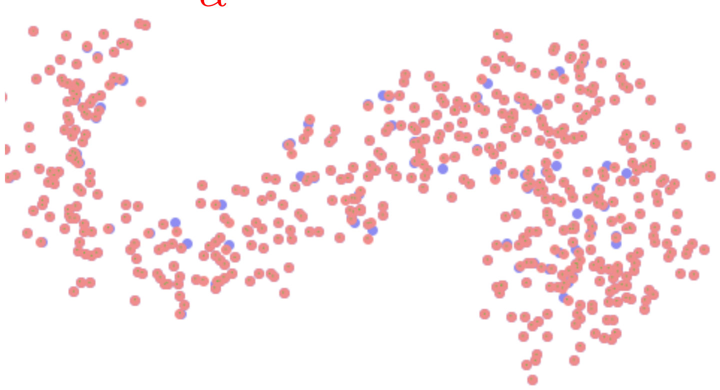
If $e^{-\frac{d^p}{\varepsilon}}$ is positive:

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

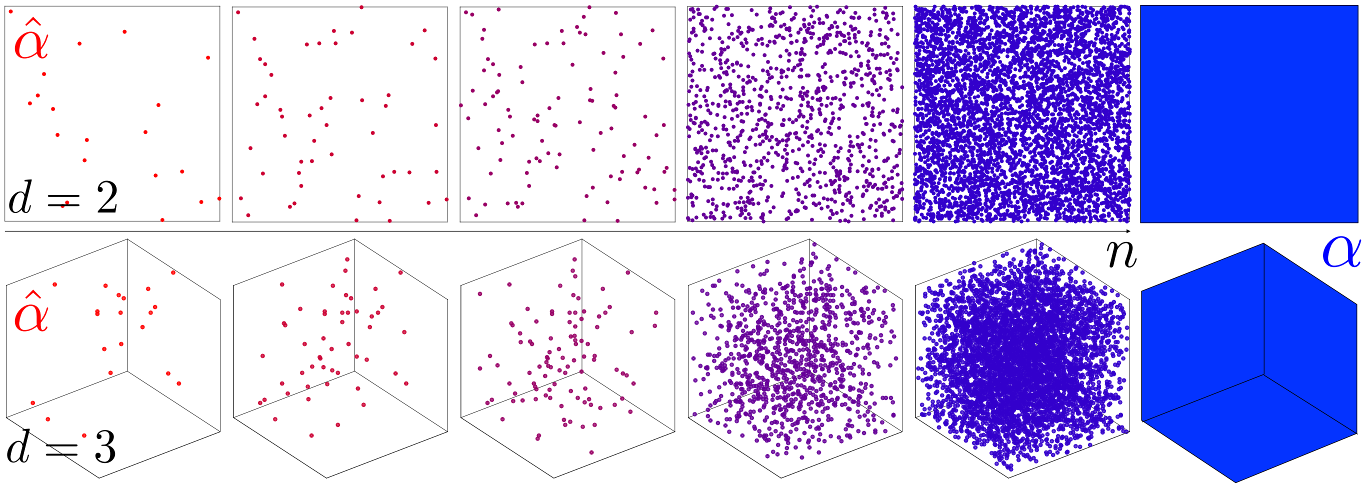
$$\overline{W}_{\varepsilon,p}(\alpha_n,\beta) \to 0 \iff \alpha_n \stackrel{\text{weak}*}{\rightharpoonup} \beta$$

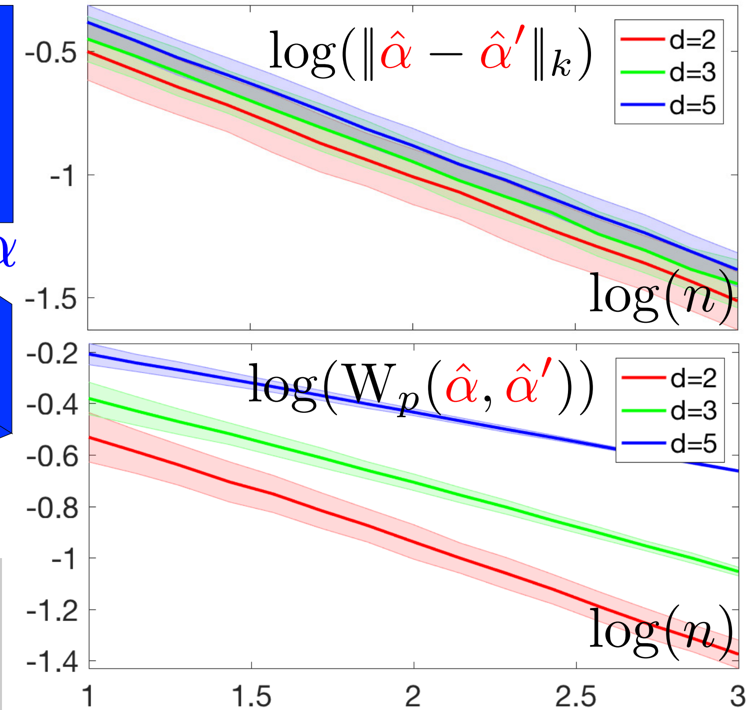
$\min_{\alpha} W^p_{\varepsilon,p}(\alpha,\beta)$ $\min_{\alpha} \overline{W}^p_{\varepsilon,p}(\alpha,\beta)$





Sample Complexity





Theorem:

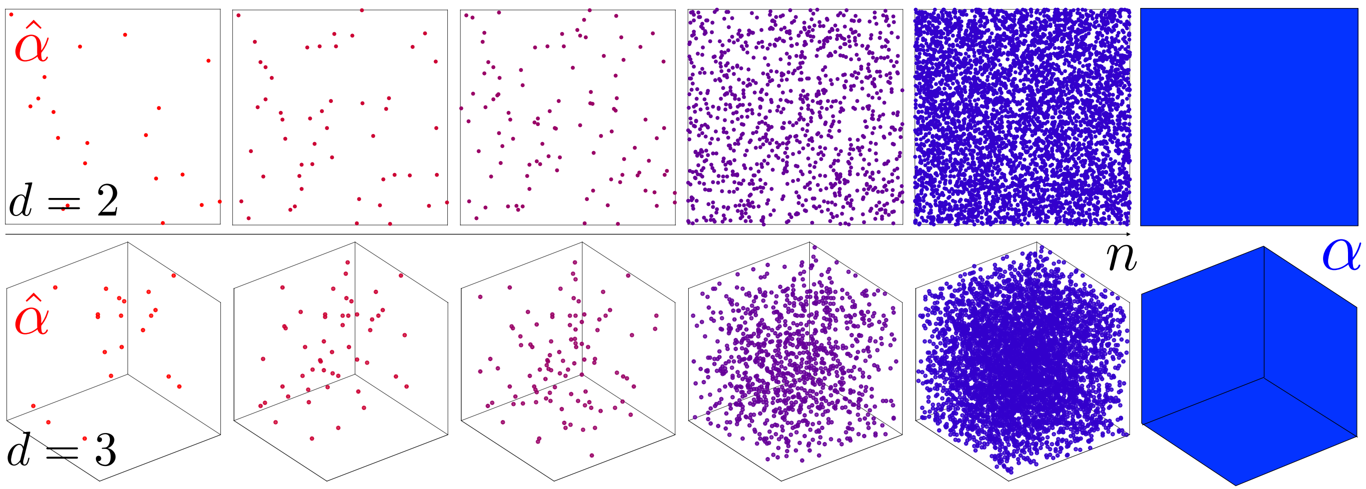
$$\mathbb{E}(|\mathrm{W}_p(\hat{\alpha}, \hat{\beta}) - \mathrm{W}_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$$

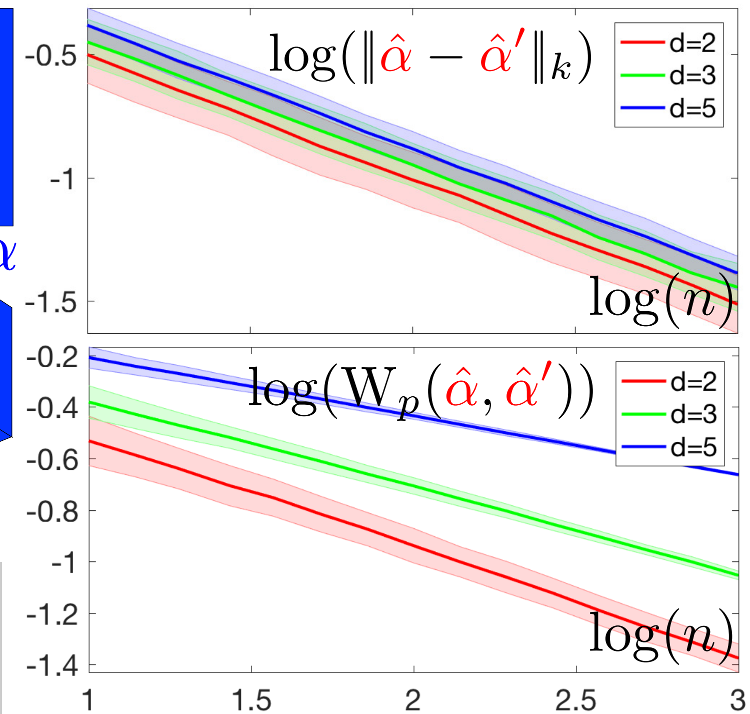
$$\mathbb{E}(|\|\hat{\alpha} - \hat{\beta}\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}})$$

Optimal transport: suffers from curse of dimensionality.

→ Adapt to support dimensionality [Weed, Bach 2017]

Sample Complexity

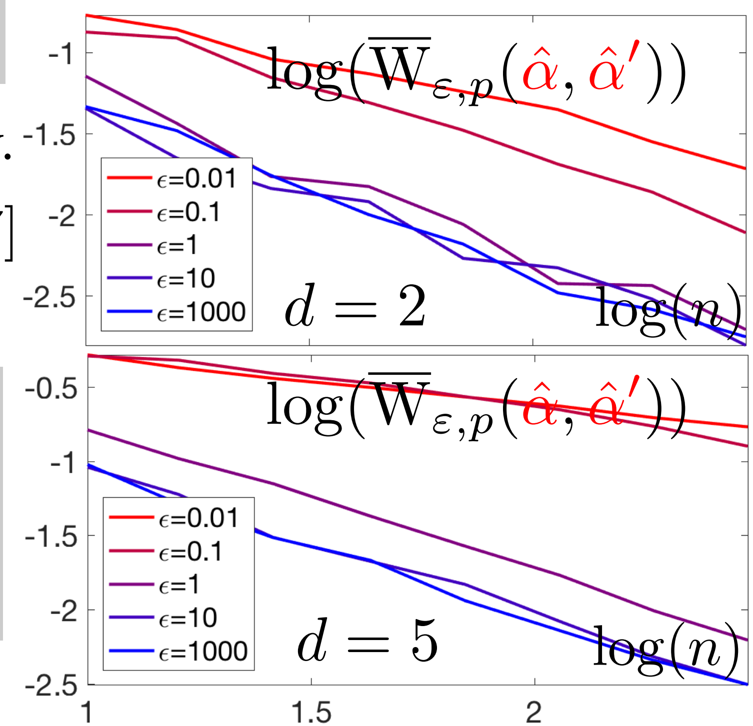




Theorem:
$$\mathbb{E}(|W_p(\hat{\alpha}, \hat{\beta}) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$$
$$\mathbb{E}(|\|\hat{\alpha} - \hat{\beta}\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}})$$

Optimal transport: suffers from curse of dimensionality.

→ Adapt to support dimensionality [Weed, Bach 2017]



Theorem: [Genevay, Bach, P, Cuturi]
$$\mathbb{E}(|\overline{W}_{\varepsilon,p}(\hat{\alpha}, \hat{\beta}) - \overline{W}_{\varepsilon,p}(\alpha, \beta)|) = O(\varepsilon^{-\frac{d}{2}} n^{-\frac{1}{2}})$$

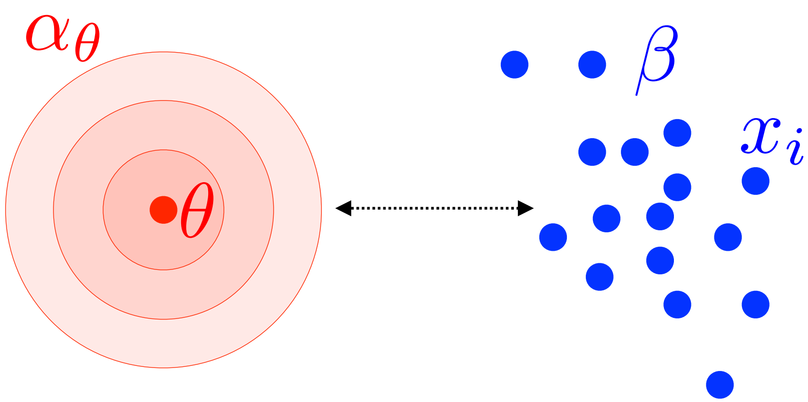
Overview

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

Density Fitting and Generative Models

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

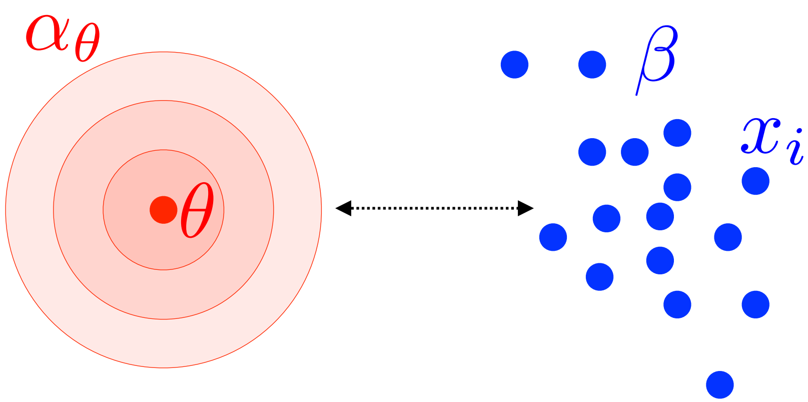
Parametric model: $\theta \mapsto \alpha_\theta$



Density Fitting and Generative Models

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

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



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

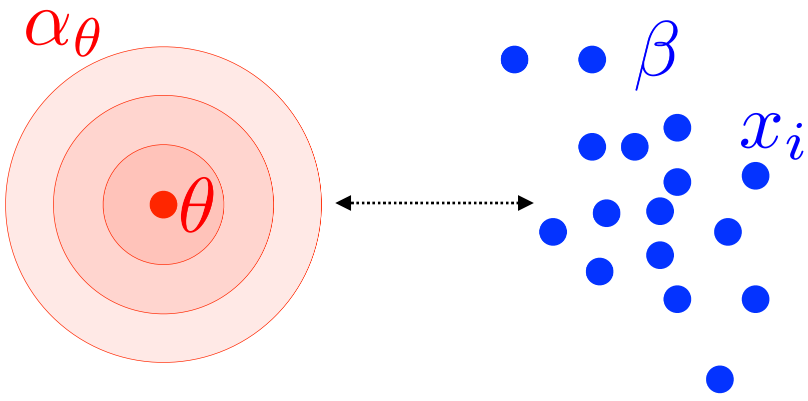
$$\min_\theta \widehat{\mathrm{KL}}(\alpha_\theta|\beta) \stackrel{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i))$$

Maximum likelihood (MLE)

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

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

Maximum likelihood (MLE)

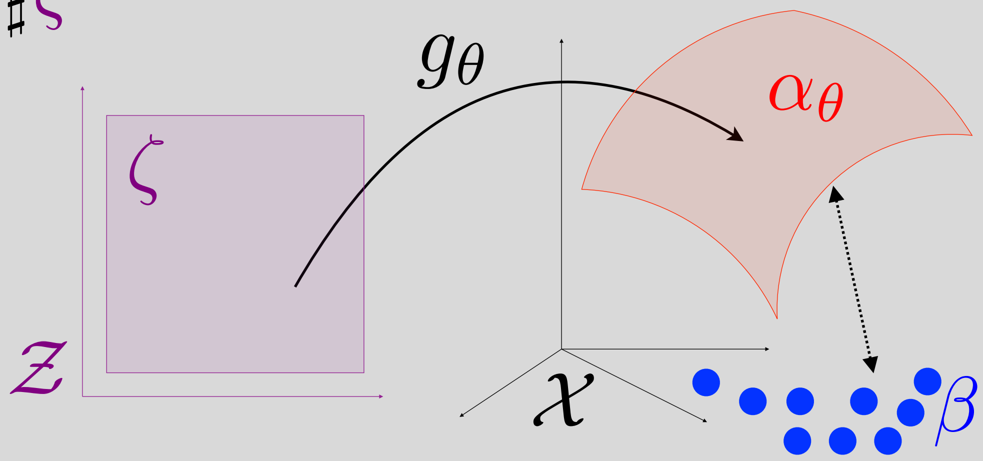
Generative model fit: $\alpha_\theta = g_{\theta,\sharp}\zeta$

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

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

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

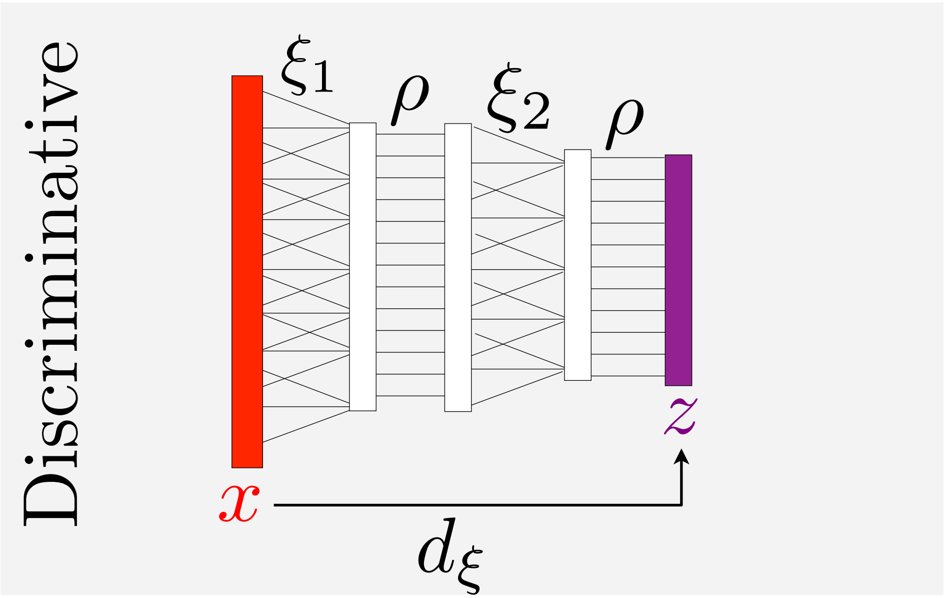


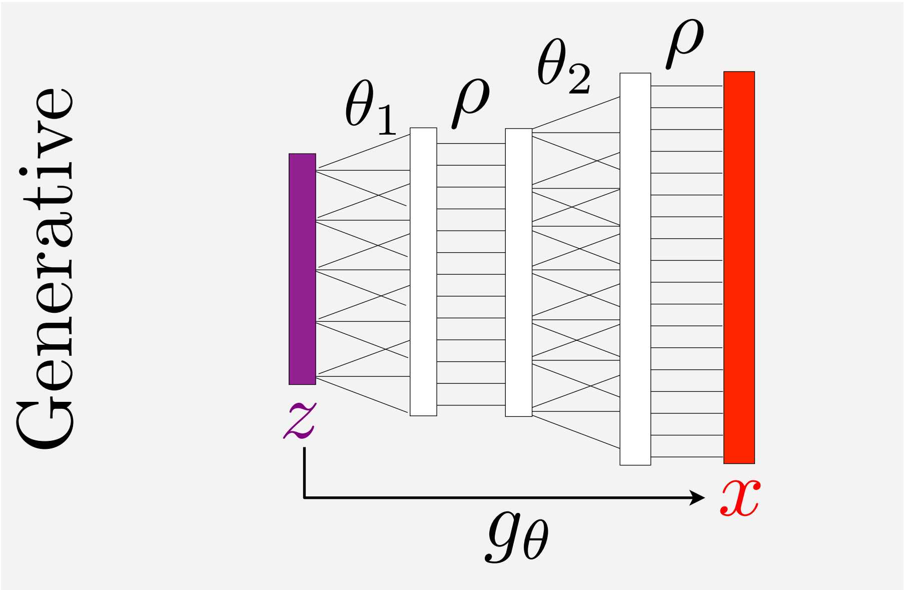
Deep Discriminative vs Generative Models

Deep networks:

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

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



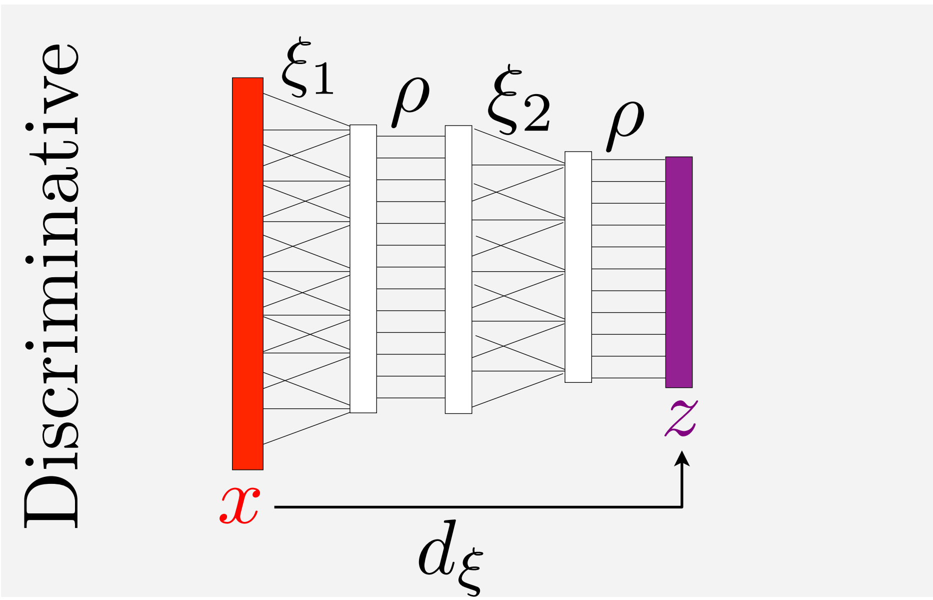


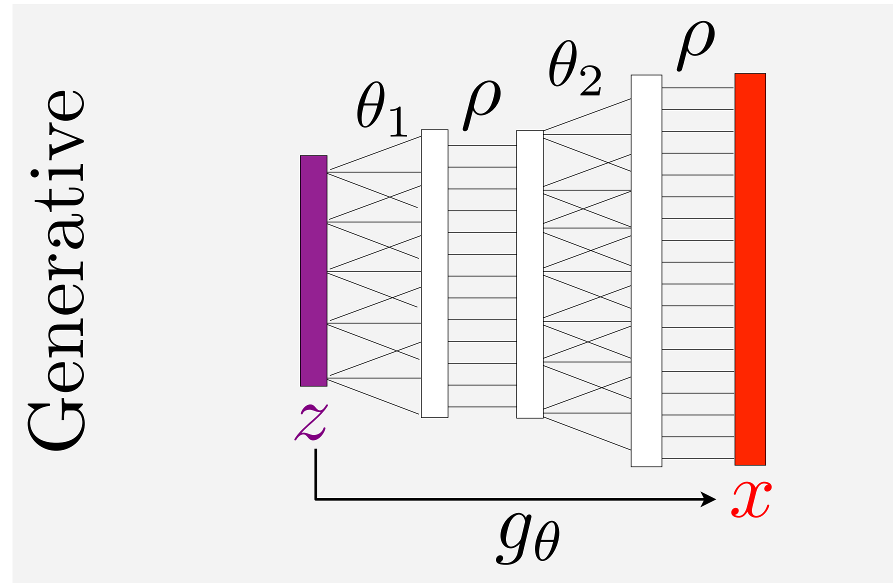
Deep Discriminative vs Generative Models

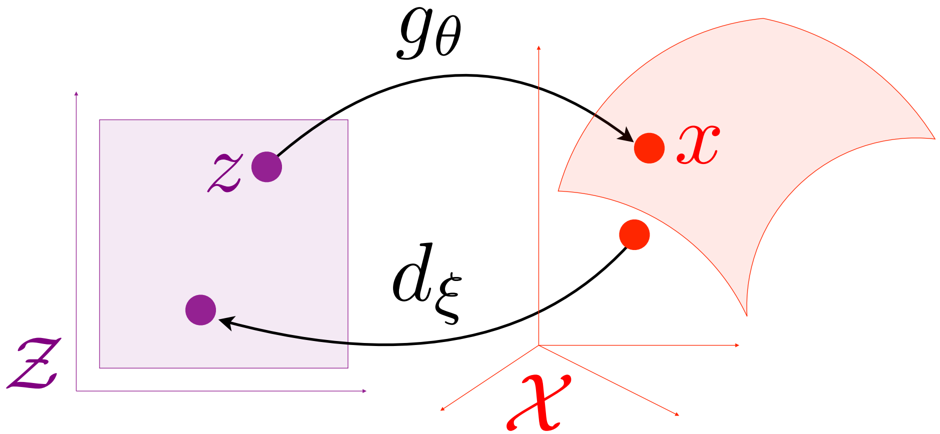
Deep networks:

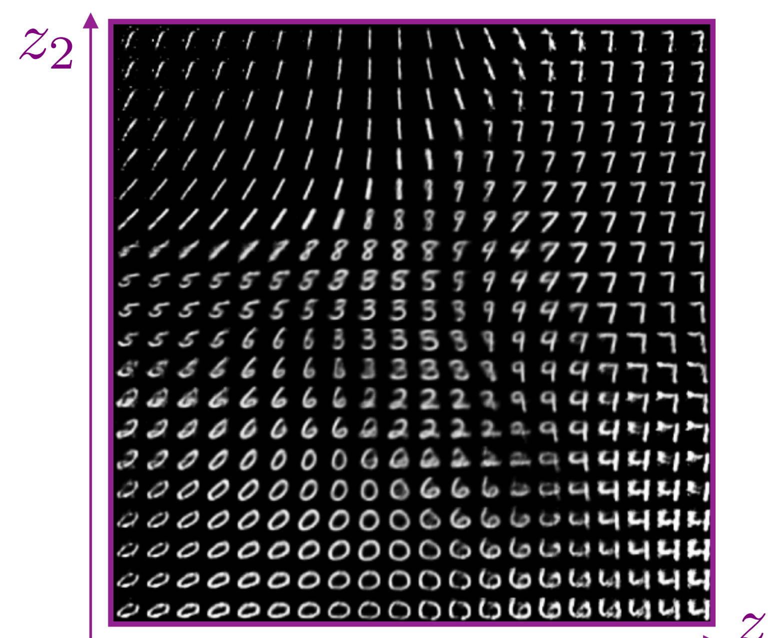
$$d_\xi(x) = \rho(\xi_K(\ldots\rho(\xi_2(\rho(\xi_1(x)\ldots)$$

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

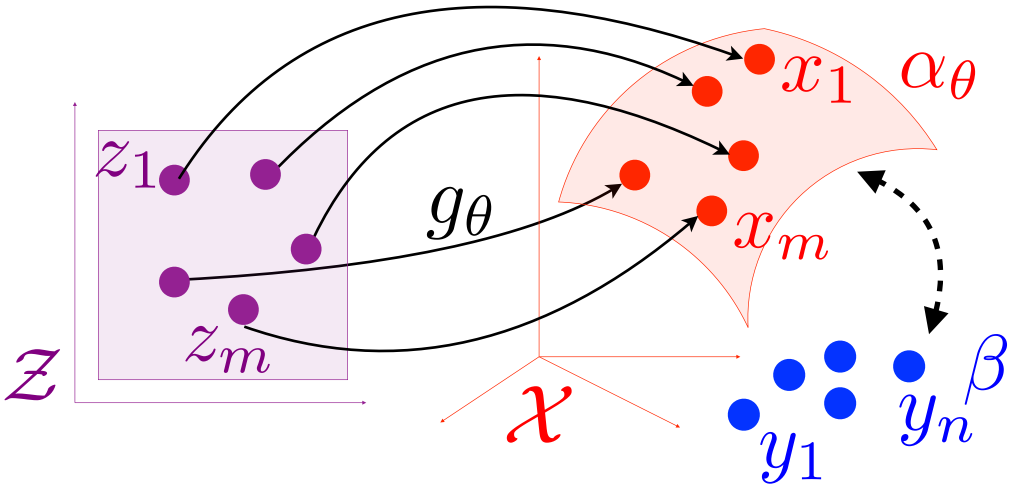








Training Architecture



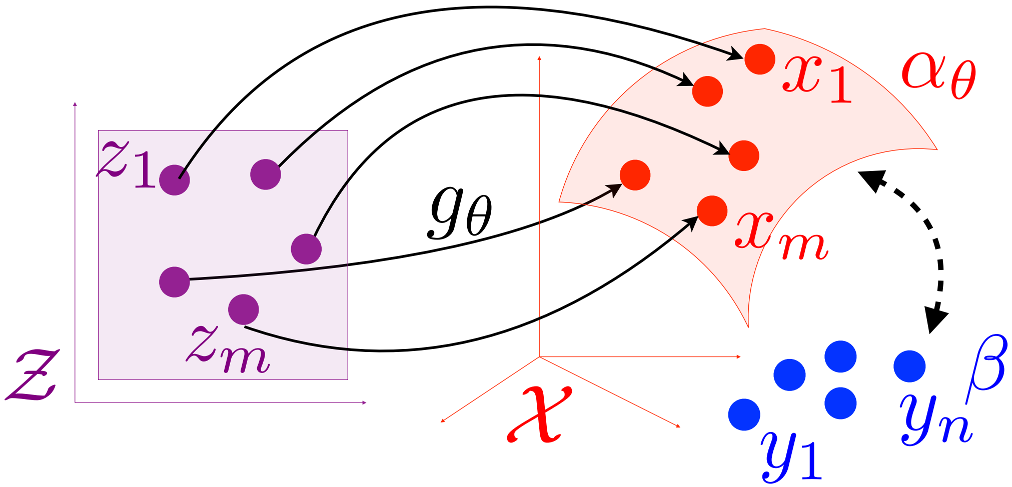
$$\min_\theta \mathcal{E}(\theta) \overset{\text{def.}}{=} \overline{\mathrm{W}}_{\varepsilon,p}^p(\alpha_\theta, \beta)$$

Stochastic gradient descent

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

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

Training Architecture

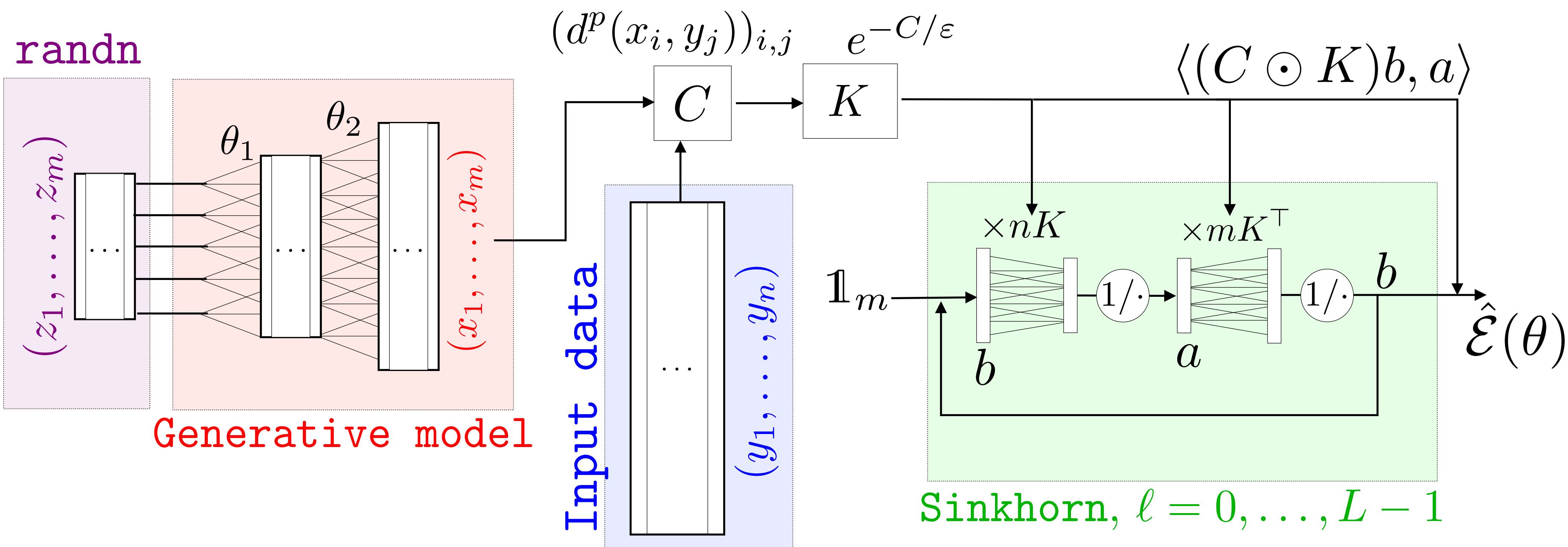


$$\min_\theta \mathcal{E}(\theta) \overset{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

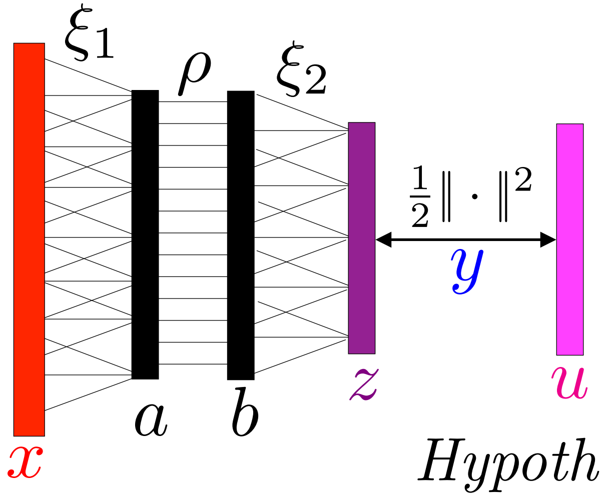
$$\theta \leftarrow \theta - \tau \nabla \hat{\mathcal{E}}(\theta)$$

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



Automatic Differentiation

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



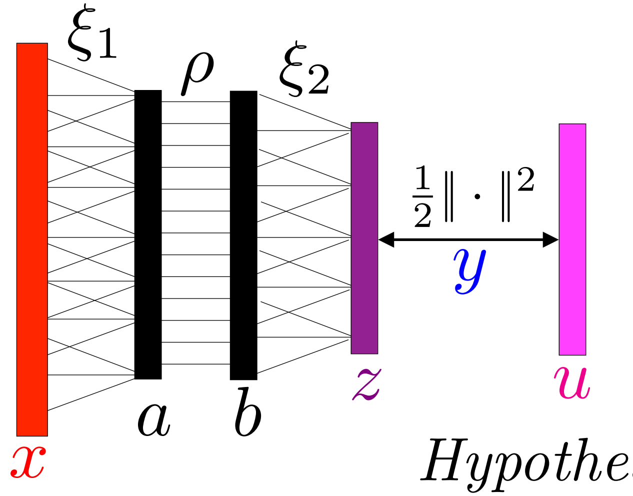

```
function y = E(x)
  a = xi1*x
  b = rho(a)
  z = xi2*b
  y = 1/2*norm(z-u)^2
```

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \ldots)$ 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: $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ 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 $\nabla\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^n$?

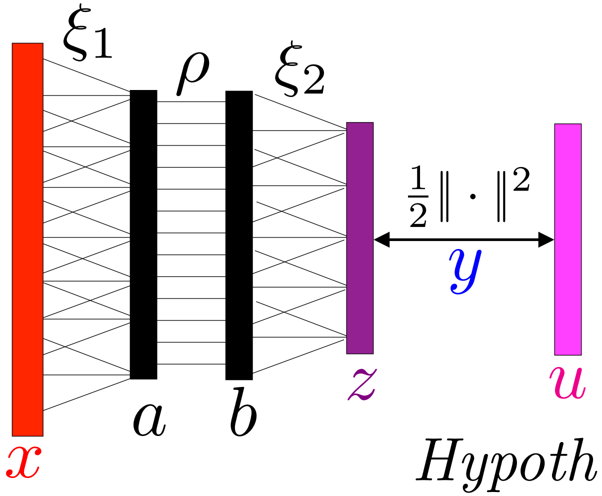
Finite differences:

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

$K(n+1)$ 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.

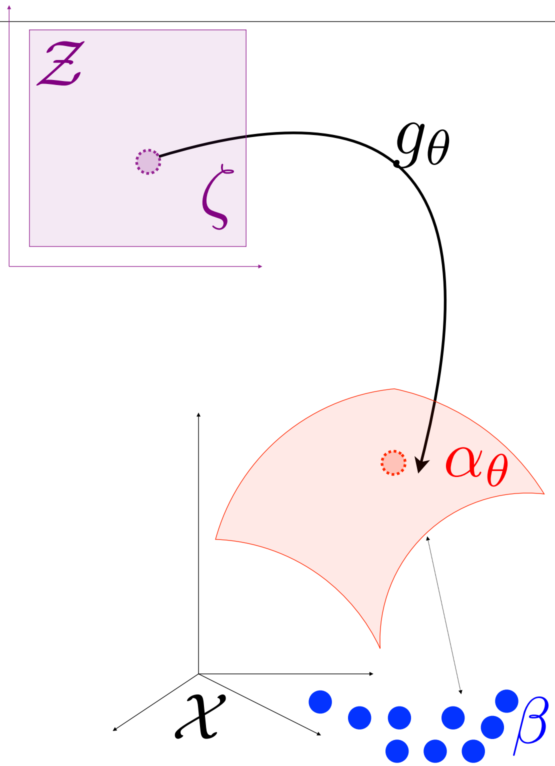
Theorem: there is an algorithm to compute $\nabla\mathcal{E}$ in $O(K)$ operations. [Seppo Linnainmaa, 1970]

This algorithm is reverse mode automatic differentiation

Examples of Images Generation

Inputs β

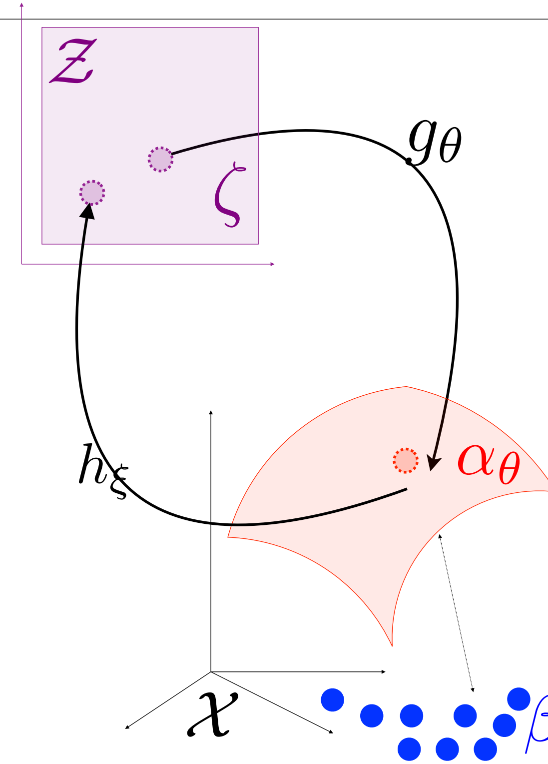
Generated α_θ

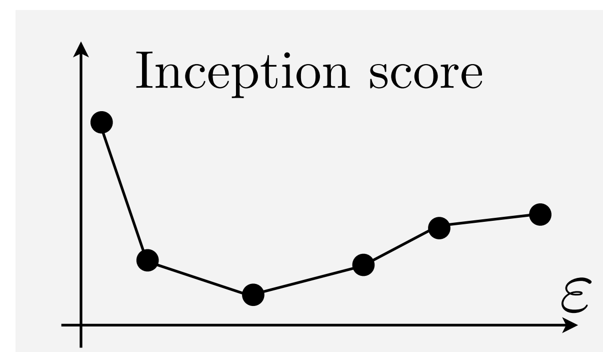


Examples of Images Generation

Inputs β

Generated α_θ



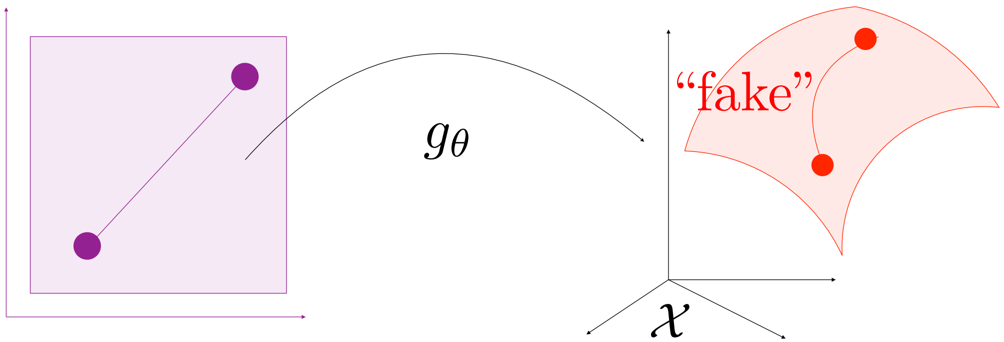


→ Need to learn the metric $d(x,y) = \|h_\xi(x) - h_\xi(y)\|$ (GANs)

→ 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