Fitting Generative Models with Optimal Transport

Marco Cuturi

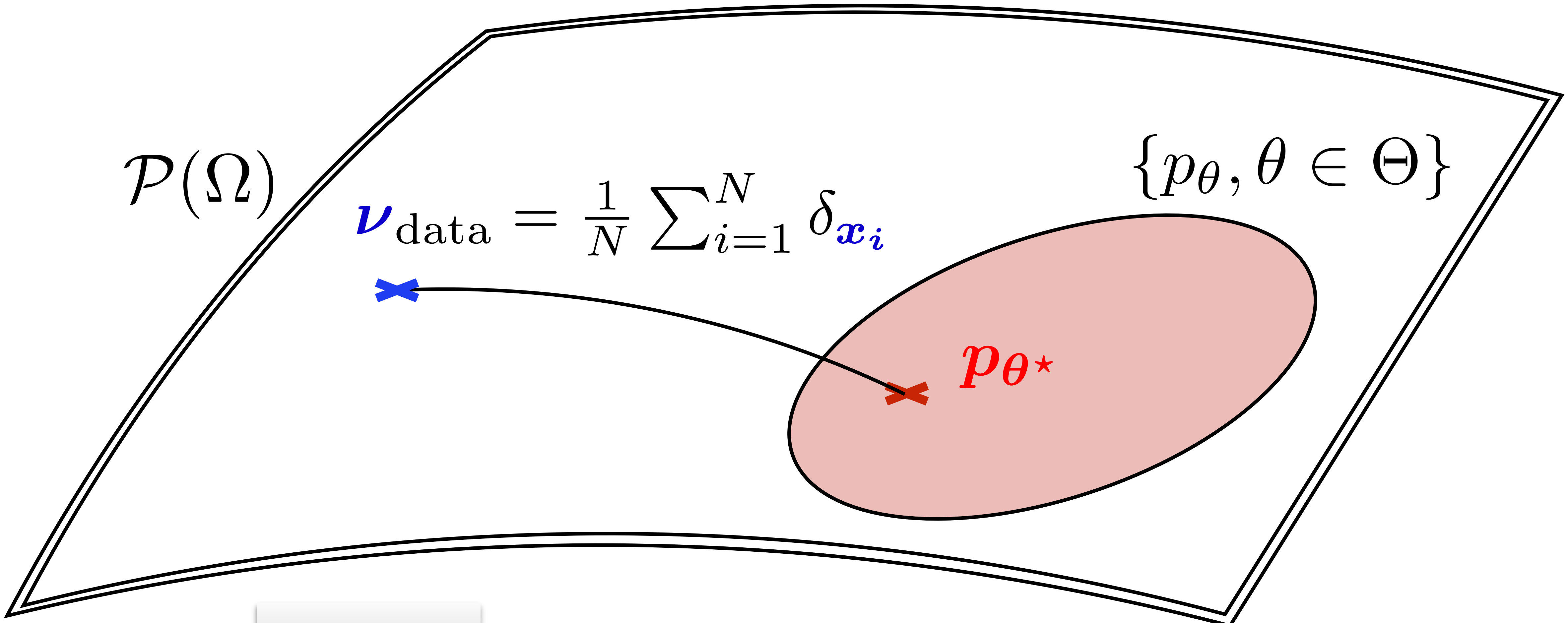
École nationale de la statistique et de l'administration économique

ENSAE ParisTech

université PARIS-SACLAY

Joint work / work in progress with
G. Peyré, A. Genevay *(ENS)*, F. Bach *(INRIA)*,
G. Montavon, K-R Müller *(TU Berlin)*

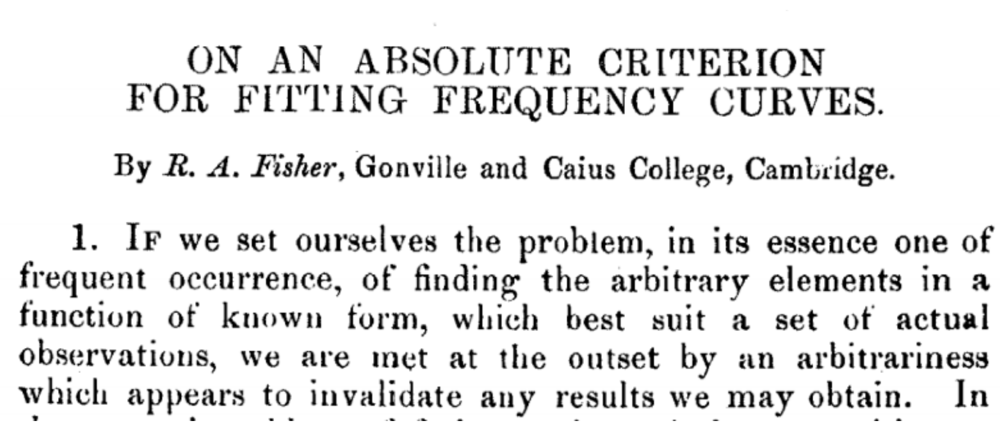
Maximum Likelihood Estimation



MLE

$$\min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\boldsymbol{\nu}_{\text{data}} \| \boldsymbol{p}_{\boldsymbol{\theta}})$$

$$\min_{\boldsymbol{\theta}\in\Theta} -\frac{1}{N}\sum_{i=1}^{N} \log \boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

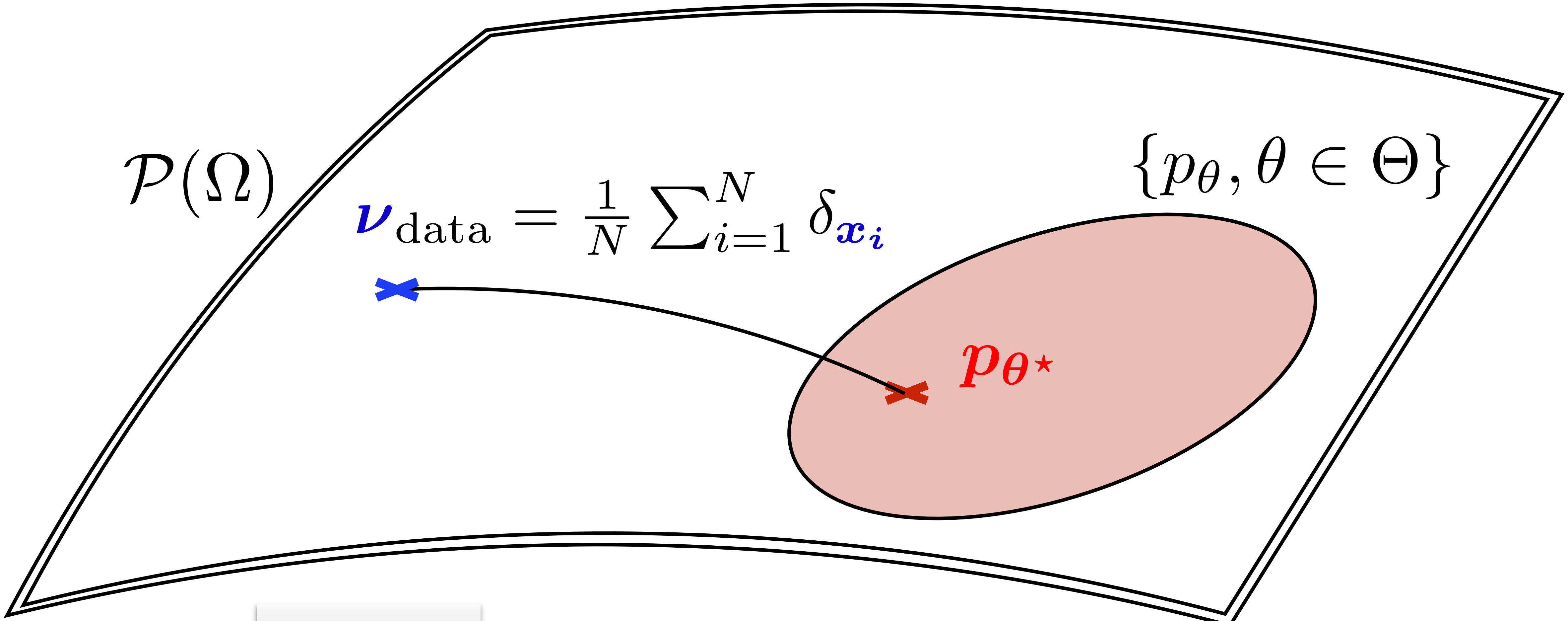


ON AN ABSOLUTE CRITERION
FOR FITTING FREQUENCY CURVES.

By *R. A. Fisher*, Gonville and Caius College, Cambridge.

1. If we set ourselves the problem, in its essence one of frequent occurrence, of finding the arbitrary elements in a function of known form, which best suit a set of actual observations, we are met at the outset by an arbitrariness which appears to invalidate any results we may obtain. In

Maximum Likelihood Estimation



MLE

$$\min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\boldsymbol{\nu}_{\text{data}} \| \boldsymbol{p_\theta})$$

$$\min_{\boldsymbol{\theta}\in\Theta} -\frac{1}{N}\sum_{i=1}^{N} \log \boldsymbol{p_\theta}(\boldsymbol{x_i})$$

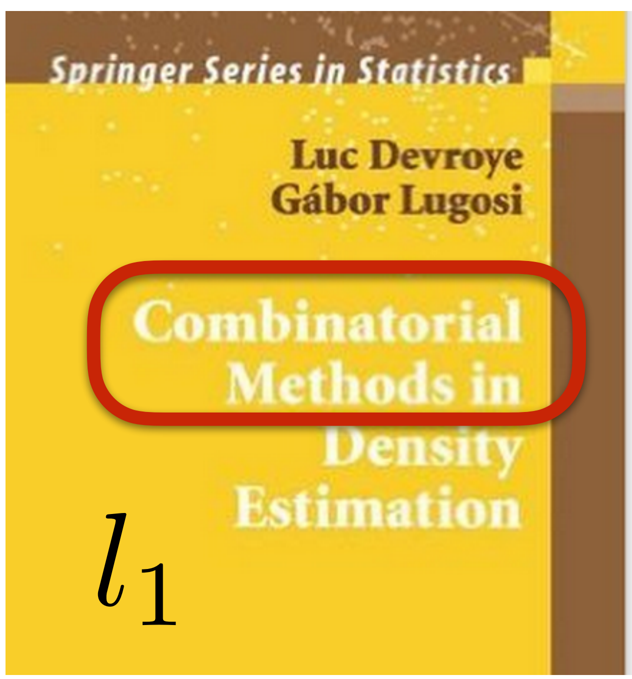
Minimum * Estimation

The Annals of Statistics
1980, Vol. 8, No. 3, 457–487

MINIMUM CHI-SQUARE, NOT MAXIMUM LIKELIHOOD!

BY JOSEPH BERKSON
Mayo Clinic, Rochester, Minnesota

ELSEVIER Computational Statistics & Data Analysis 29 (1998) 81–103
COMPUTATIONAL STATISTICS & DATA ANALYSIS



Minimum Hellinger distance estimation for Poisson mixtures

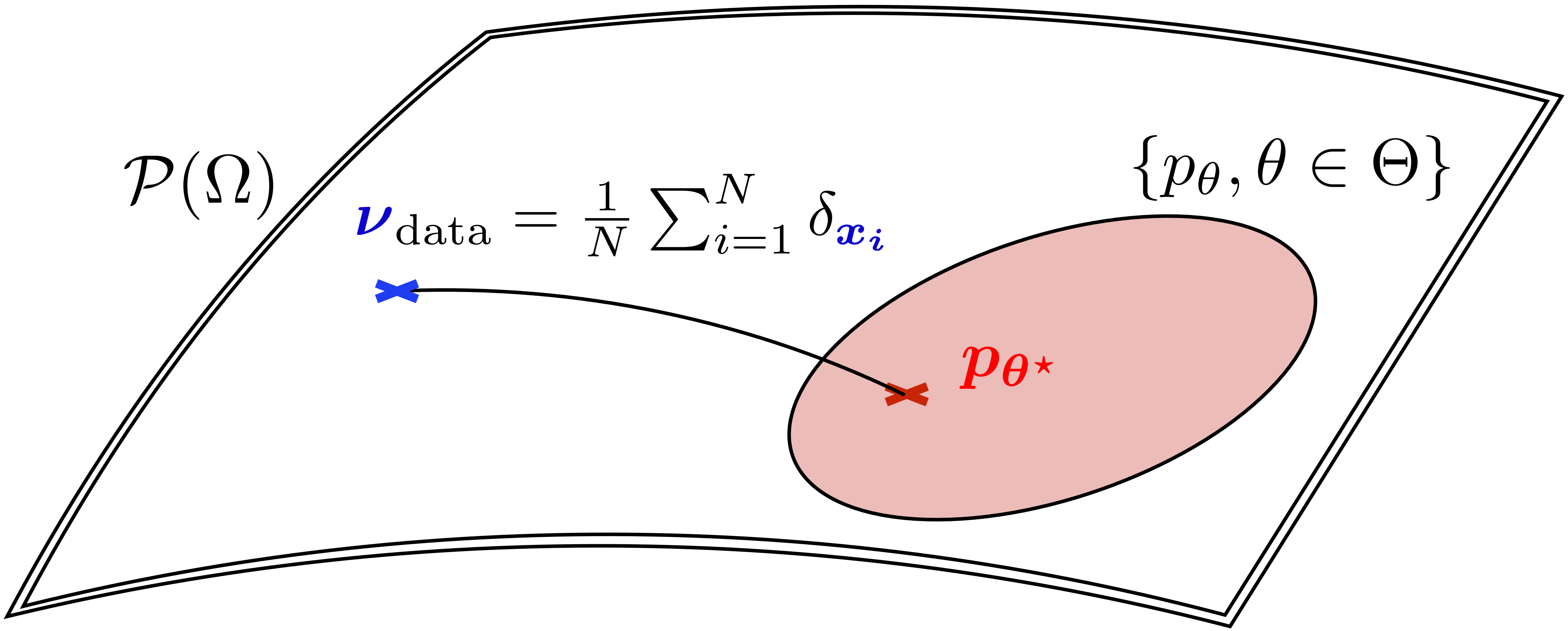
Dimitris Karlis, Evdokia Xekalaki*
Department of Statistics, Athens University of Economics and Business, 76 Patission Str., 104 34 Athens, Greece

ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT
Statistics & Probability Letters 76 (2006) 1298–1302
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

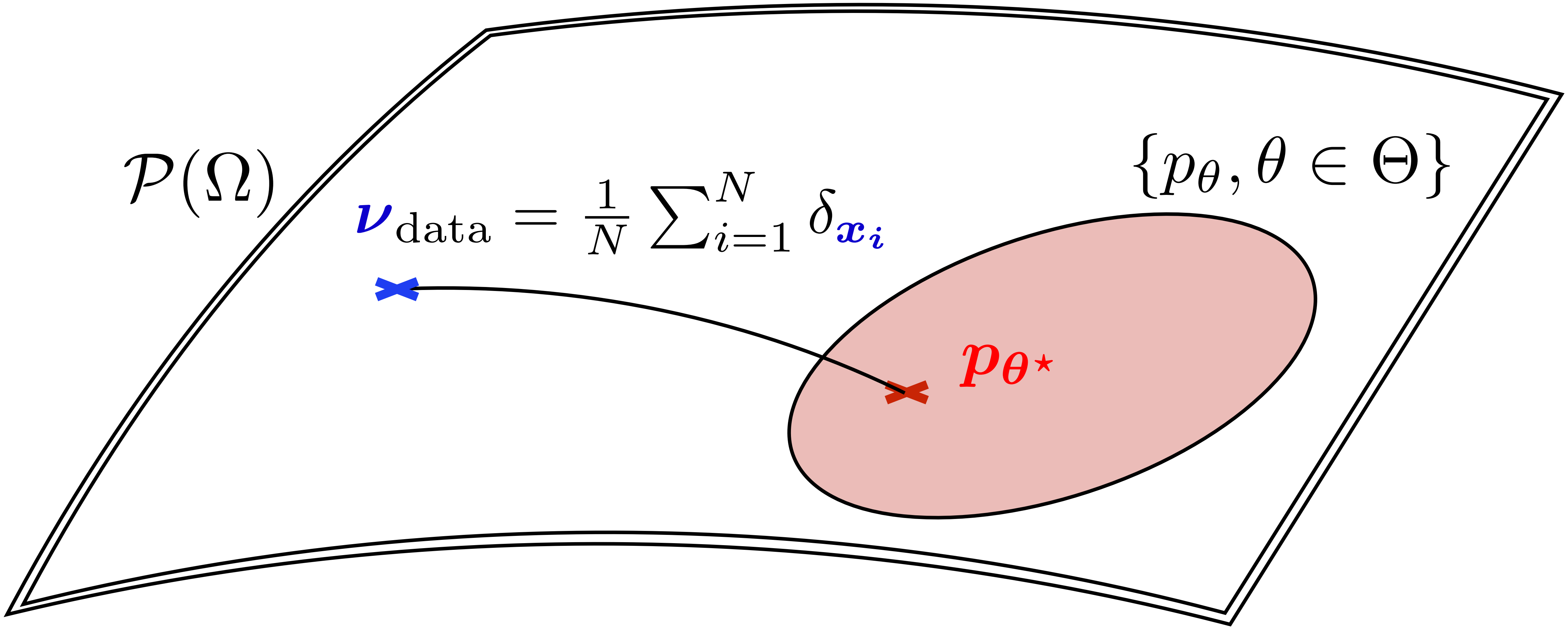
On minimum Kantorovich distance estimators

Federico Bassetti[a], Antonella Bodini[b], Eugenio Regazzini[a,*]

Statistical Estimation



Statistical Estimation



$$\min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\boldsymbol{\nu}_{\text{data}} \| p_{\boldsymbol{\theta}} \|)$$

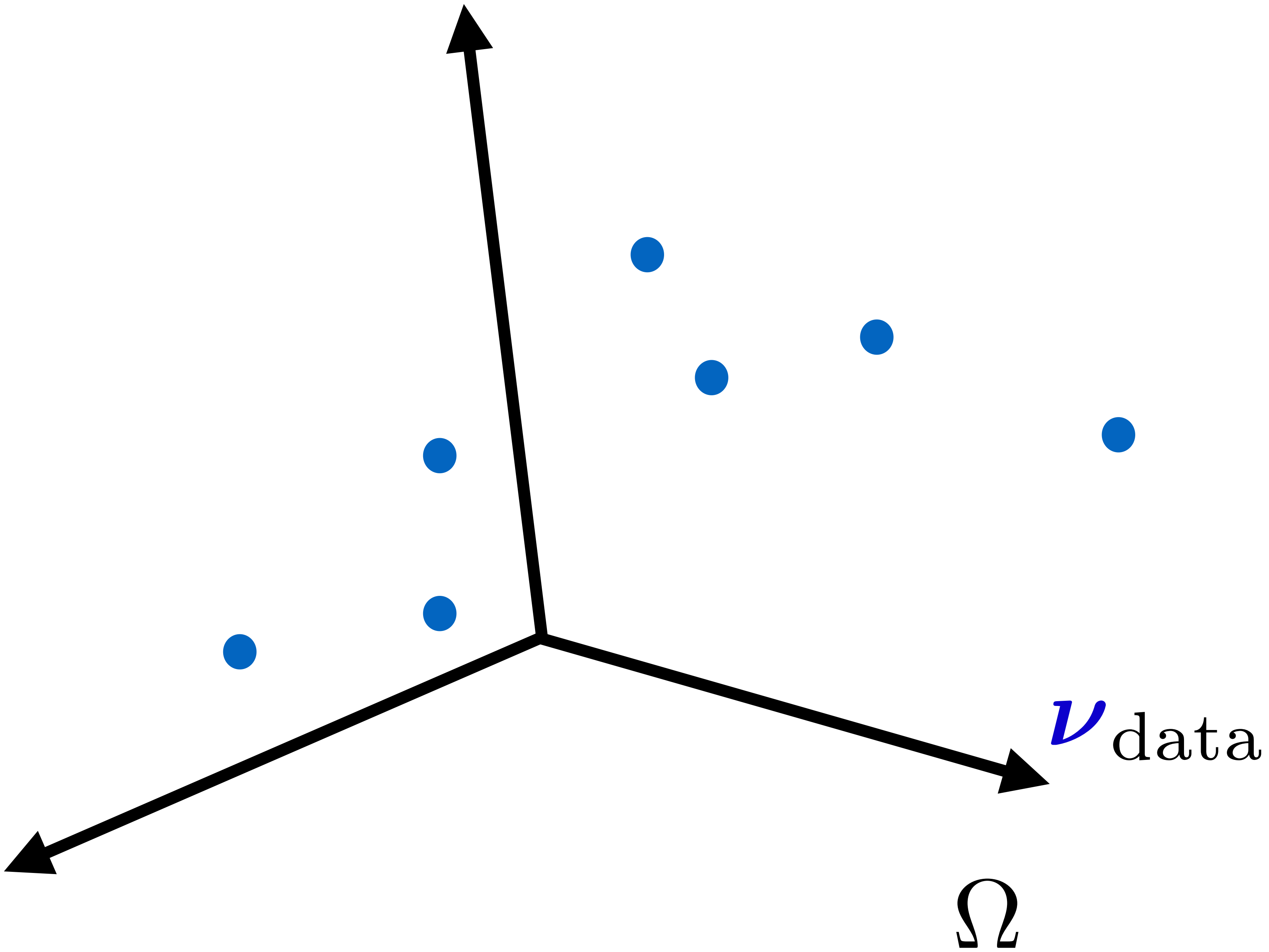
MLE

$$\min_{\boldsymbol{\theta}\in\Theta} W(\boldsymbol{\nu}_{\text{data}}, p_{\boldsymbol{\theta}})$$

MKE

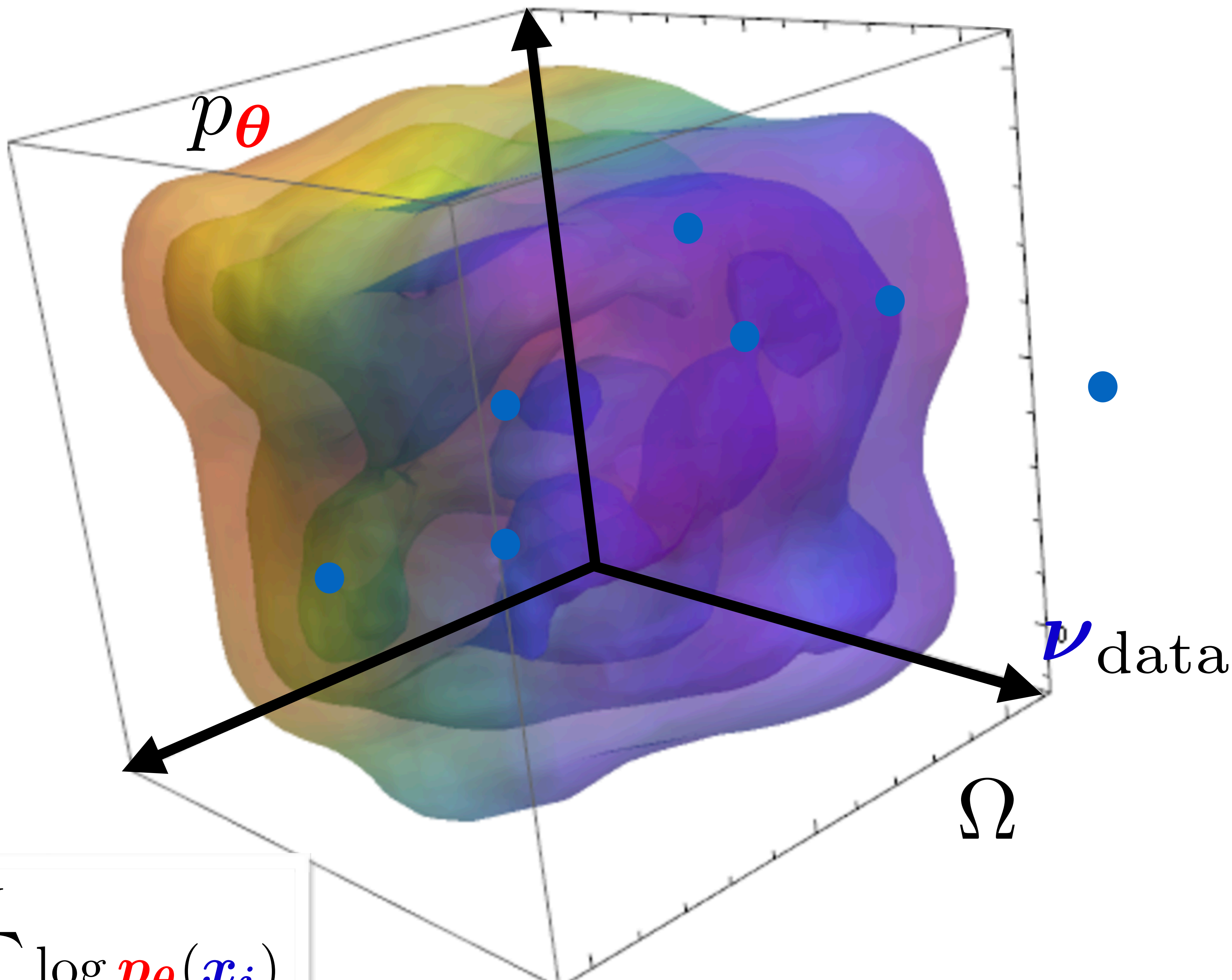
[Bassetti'06]

Model = positive densities



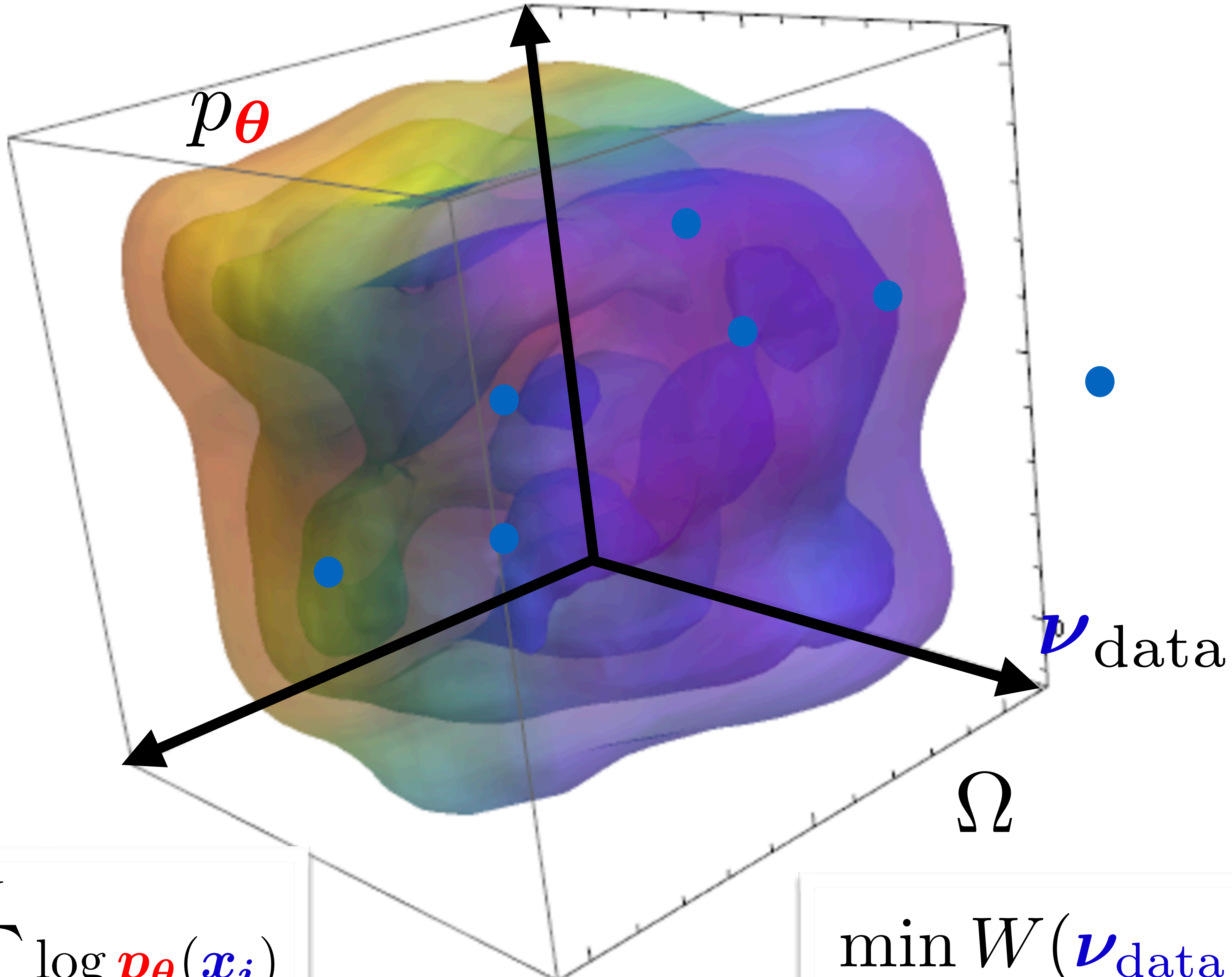
$$\min_{\theta \in \Theta} - \sum_{i=1}^{N} \log p_{\theta}(x_i)$$

Model = positive densities



$$\min_{\boldsymbol{\theta}\in\Theta} -\sum_{i=1}^{N} \log \boldsymbol{p_{\theta}}(\boldsymbol{x_i})$$

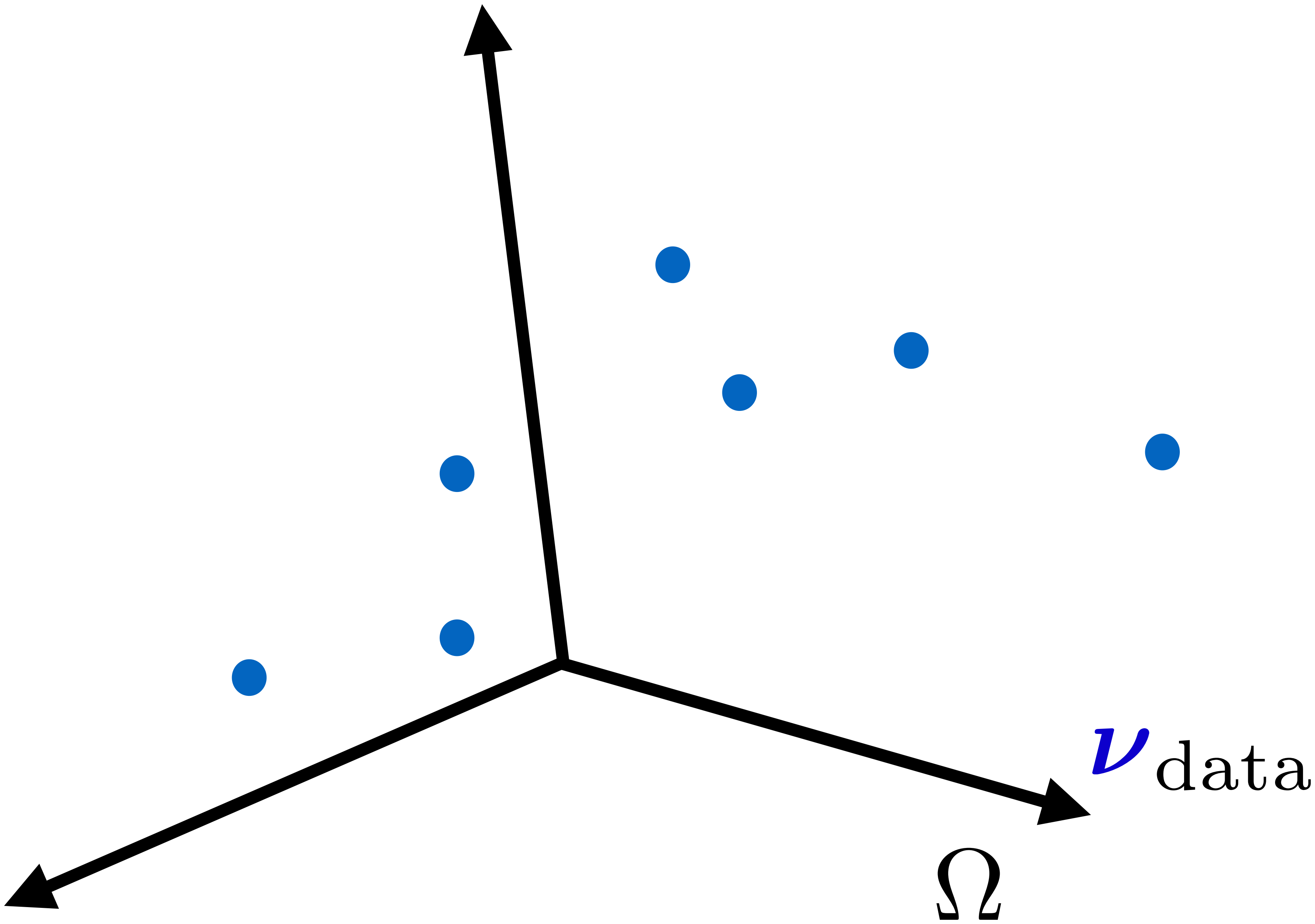
Model = positive densities



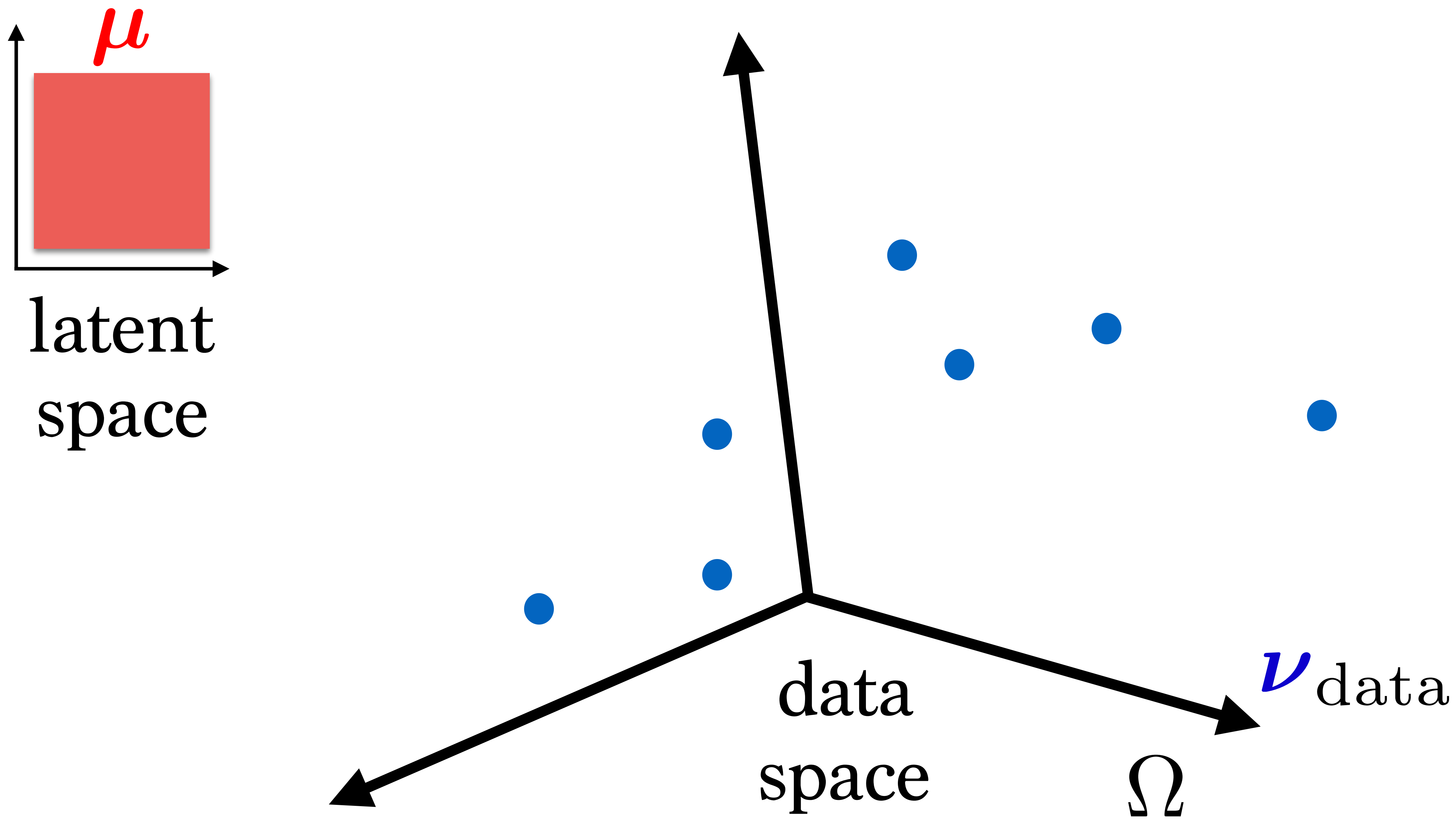
$$\min_{\theta \in \Theta} - \sum_{i=1}^{N} \log p_\theta(x_i)$$

$$\min_{\theta \in \Theta} W(\nu_{\text{data}}, p_\theta)$$

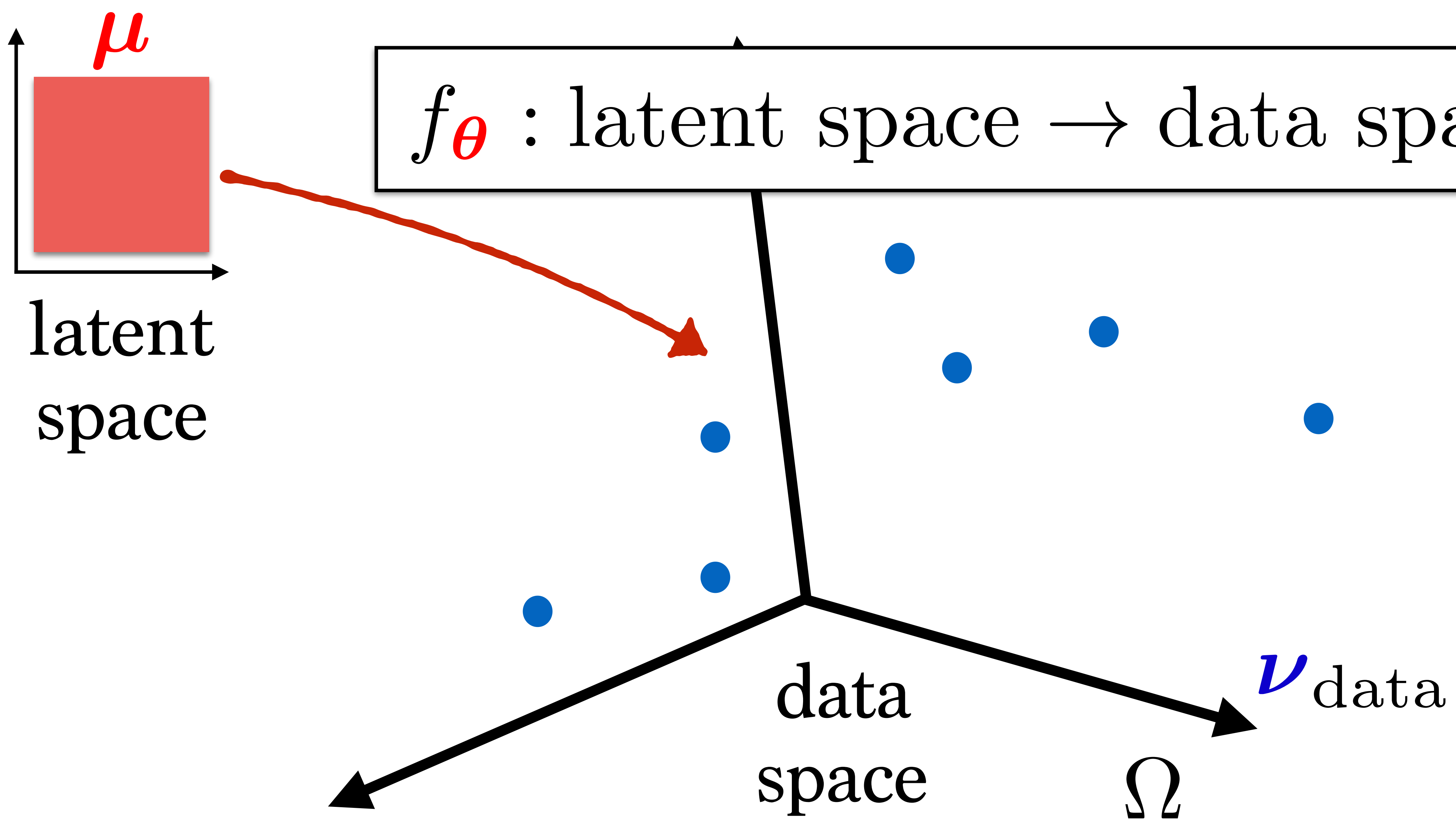
Model = generative



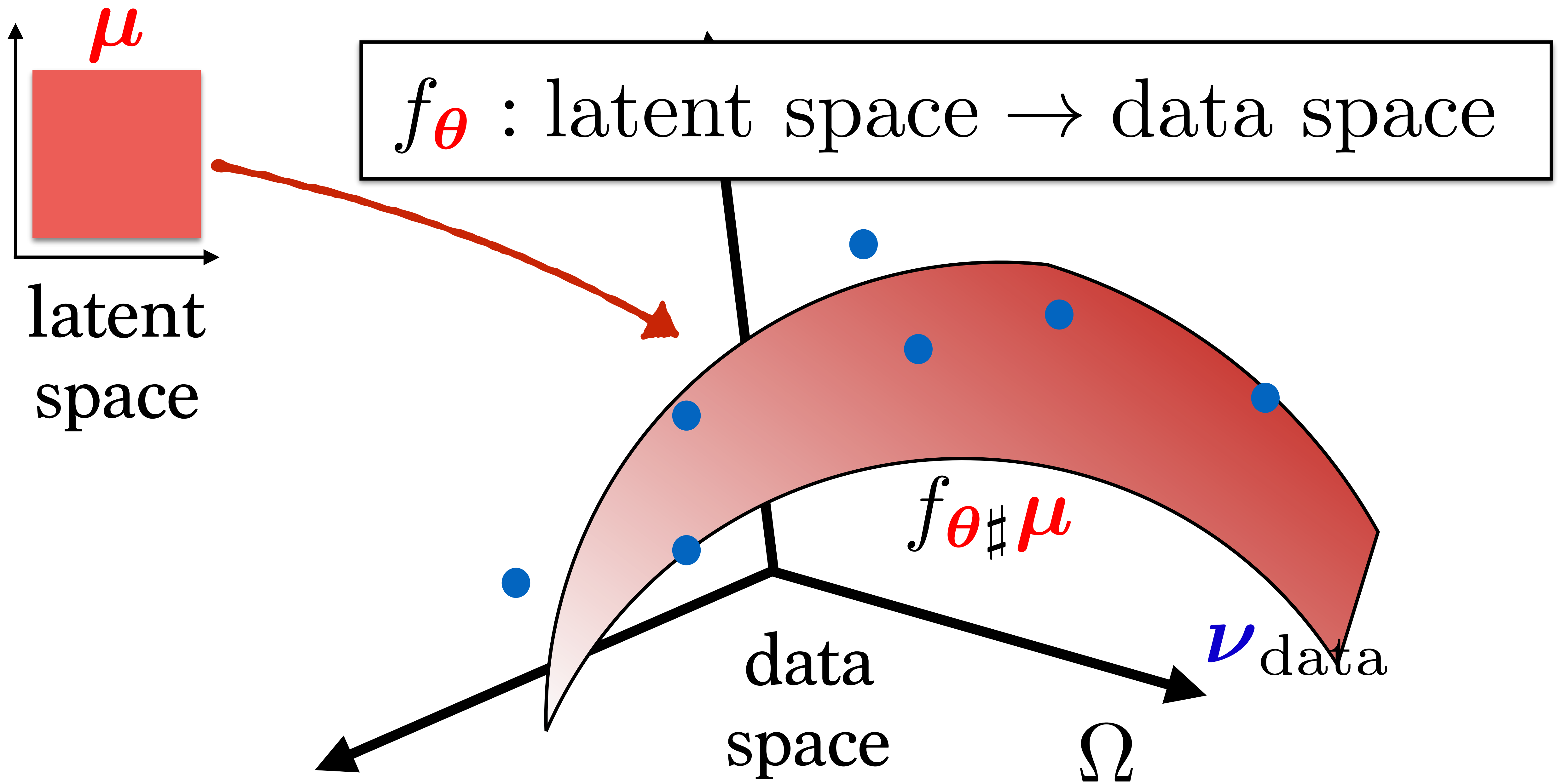
Model = generative



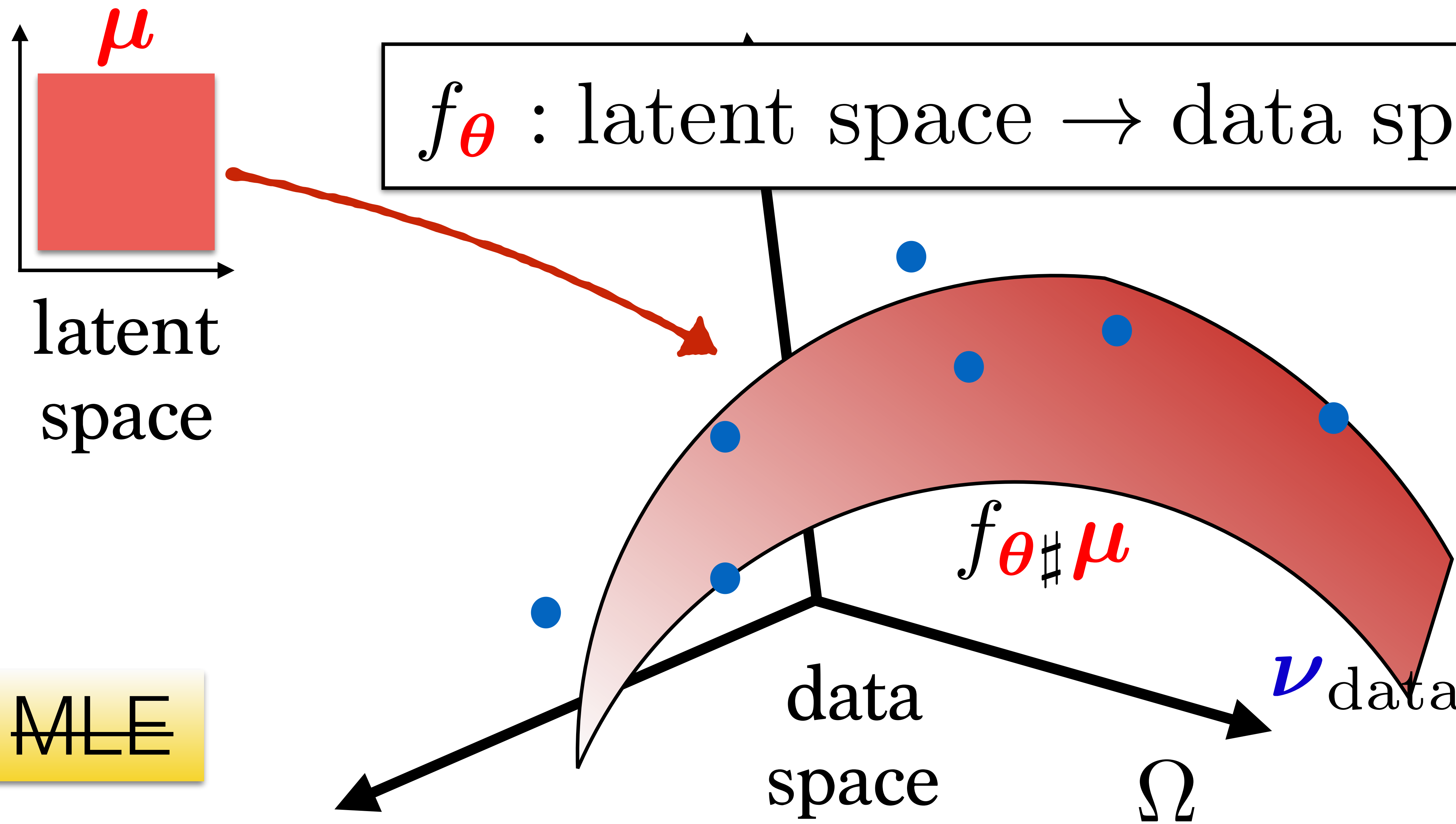
Model = generative



Model = generative



Model = generative



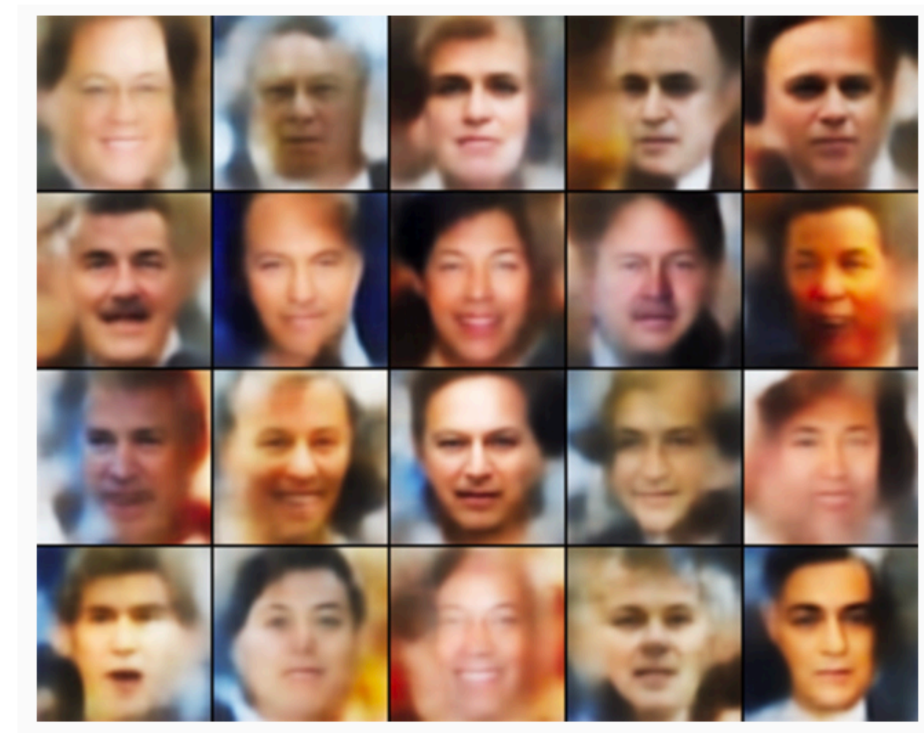
~~MLE~~

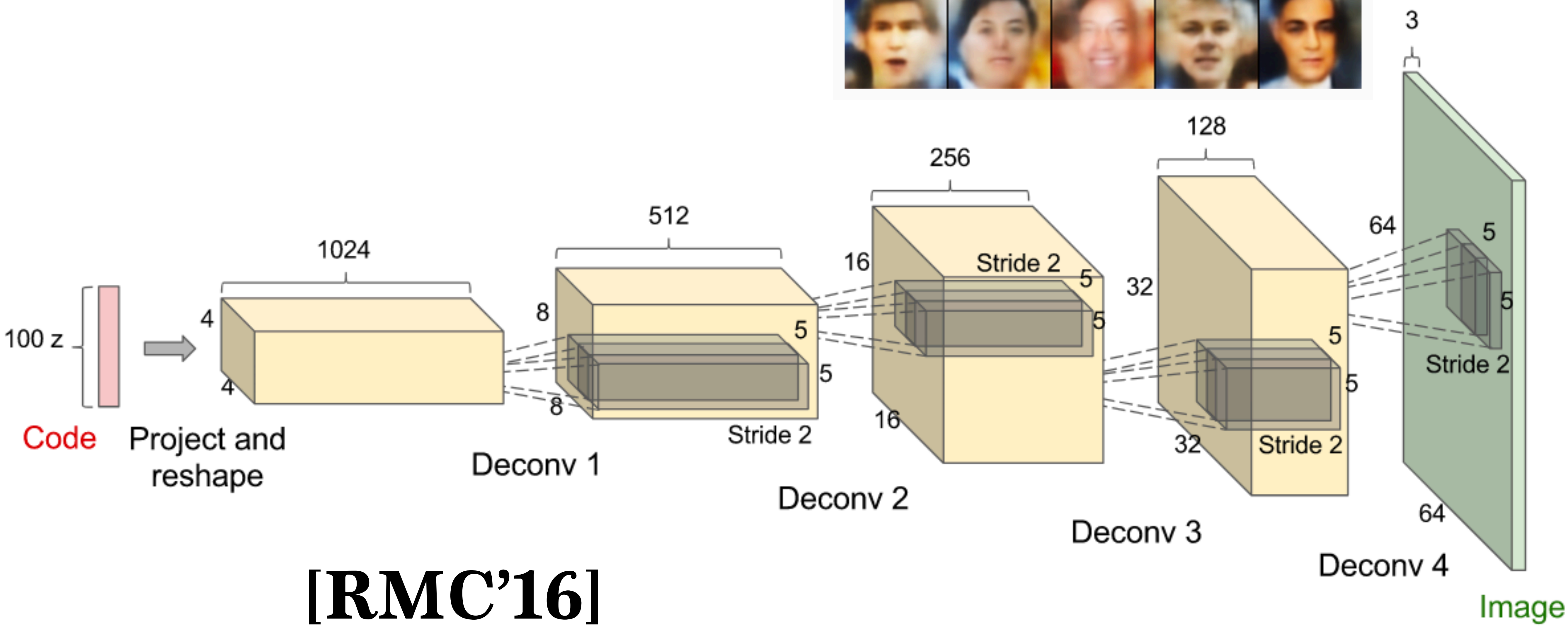
$$\min_{\theta \in \Theta} W(\nu_{\text{data}}, f_{\theta\sharp}\mu)$$

GM-MKE
W-GAN

Illustration - GAN

latent

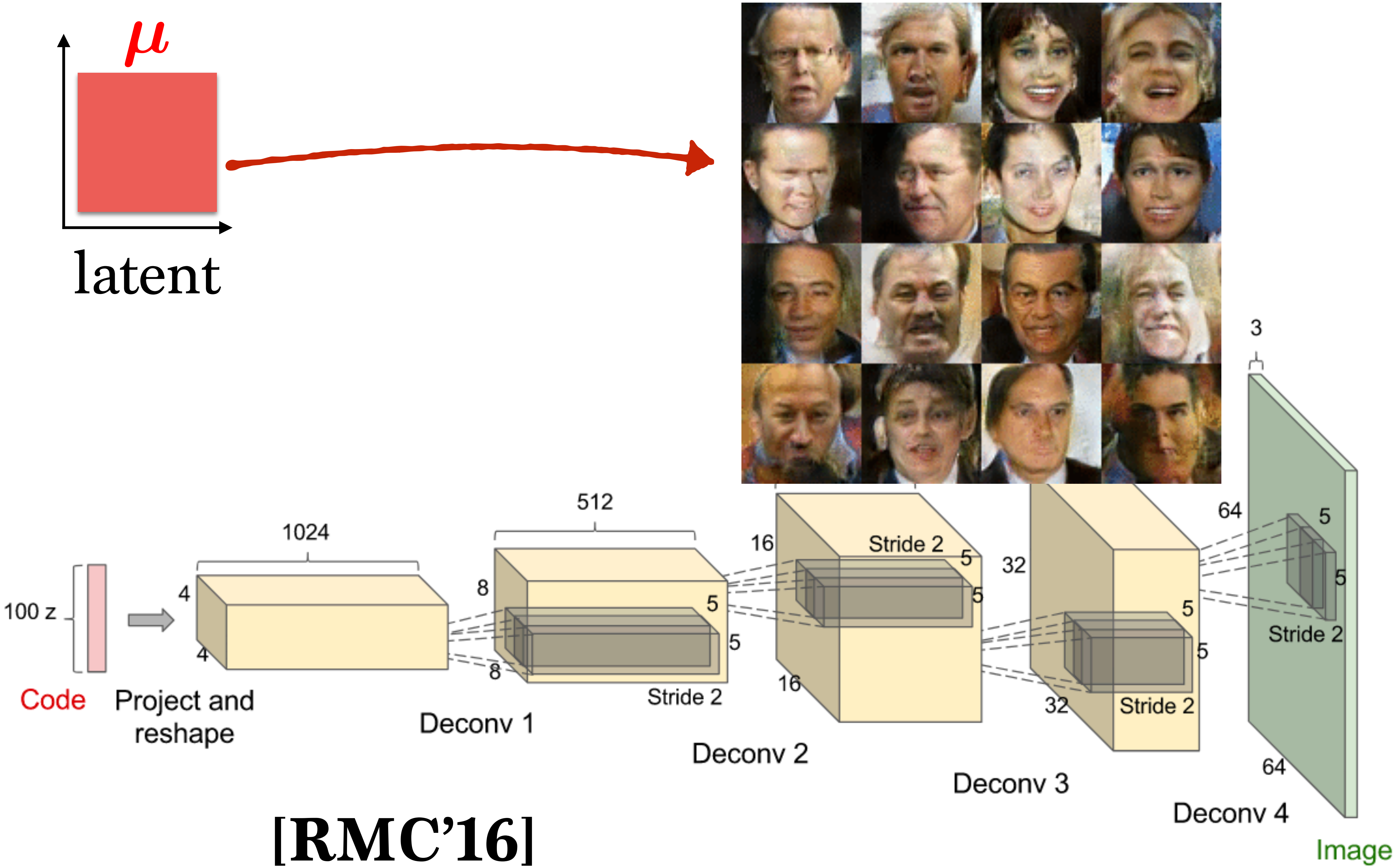




[RMC'16]

http://torch.ch/blog/2015/11/13/gan.html

Illustration - GAN



[RMC'16]

http://torch.ch/blog/2015/11/13/gan.html

Wasserstein Distances

Def. For $p \geq 1$, the p-Wasserstein distance between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$, defined by a metric $\boldsymbol{D}$ on Ω,

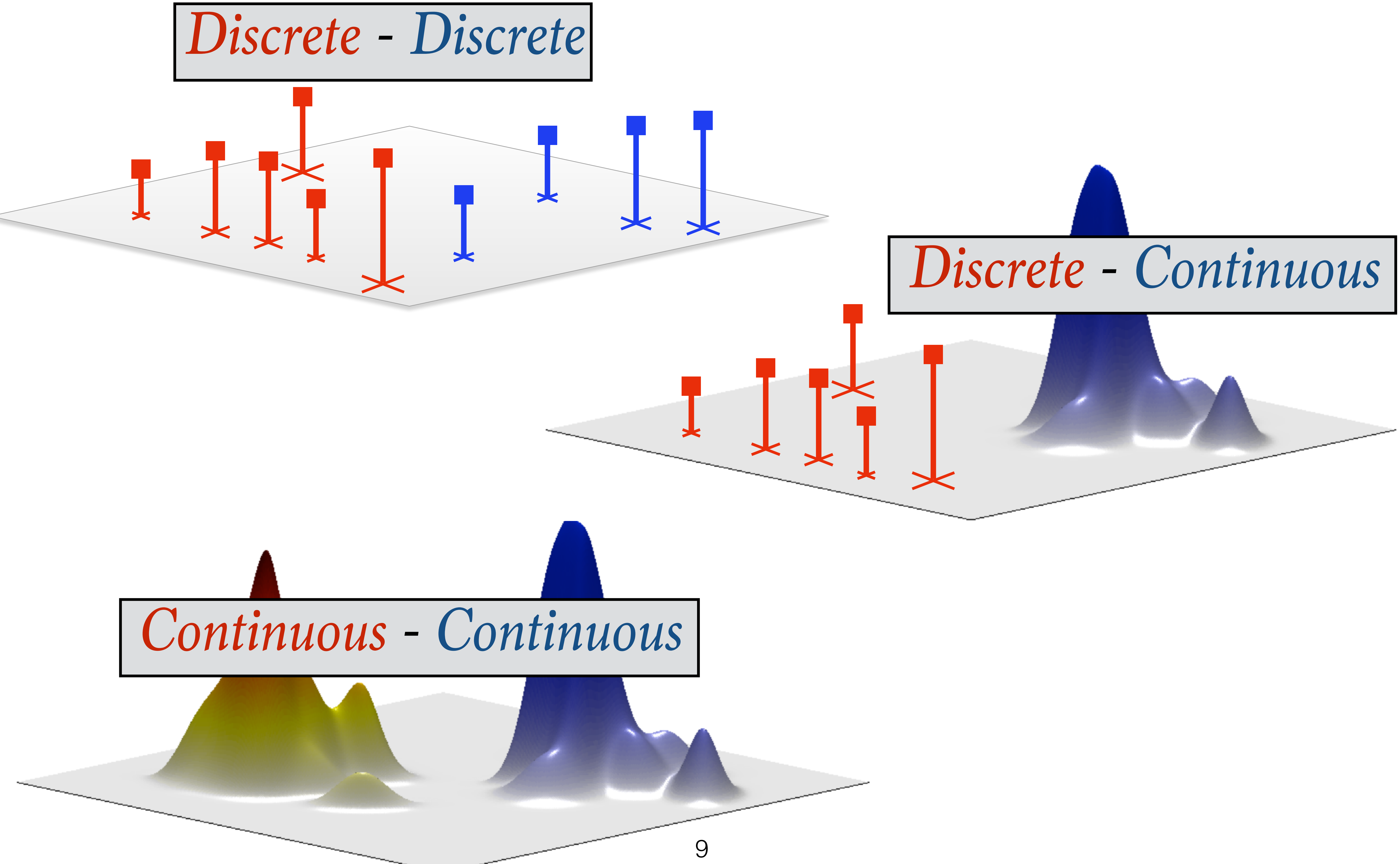
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \int\!\!\int \boldsymbol{D}(\boldsymbol{X}, \boldsymbol{Y})^p d\boldsymbol{P}(X, Y).$$

PRIMAL

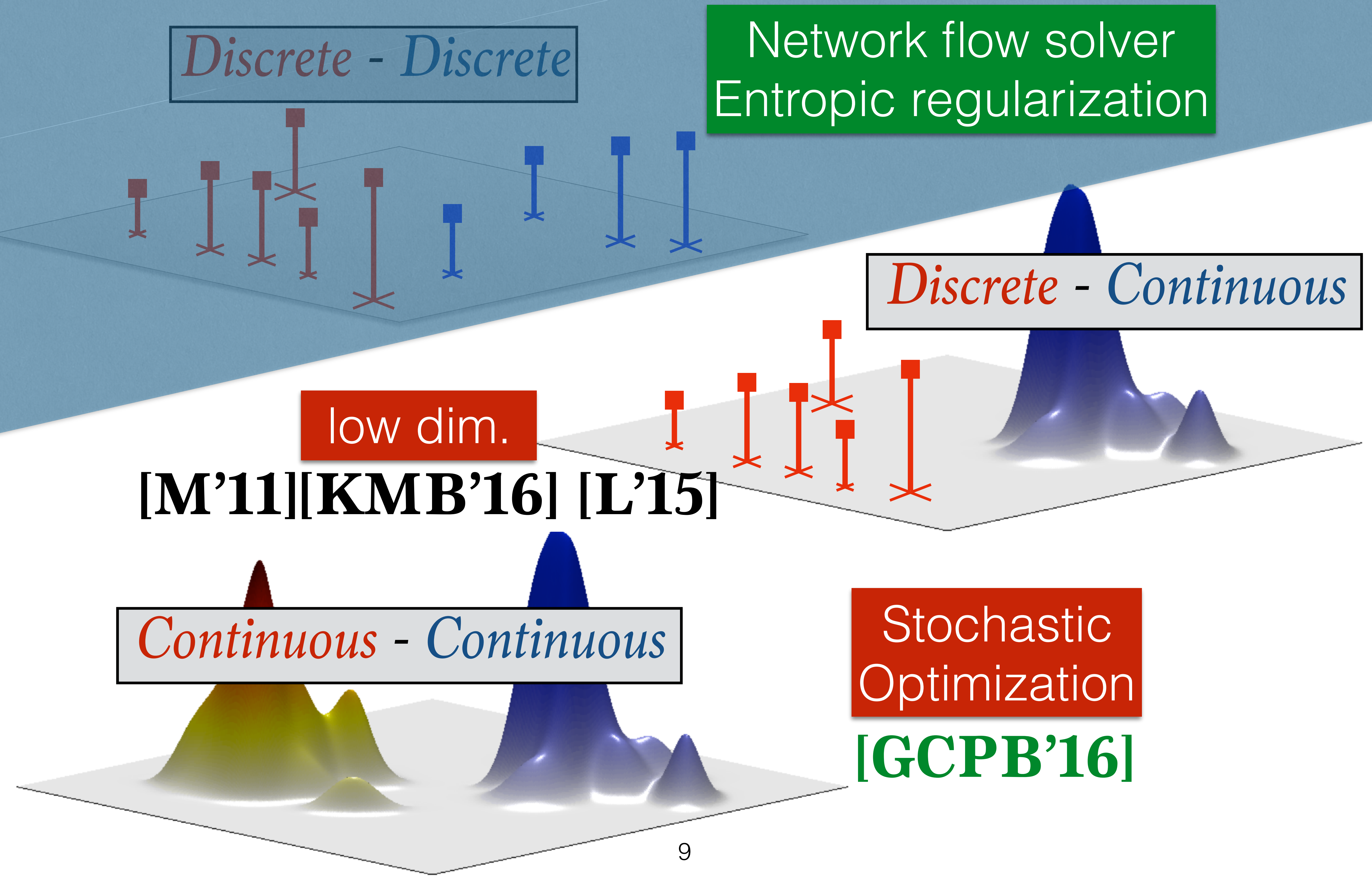
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{D}^p(x, y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

W is versatile



W is versatile

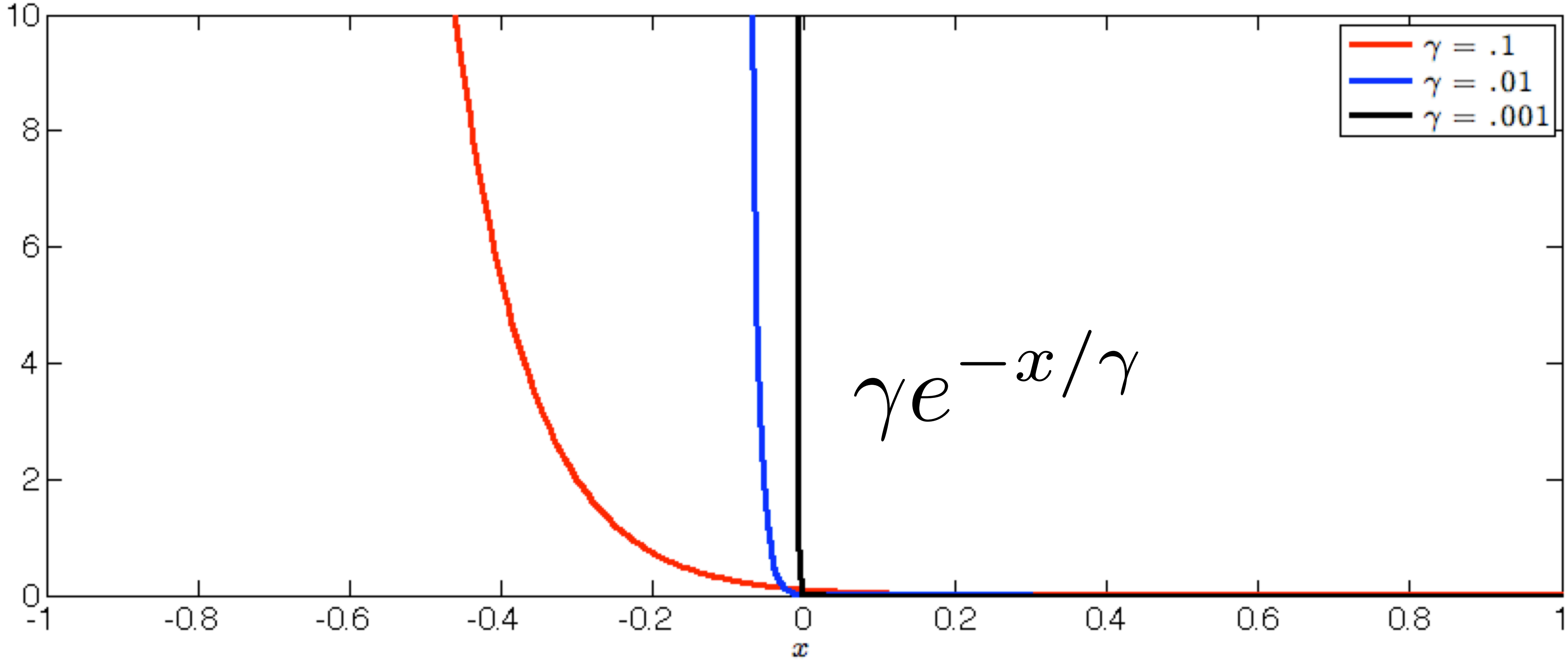


Dual regularization

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi}, \boldsymbol{\psi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iota_C(\boldsymbol{\varphi}, \boldsymbol{\psi})$$

$$C = \{(\boldsymbol{\varphi}, \boldsymbol{\psi}) | \boldsymbol{\varphi} \oplus \boldsymbol{\psi} \leq \boldsymbol{D}^p\}$$

DUAL

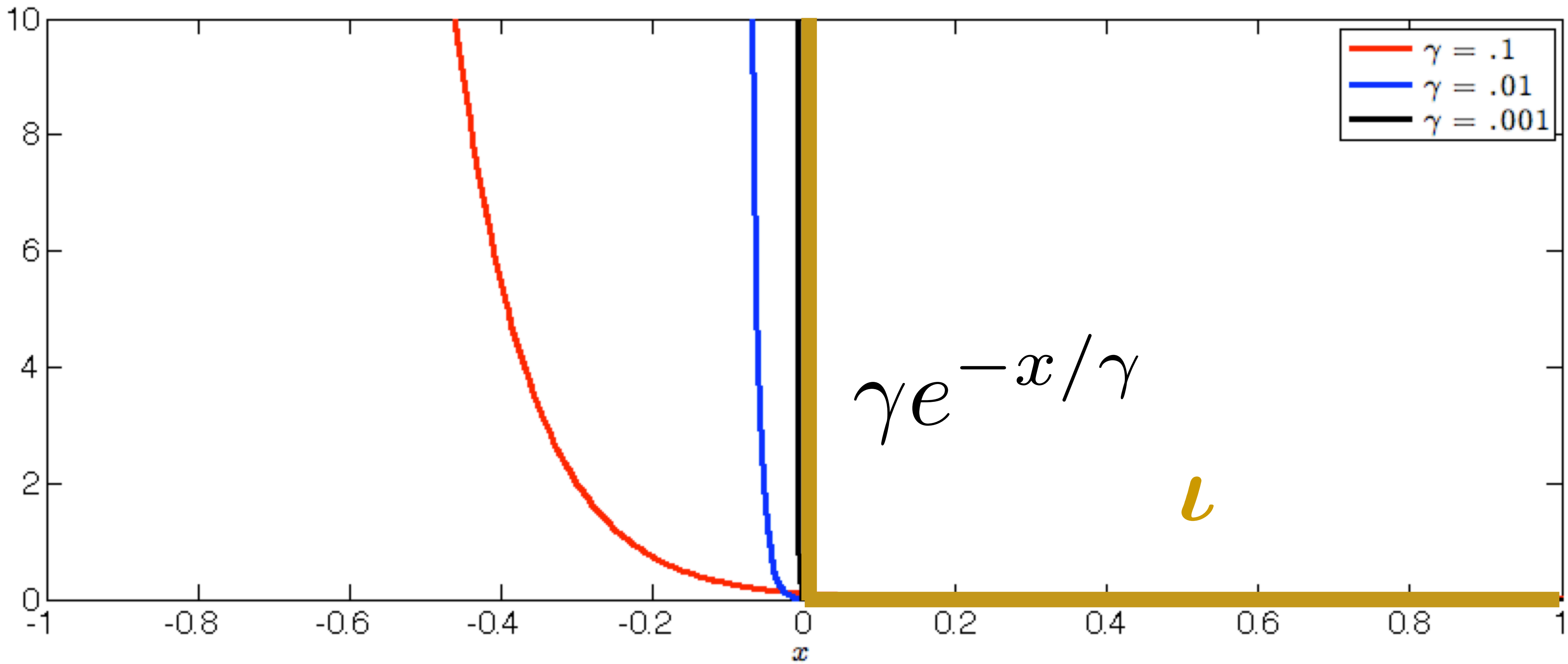


Dual regularization

$$W_p^p(\mu, \nu) = \sup_{\varphi, \psi} \int \varphi d\mu + \int \psi d\nu - \iota_C(\varphi, \psi)$$

$$C = \{(\varphi, \psi) | \varphi \oplus \psi \leq D^p\}$$

DUAL



Dual regularization

$$W_p^p(\mu,\nu) = \sup_{\varphi,\psi} \int \varphi \, d\mu + \int \psi \, d\nu - \iota_C(\varphi,\psi)$$

$$C = \{(\varphi,\psi) | \varphi \oplus \psi \le D^p\}$$

DUAL

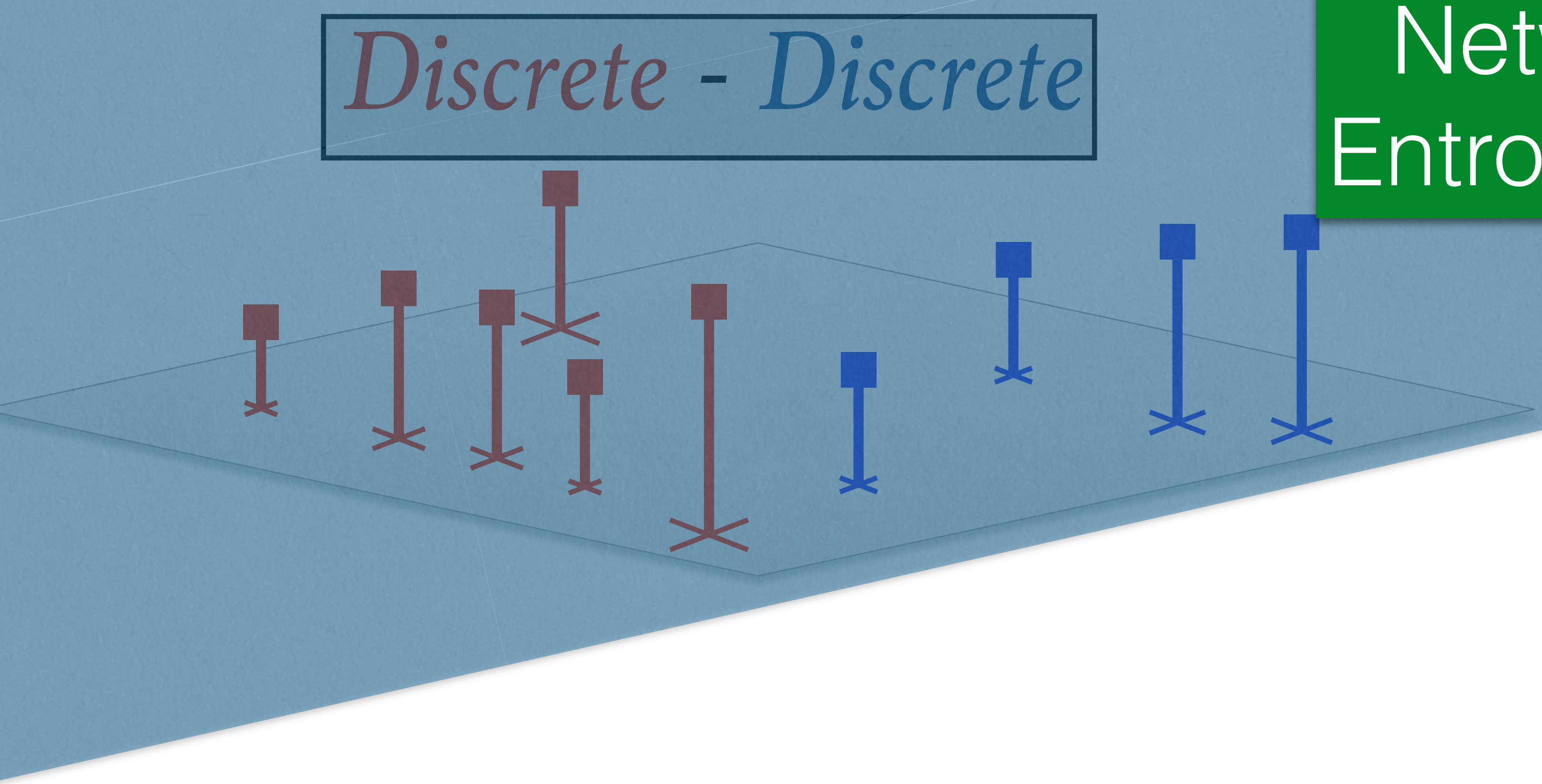
regularizing dual ↓ *constraints* $\gamma > 0$

$$W_\gamma(\mu,\nu) = \sup_{\varphi,\psi} \int \varphi \, d\mu + \int \psi \, d\nu - \iota_C^\gamma(\varphi,\psi)$$

$$\iota_C^\gamma(\varphi,\psi) = \gamma \iint e^{(\varphi \oplus \psi - D^p)/\gamma} \, d\mu \, d\nu$$

REGULARIZED DUAL

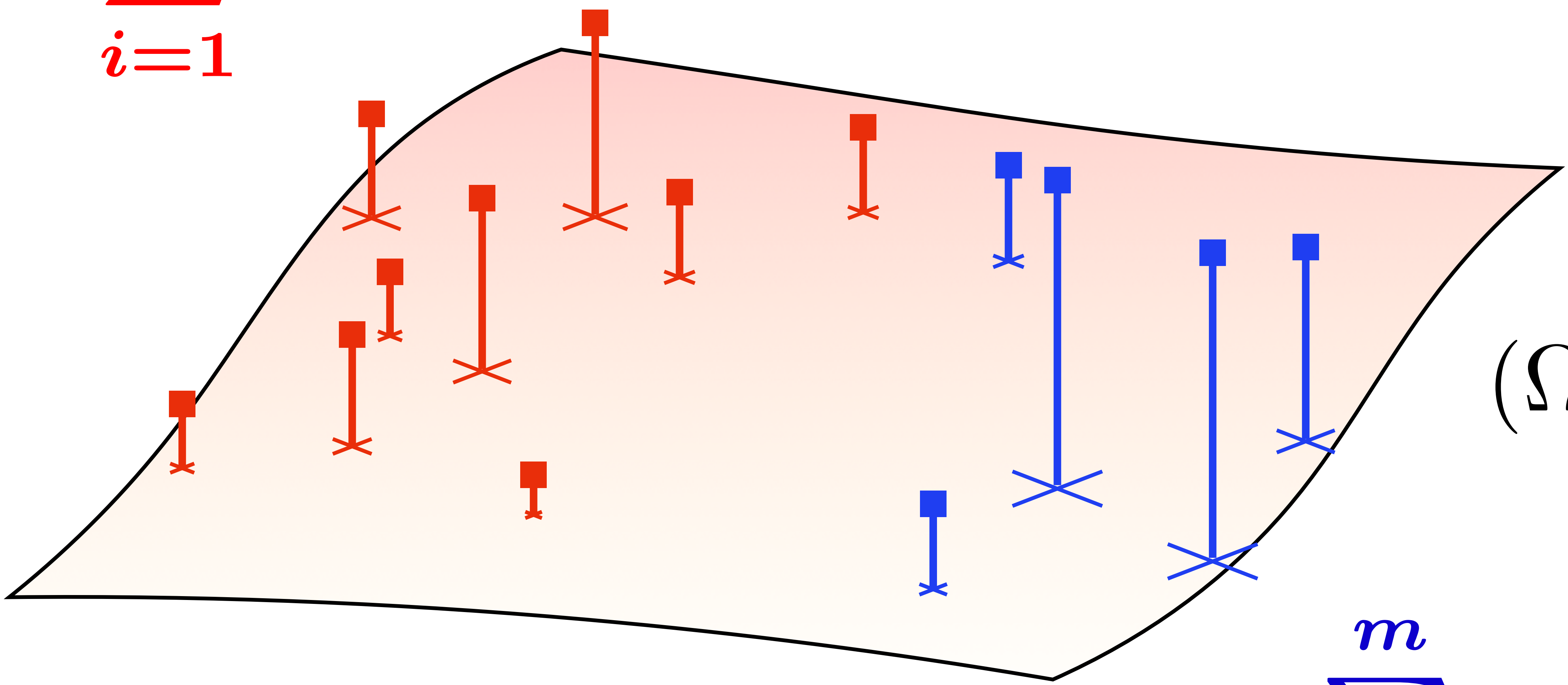
W is versatile



Network flow solver
Entropic regularization

OT on Two Empirical Measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$

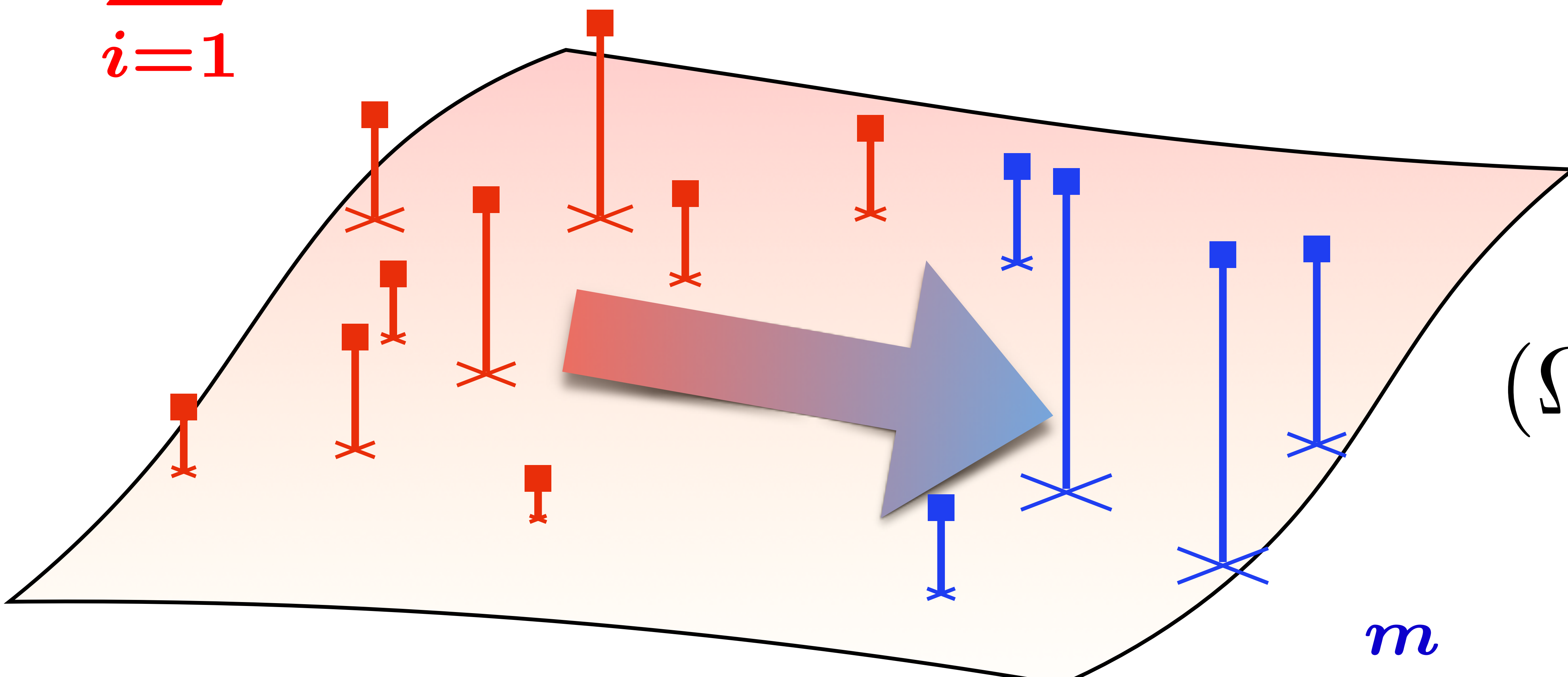


(Ω, D)

$$\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

OT on Two Empirical Measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



(Ω, D)

$$\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Dual regularization, *Discrete*

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iota_C^\gamma(\boldsymbol{\varphi},\boldsymbol{\psi})$$

$$\iota_C^\gamma(\boldsymbol{\varphi},\boldsymbol{\psi}) = \gamma \iint e^{(\boldsymbol{\varphi}\oplus\boldsymbol{\psi}-\boldsymbol{D}^p)/\gamma} d\boldsymbol{\mu} d\boldsymbol{\nu}$$

REGULARIZED DUAL

$$\boldsymbol{\mu} = \sum_{i=1}^n \boldsymbol{a_i} \delta_{\boldsymbol{x_i}}$$

$$\boldsymbol{\nu} = \sum_{j=1}^m \boldsymbol{b_j} \delta_{\boldsymbol{y_j}}$$

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma \sum_{\boldsymbol{ij}} \boldsymbol{a_i} \boldsymbol{b_j}\, e^{\frac{\boldsymbol{\alpha_i}+\boldsymbol{\beta_j}-\boldsymbol{D}^p(\boldsymbol{x_i},\boldsymbol{y_j})}{\gamma}}$$

REGULARIZED DISCRETE DUAL

Dual regularization, *Discrete*

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iota_C^\gamma(\boldsymbol{\varphi},\boldsymbol{\psi})$$

$$\iota_C^\gamma(\boldsymbol{\varphi},\boldsymbol{\psi}) = \gamma \iint e^{(\boldsymbol{\varphi}\oplus\boldsymbol{\psi}-\boldsymbol{D}^p)/\gamma} d\boldsymbol{\mu} d\boldsymbol{\nu}$$

REGULARIZED DUAL

$\boldsymbol{\mu} = \sum_{i=1}^n \boldsymbol{a_i}\delta_{\boldsymbol{x_i}}$

$\boldsymbol{\nu} = \sum_{j=1}^m \boldsymbol{b_j}\delta_{\boldsymbol{y_j}}$

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b} - \gamma(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})^T \; K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

$$\text{where} \quad K = \left[e^{-\frac{\boldsymbol{D}^p(\boldsymbol{x_i},\boldsymbol{y_j})}{\gamma}} \right]_{ij}$$

REGULARIZED DISCRETE DUAL

Algorithm: Block Coordinate Ascent

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma(\boldsymbol{a} \odot e^{\boldsymbol{\alpha}/\gamma})^T \; K(\boldsymbol{b} \odot e^{\boldsymbol{\beta}/\gamma})$$

REGULARIZED DISCRETE DUAL

$$\mathcal{E}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma(\boldsymbol{a} \odot e^{\boldsymbol{\alpha}/\gamma})^T \; K(\boldsymbol{b} \odot e^{\boldsymbol{\beta}/\gamma})$$

$$\nabla_{\boldsymbol{\alpha}} \mathcal{E} = \boldsymbol{a} - \boldsymbol{a} \odot e^{\boldsymbol{\alpha}/\gamma} \odot \; K(\boldsymbol{b} \odot e^{\boldsymbol{\beta}/\gamma})$$

$$\nabla_{\boldsymbol{\beta}} \mathcal{E} = \boldsymbol{b} - \boldsymbol{b} \odot e^{\boldsymbol{\beta}/\gamma} \odot \; K^T(\boldsymbol{a} \odot e^{\boldsymbol{\alpha}/\gamma})$$

Algorithm: Block Coordinate Ascent

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b} - \gamma(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})^T\ K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

REGULARIZED DISCRETE DUAL

$$\mathcal{E}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b} - \gamma(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})^T\ K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

$$\nabla_{\boldsymbol{\alpha}}\mathcal{E} = \boldsymbol{a} - \boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma}\odot\ K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

$$\boldsymbol{\alpha} \leftarrow -\gamma\log\ K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

$$\nabla_{\boldsymbol{\beta}}\mathcal{E} = \boldsymbol{b} - \boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma}\odot\ K^T(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})$$

$$\boldsymbol{\beta} \leftarrow -\gamma\log\ K^T(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})$$

Algorithm: Block Coordinate Ascent

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma (\boldsymbol{a} \odot e^{\boldsymbol{\alpha}/\gamma})^T \; K (\boldsymbol{b} \odot e^{\boldsymbol{\beta}/\gamma})$$

REGULARIZED DISCRETE DUAL

Algorithm: Block Coordinate Ascent

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b} - \gamma(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})^T \; K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

REGULARIZED DISCRETE DUAL

$$\boldsymbol{u} \leftarrow \frac{\boldsymbol{a}}{K\boldsymbol{v}}$$

$$\boldsymbol{v} \leftarrow \frac{\boldsymbol{b}}{K^T\boldsymbol{u}}$$

Algorithm: Block Coordinate Ascent

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b} - \gamma(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})^T\ K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

REGULARIZED DISCRETE DUAL

$$(\boldsymbol{u},\boldsymbol{v}) \stackrel{\text{def}}{=} (\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma}, \boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

$$\boldsymbol{\alpha} \leftarrow -\gamma \log\ K(\boldsymbol{b}\odot e^{\boldsymbol{\beta}/\gamma})$$

$$\boldsymbol{u} \leftarrow \frac{\boldsymbol{a}}{K\boldsymbol{v}}$$

$$\boldsymbol{\beta} \leftarrow -\gamma \log\ K^T(\boldsymbol{a}\odot e^{\boldsymbol{\alpha}/\gamma})$$

$$\boldsymbol{v} \leftarrow \frac{\boldsymbol{b}}{K^T\boldsymbol{u}}$$

Entropic Regularization [Wilson'62]

Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle + \gamma \mathrm{KL}\left(\boldsymbol{P} \| \boldsymbol{a}\boldsymbol{b}^T\right)$$

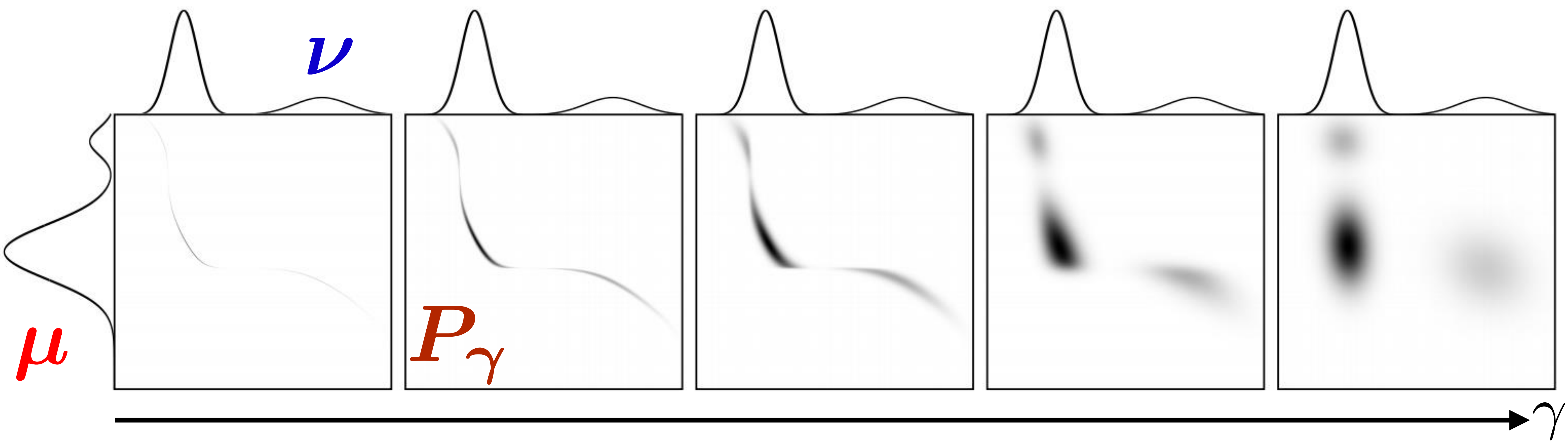
$$\mathrm{KL}\left(\boldsymbol{P} \| \boldsymbol{a}\boldsymbol{b}^T\right) = E(\boldsymbol{a}) + E(\boldsymbol{b}) - E(\boldsymbol{P})$$

Note: Unique optimal solution because of strong concavity of Entropy

Entropic Regularization [**Wilson'62**]

Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle + \gamma \mathrm{KL}\left(\boldsymbol{P} || \boldsymbol{a}\boldsymbol{b}^T\right)$$



Note: Unique optimal solution because of strong concavity of Entropy

Fast & Scalable Algorithm

Prop. If $P_\gamma \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \stackrel{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

Fast & Scalable Algorithm

Prop. If $P_\gamma \stackrel{\text{def}}{=} \underset{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})}{\text{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \textbf{diag}(\boldsymbol{u}) K \textbf{diag}(\boldsymbol{v}), \quad K \stackrel{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

$$\mathcal{L}(P,\alpha,\beta) = \sum_{ij} P_{ij}M_{ij} + \gamma P_{ij}(\log P_{ij} - \log(a_i b_j) - 1) - \alpha^T(P\mathbf{1} - \boldsymbol{a}) - \beta^T(P^T\mathbf{1} - \boldsymbol{b})$$

$$\partial L/\partial P_{ij} = M_{ij} + \gamma(\log P_{ij} - \log a_i - \log b_j) - \alpha_i - \beta_j$$

$$(\boldsymbol{\partial L/\partial P_{ij}} = 0) \Rightarrow P_{ij} = a_i e^{\frac{\alpha_i}{\gamma}} e^{-\frac{M_{ij}}{\gamma}} b_j e^{\frac{\beta_j}{\gamma}} = \boldsymbol{u_i} K_{ij} \boldsymbol{v_j}$$

Fast & Scalable Algorithm

Prop. If $P_\gamma \stackrel{\text{def}}{=} \underset{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists! \boldsymbol{u} \in \mathbb{R}^n_+, \boldsymbol{v} \in \mathbb{R}^m_+$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \stackrel{\text{def}}{=} e^{-M_{\mathbf{XY}}/\gamma}$$

- **[Sinkhorn'64]** fixed-point iterations for $(\boldsymbol{u}, \boldsymbol{v})$

$$\boldsymbol{u} \leftarrow \boldsymbol{a}/K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b}/K^T\boldsymbol{u}$$

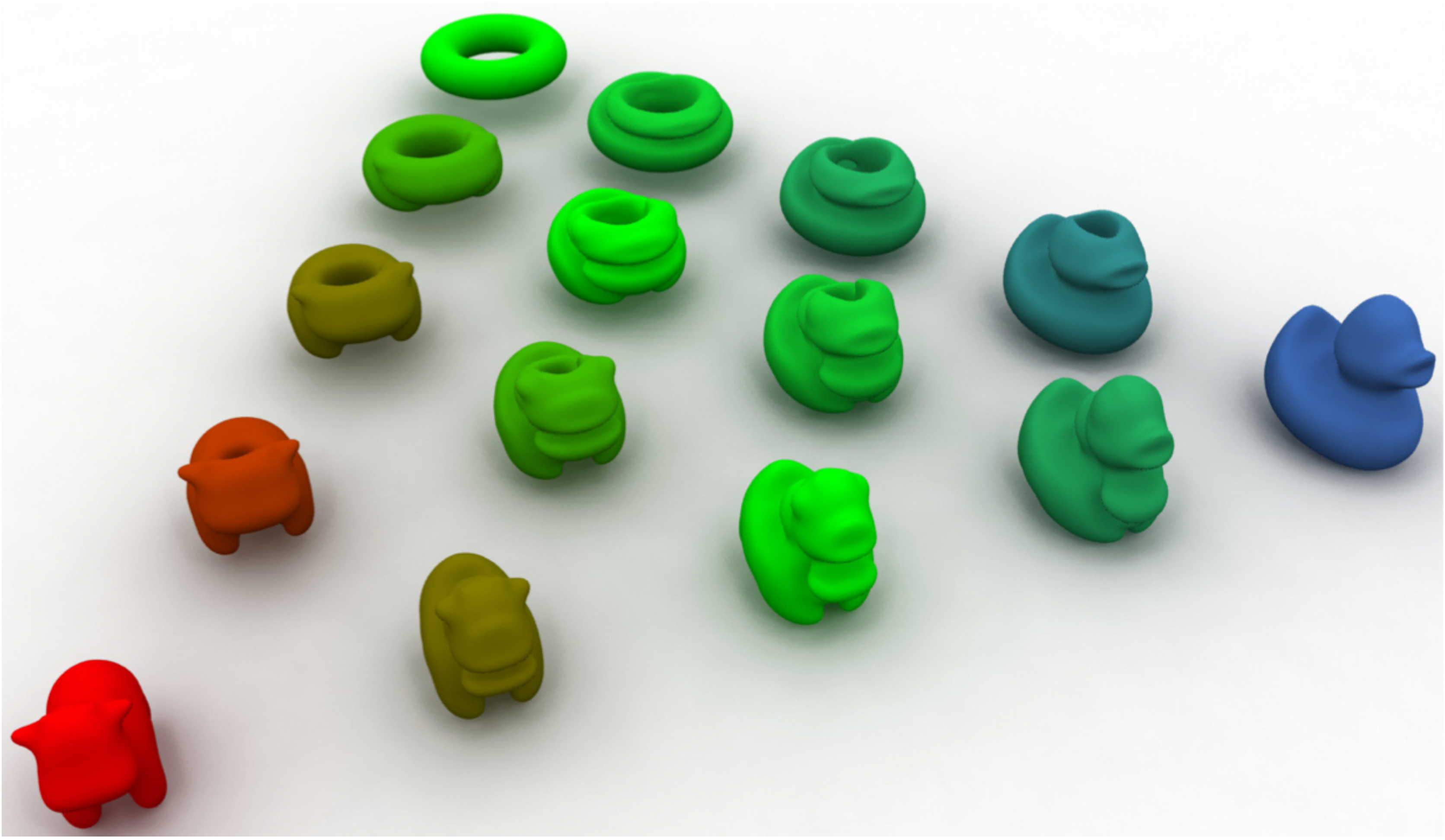
- $O(nm)$ complexity, GPGPU parallel **[C'13]** .
- $O(n^{d+1})$ if $\Omega = \{1, \ldots, n\}^d$ and D^p separable.

[S..C..'15]

(Application: Barycenters)

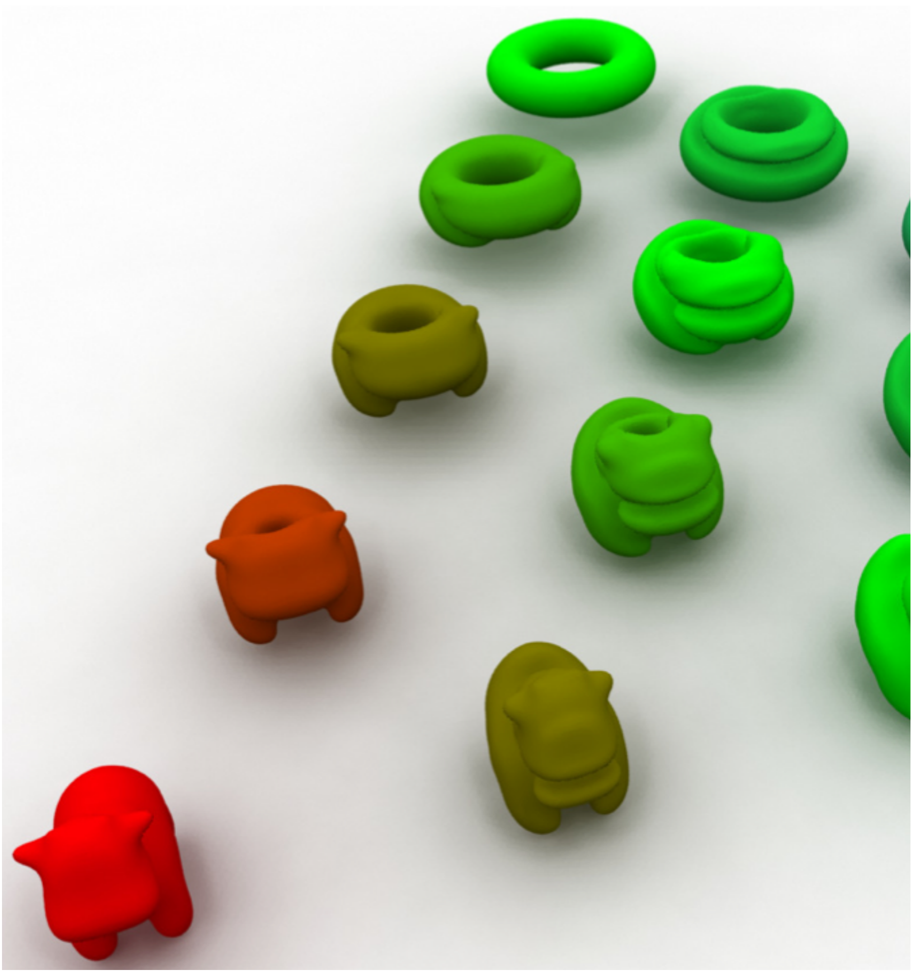
Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15 **[S..C..'15]**

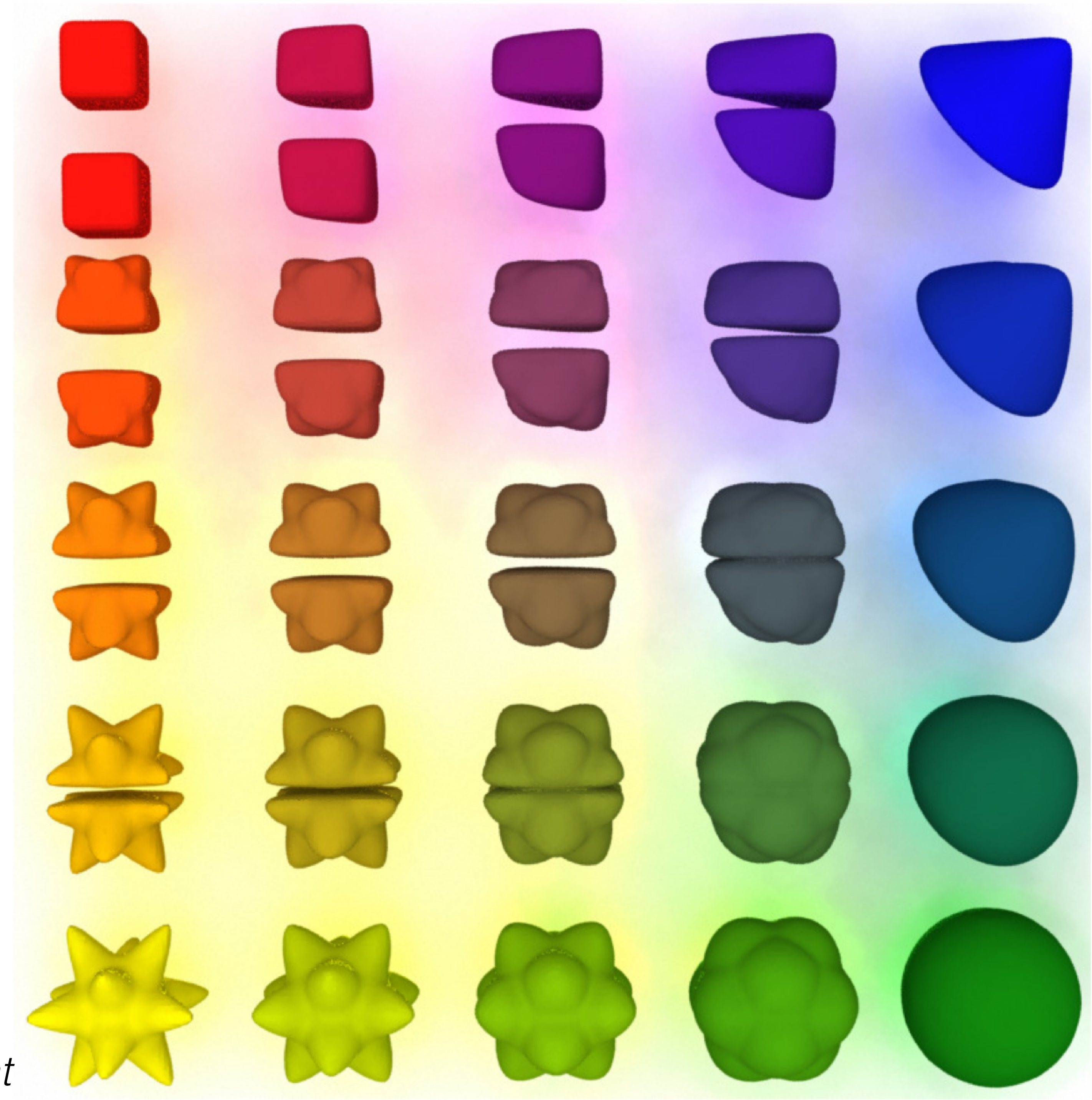
(Application: Barycenters)



Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15 **[S..C..'15]**

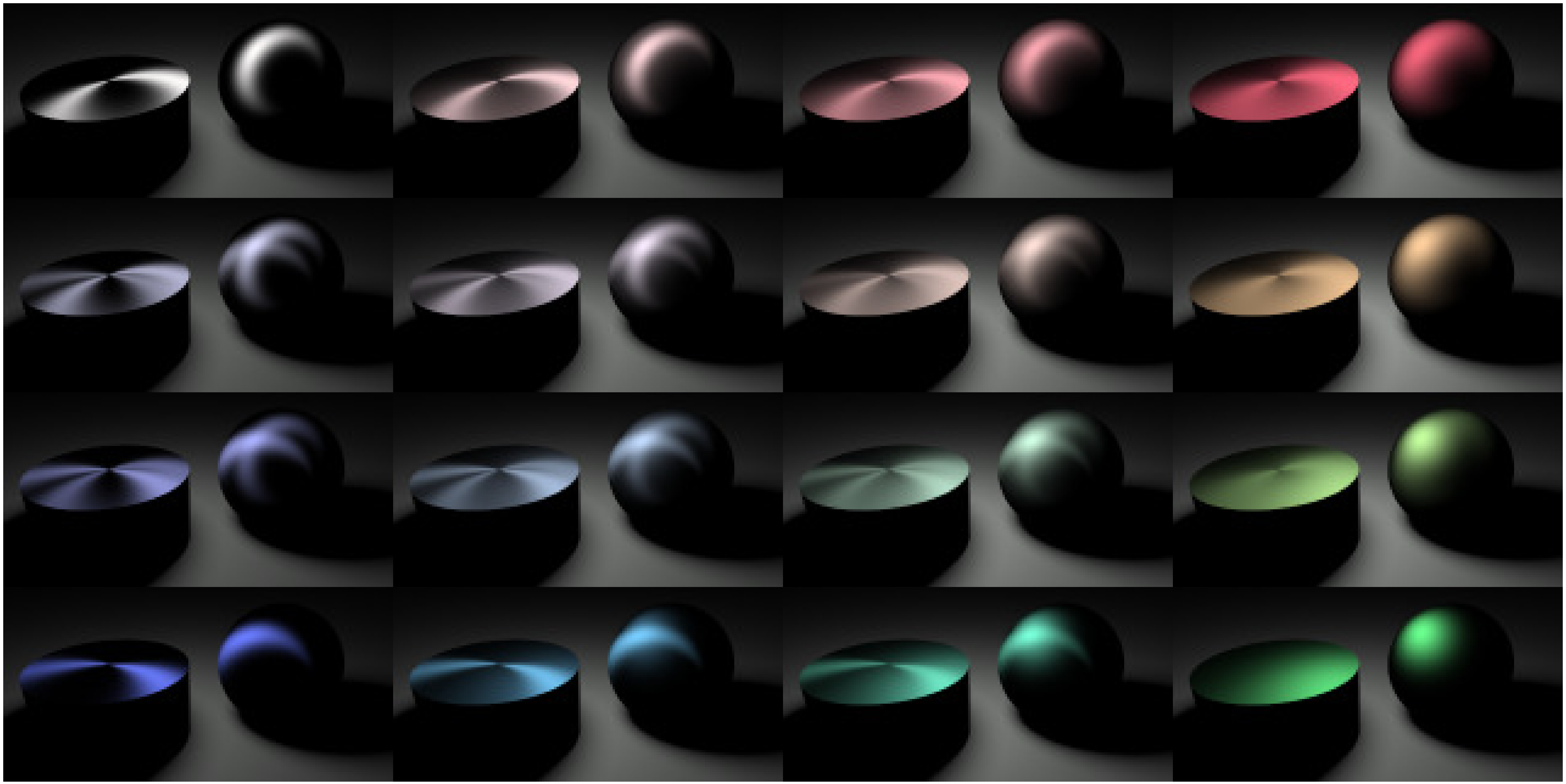
(Application: Barycenters)





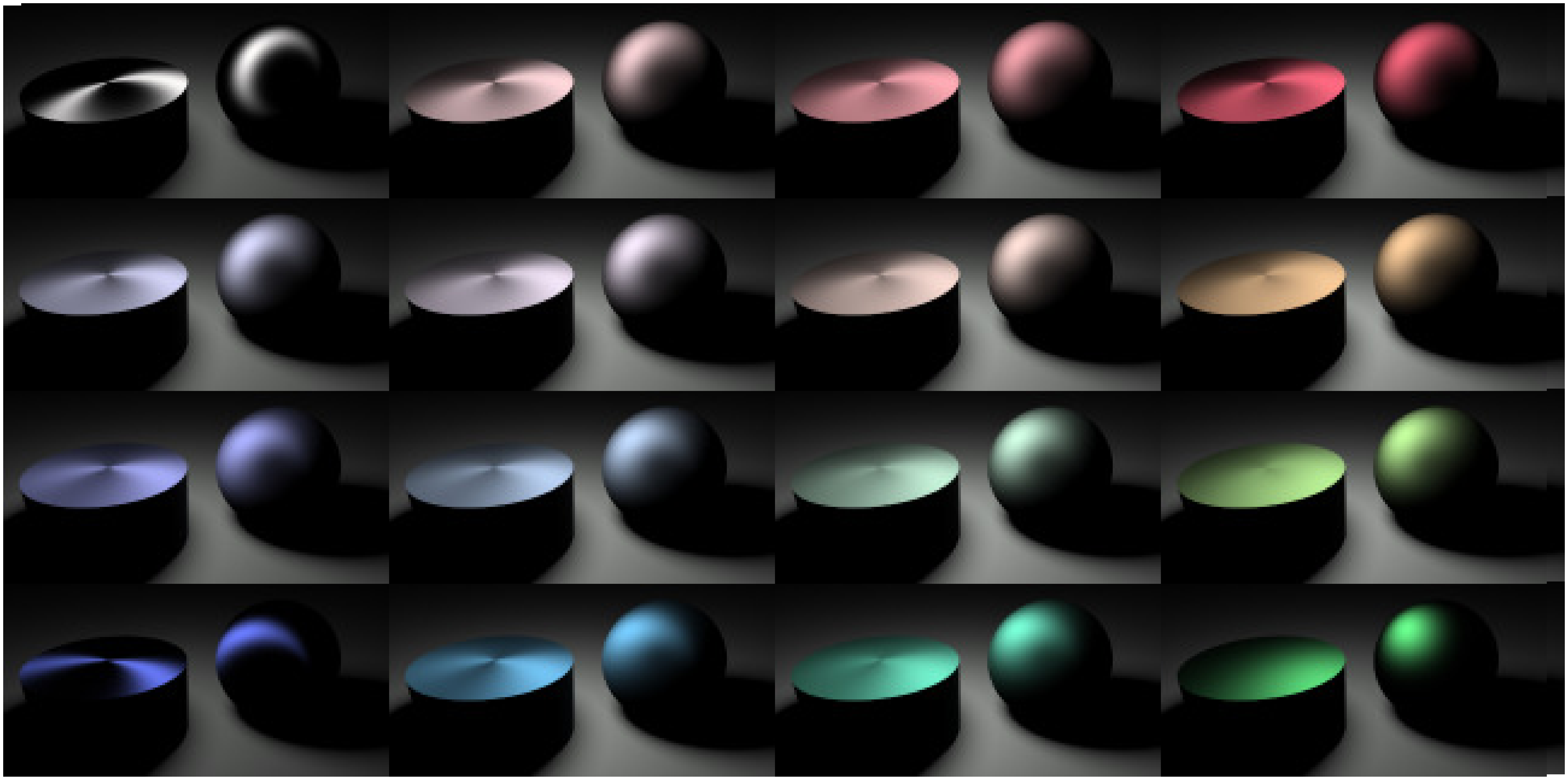
Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15 **[S..C..'15]**

(Application: Barycenters)



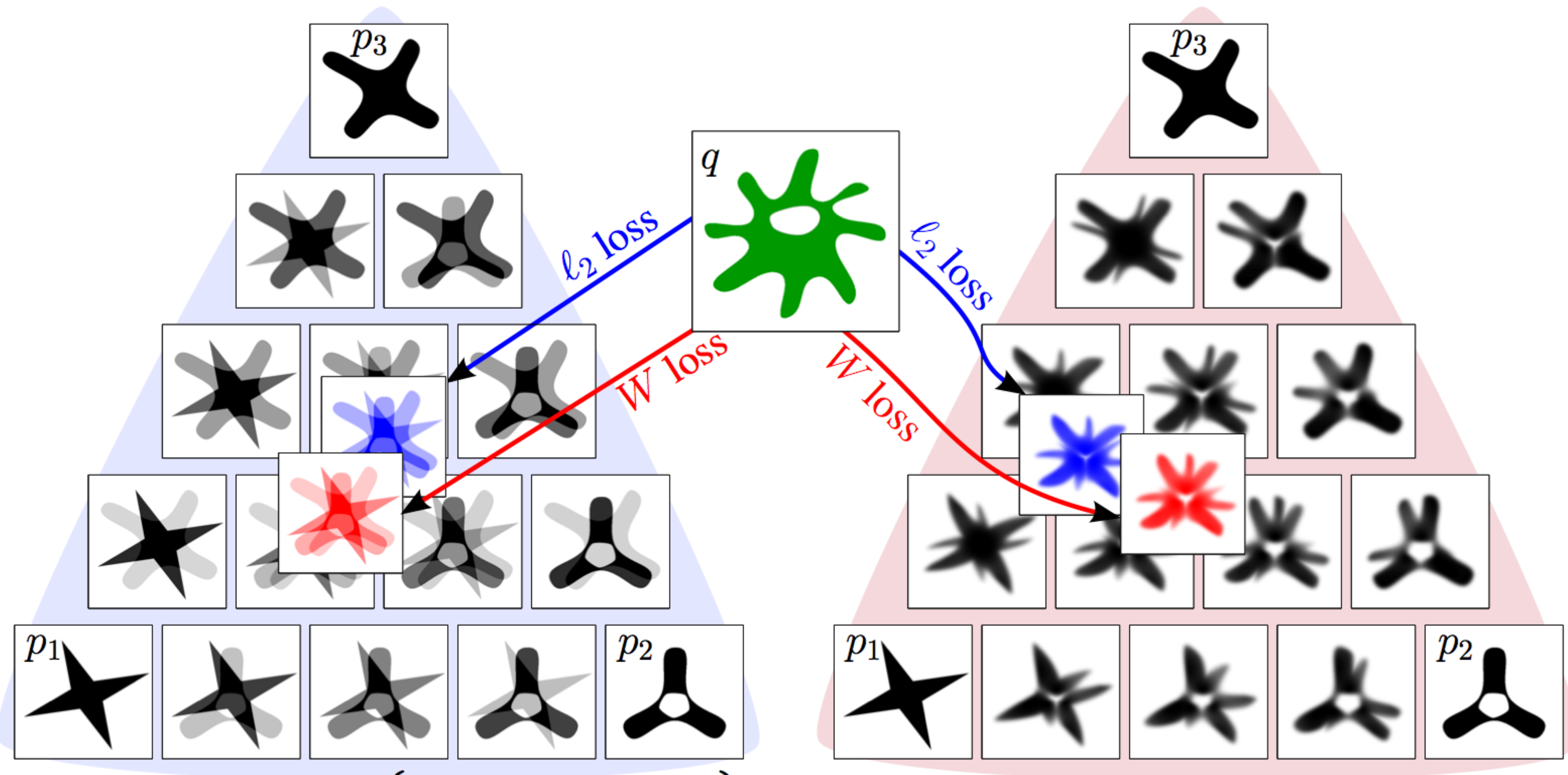
Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15 **[S..C..'15]**

(Application: Barycenters)



Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15 **[S..C..'15]**

(Application: Wasserstein Regression)

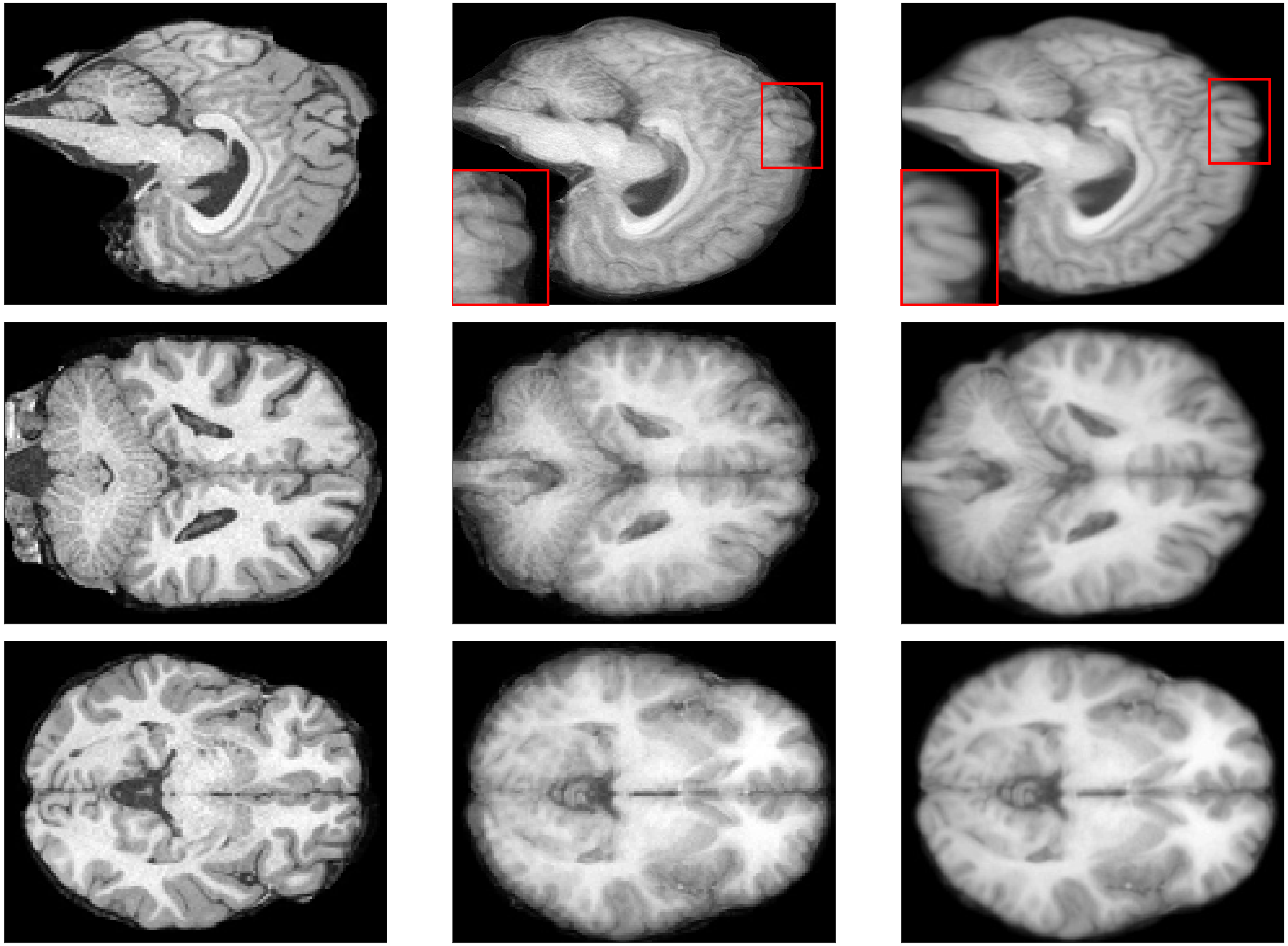


Euclidean Simplex: $\left\{\sum_{i=1}^{3} \lambda_i p_i, \lambda \in \Sigma_3\right\}$

Wasserstein simplex: $\{P(\lambda), \lambda \in \Sigma_3\}$

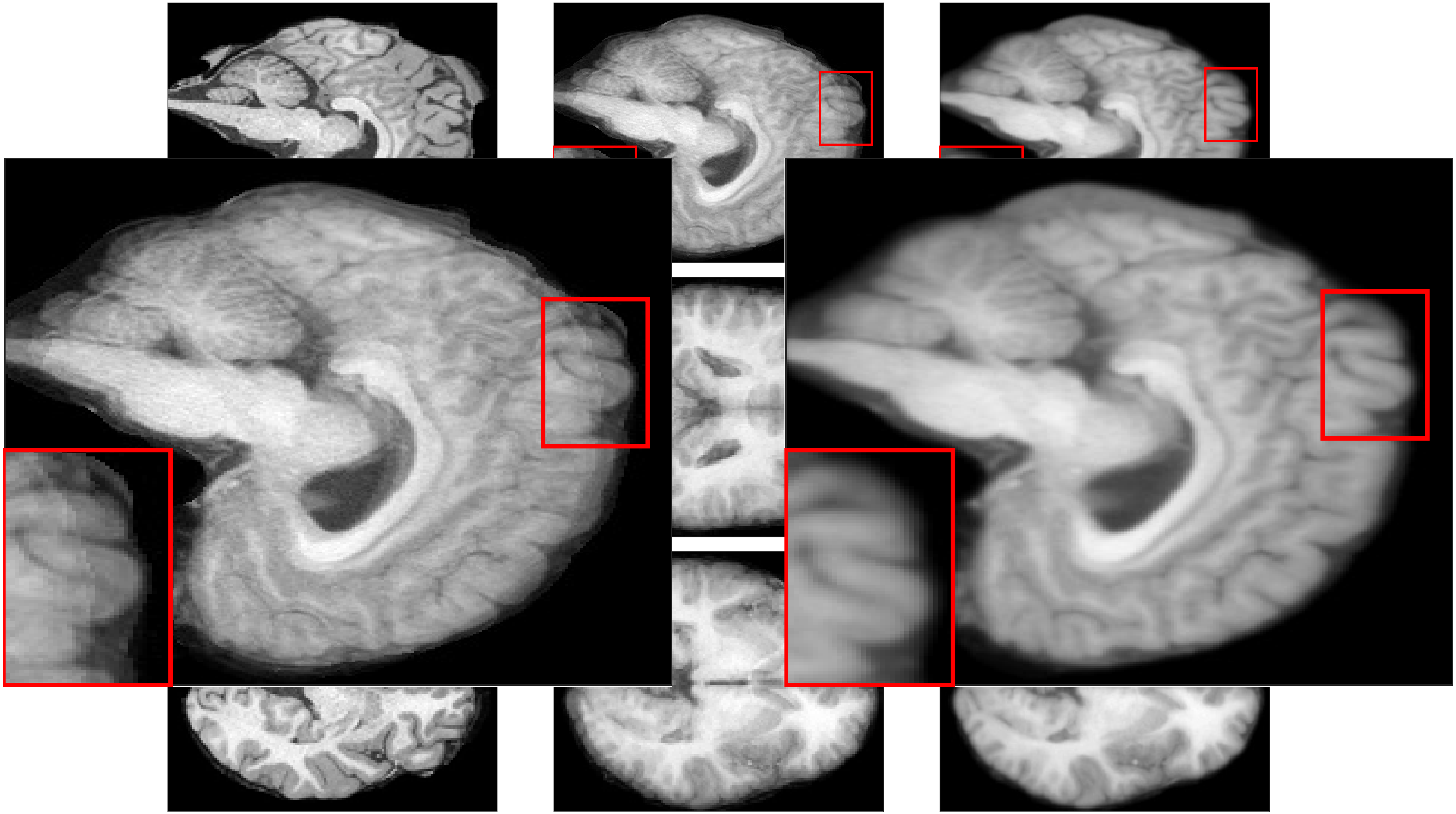
Wasserstein Barycentric Coordinates: Histogram Regression using Optimal Transport, **SIGGRAPH'16**

(Application: Brain Regression)



Original | Euclidean projection | Wasserstein projection

(Application: Brain Regression)



Original

Euclidean projection

Wasserstein projection

Algorithmic Formulation

Def. For $L \geq 1$, define

$$W_L(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \langle \boldsymbol{P_L}, M_{\mathbf{XY}} \rangle,$$

where $\boldsymbol{P_L} \stackrel{\text{def}}{=} \mathbf{diag}(\boldsymbol{u_L}) K \mathbf{diag}(\boldsymbol{v_L}),$

$\boldsymbol{v_0} = \mathbf{1}_m; l \geq 0, \boldsymbol{u_l} \stackrel{\text{def}}{=} \boldsymbol{a} / K \boldsymbol{v_l}, \boldsymbol{v_{l+1}} \stackrel{\text{def}}{=} \boldsymbol{b} / K^T \boldsymbol{u_l}.$

Prop. $\frac{\partial W_L}{\partial \mathbf{X}}, \frac{\partial W_L}{\partial \boldsymbol{a}}$ can be computed recursively, in $O(L)$ kernel $K \times$vector products.

Algorithmic Formulation

Def. For $L \geq 1$, define

$$\underline{\mathrm{W}}_L(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \gamma \boldsymbol{a}^T \log \boldsymbol{u_L} + \gamma \boldsymbol{b}^T \log \boldsymbol{v_L},$$

$$\boldsymbol{v_0} = \mathbf{1}_m; l \geq 0, \boldsymbol{u_l} \stackrel{\text{def}}{=} \boldsymbol{a}/K\boldsymbol{v_l}, \boldsymbol{v_{l+1}} \stackrel{\text{def}}{=} \boldsymbol{b}/K^T\boldsymbol{u_l}.$$

Prop. $\frac{\partial \underline{\mathrm{W}}_L}{\partial \mathbf{X}}, \frac{\partial \underline{\mathrm{W}}_L}{\partial \boldsymbol{a}}$ can be computed recursively, in $O(L)$ kernel $K \times$vector products.

Algorithmic Formulation

Example: Differentiability w.r.t. a

$$\left(\frac{\partial \boldsymbol{v_0}}{\partial a}\right)^T = \mathbf{0}_{m\times n},$$

$$\left(\frac{\partial \boldsymbol{u_l}}{\partial a}\right)^T \boldsymbol{x} = \frac{\boldsymbol{x}}{K\boldsymbol{v_l}} - \left(\frac{\partial \boldsymbol{v_l}}{\partial a}\right)^T K^T \frac{\boldsymbol{x}\circ a}{(K\boldsymbol{v_l})^2},$$

$$\left(\frac{\partial \boldsymbol{v_{l+1}}}{\partial a}\right)^T \boldsymbol{y} = -\left(\frac{\partial \boldsymbol{u_l}}{\partial a}\right)^T K \frac{\boldsymbol{y}\circ b}{(K^T\boldsymbol{u_l})^2}.$$

4. Algorithmic Formulation

Example: Differentiability w.r.t. a

$$N = K \circ M_{\boldsymbol{XY}}$$

$$\nabla_{\boldsymbol{a}} W_L(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left(\frac{\partial \boldsymbol{u_L}}{\partial a}\right)^T N \boldsymbol{v_L} + \left(\frac{\partial \boldsymbol{v_L}}{\partial a}\right)^T N^T \boldsymbol{u_L}$$

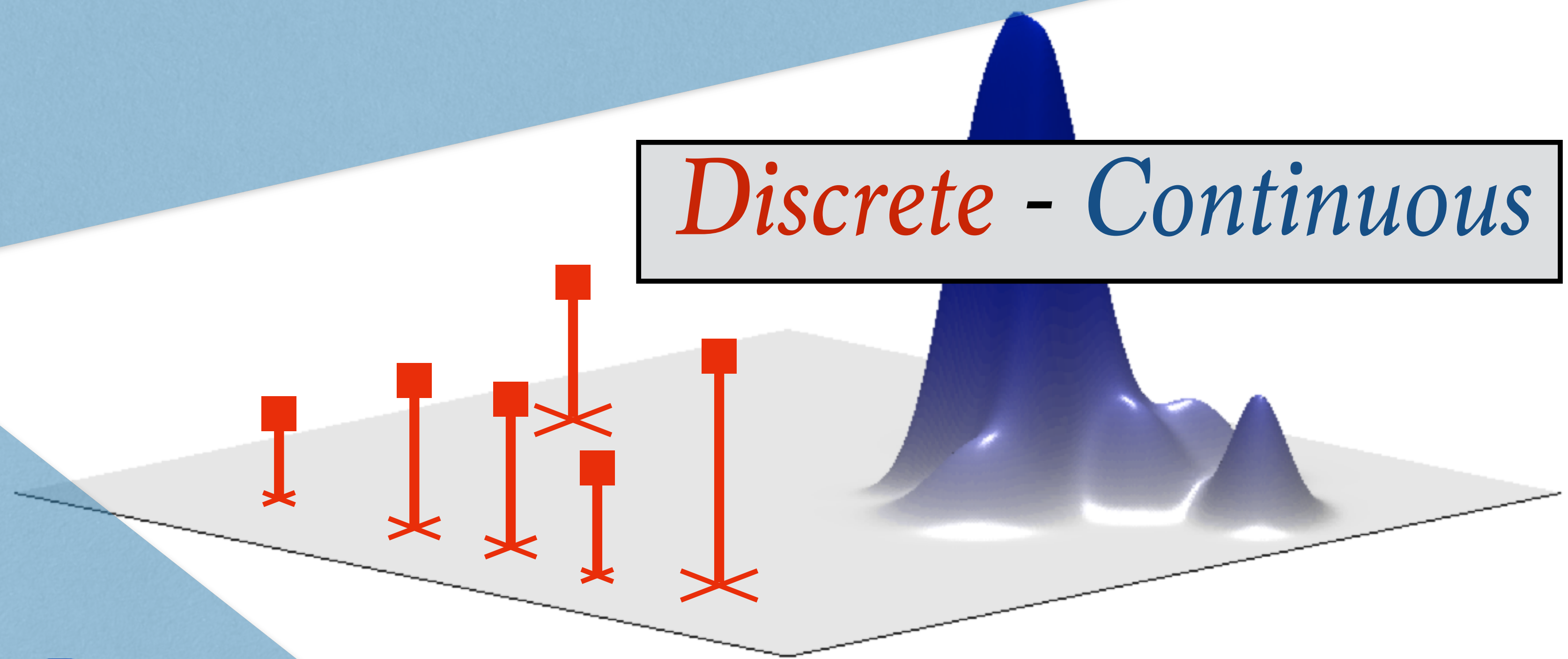
```
function [d,grad_a,grad_b,hess_a,hess_b] = sinkhornObjGradHess(a,b,K,M,niter)

u_update = @(v,a) a./(K*v);
v_update = @(u,b) b./(K'*u);

% DuDa = @(eps,dvda,a,v)  (eps./(K*v))- (a./((K*v).^2)).*(K*dvda(eps));
%
% DvDa = @(eps,duda,b,u)  -(b./((K'*u).^2)).*(K'*duda(eps));
%
% DuDb = @(eps,dvdb,a,v)  -(a./((K*v).^2)).*(K*dvdb(eps));
%
% DvDb = @(eps,dudb,b,u)  (eps./(K'*u))-(b./((K'*u).^2)).*(K'*dudb(eps));


DuDat = @(x,dvdat,a,v)  bsxfun(@rdivide,x,K*v)... (x./(K*v))
    -dvdat(K'*( bsxfun(@times,x,(a./((K*v).^2)))));...-dvdat(K'*( (a./((K*v).^2)).*x));


DvDat = @(x,dudat,b,u)  -dudat(K*(bsxfun(@times,x,(b./((K'*u).^2))))); ...(b./((K'*u).^2)).*x))


JDuDat= @(x,Jdvdat,dvdat,a,v) -diag((x'*dvdat(K'))'./((K*v).^2)) ...(K*dvda(x))
    - Jdvdat(x)*K'*diag(a./((K*v).^2))...
    - dvdat(K'* ...
    ( diag(a.*( (-2*(x'*dvdat(K'))')./((K*v).^3)))+...
    diag(x./((K*v).^2))  ));          %1

JDvDat = @(x,Jdudat,dudat,b,u) ...
    -Jdudat(x)*K*diag(b./((K'*u).^2))...
    - dudat(K)* ( ...
    diag(b.*( (-2* (x'*dudat(K))')./((K'*u).^3)))) ;...
```

```
DuDbt = @(x,dvdbt,a,v)  -dvdbt(K'*(bsxfun(@times,x,(a./((K*v).^2))))); ...(a./((K*v).^2)).*x));

DvDbt = @(x,dudbt,b,u)  bsxfun(@rdivide,x,K'*u) ... (x./(K'*u))...
    -dudbt(K*( bsxfun(@times,x,(b./((K'*u).^2)))));...( b./((K'*u).^2)) .*x));




JDvDbt= @(x,Jdudbt,dudbt,b,u) -diag((x'*dudbt(K))'./((K'*u).^2)) ...  (K'*dudb(x))
    - Jdudbt(x)*K*diag(b./((K'*u).^2))...
    - dudbt(K)* ( ...
    diag(b.*( (-2*(x'*dudbt(K))')./((K'*u).^3)))+...
    diag(x./((K'*u).^2))   ) ;

JDuDbt = @(x,Jdvdbt,dvdbt,a,v) ...
    -Jdvdbt(x)*K'*diag(a./((K*v).^2))...
    - dvdbt(K')* ( ...
    diag(a.*( (-2* (x'*dvdbt(K'))')./((K*v).^3)))) ;
```

```
n=size(a,1);
m=size(b,1);

DVDAT= @(eps) zeros(n,size(eps,2));
DVDBT= @(eps) zeros(m,size(eps,2));

JDVDAT= @(eps) zeros(n,m);
JDVDBT= @(eps) zeros(m,m);

v=ones(m,size(b,2));

for j=1:niter,
    u=u_update(v,a);
    DUDAT = @(x) DuDat(x,DVDAT,a,v);
    DUDBT = @(x) DuDbt(x,DVDBT,a,v);

    if nargout>3
        JDUDAT = @(x) JDuDat(x,JDVDAT,DVDAT,a,v);
        JDUDBT = @(x) JDuDbt(x,JDVDBT,DVDBT,a,v);
    end


    v=v_update(u,b);
    DVDAT = @(x) DvDat(x,DUDAT,b,u);
    DVDBT = @(x) DvDbt(x,DUDBT,b,u);

    if nargout>3
        JDVDAT = @(x) JDvDat(x,JDUDAT,DUDAT,b,u);
        JDVDBT = @(x) JDvDbt(x,JDUDBT,DUDBT,b,u);
    end
end
```

```
U=K.*M;
d=diag(u'*U*v);

grad_a=(DUDAT(U*v)+DVDAT(U'*u));
grad_b=(DUDBT(U*v)+DVDBT(U'*u));


if nargout>3
    hess_a= @(eps) JDUDAT(eps)*(U*v)+DUDAT((eps'*DVDAT(U'))')+...
        JDVDAT(eps)*(U'*u)+DVDAT((eps'*DUDAT(U))');
end
```

W is versatile



W is versatile

D transforms

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

D transforms

$$W_p^p(\mu, \nu) = \sup_{\substack{\varphi \in L_1(\mu), \psi \in L_1(\nu) \\ \varphi(x) + \psi(y) \leq D^p(x,y)}} \int \varphi d\mu + \int \psi d\nu.$$

DUAL

For given φ, cannot get a better ψ than

$$\varphi^D(y) \stackrel{\text{def}}{=} \inf_x D^p(x, y) - \varphi(x).$$

$$W_p^p(\mu, \nu) = \sup_{\varphi} \int \varphi d\mu + \int \varphi^D d\nu.$$

SEMI-DUAL

D transforms

$$W_p^p(\mu, \nu) = \sup_{\varphi} \int \varphi d\mu + \int \varphi^D d\nu.$$

SEMI-DUAL

$$\varphi^D(y) \stackrel{\text{def}}{=} \inf_x D^p(x,y) - \varphi(x).$$

$$\varphi^{DD}(x) = \inf_y D^p(x,y) - \varphi^D(y).$$

φ is D concave if $\exists \phi : \varphi = \phi^D$

D transforms

$$\varphi^{D}(y) \stackrel{\text{def}}{=} \inf_{x} D^p(x,y) - \varphi(x).$$

$$\varphi^{DD}(x) = \inf_{y} D^p(x,y) - \varphi^{D}(y).$$

φ is D concave if $\exists \phi : \varphi = \phi^{D}$

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\varphi \text{ is } D\text{-concave}} \int \varphi \, d\boldsymbol{\mu} + \int \varphi^{D} d\boldsymbol{\nu}.$$

SEMI-DUAL

Reminder: dual regularization

$$W_p^p(\mu,\nu) = \sup_{\varphi,\psi} \int \varphi d\mu + \int \psi d\nu - \iota_C(\varphi,\psi)$$

$$C = \{(\varphi,\psi) | \varphi \oplus \psi \leq D^p\}$$

DUAL

regularizing dual ↓ *constraints* $\gamma > 0$

$$W_\gamma(\mu,\nu) = \sup_{\varphi,\psi} \int \varphi d\mu + \int \psi d\nu - \iota_C^\gamma(\varphi,\psi)$$

$$\iota_C^\gamma(\varphi,\psi) = \gamma \iint e^{(\varphi \oplus \psi - D^p)/\gamma} d\mu d\nu$$

REGULARIZED DUAL

Smoothed D transforms

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iota_C^\gamma(\boldsymbol{\varphi},\boldsymbol{\psi})$$

$$\iota_C^\gamma(\boldsymbol{\varphi},\boldsymbol{\psi}) = \gamma \iint e^{(\boldsymbol{\varphi}\oplus\boldsymbol{\psi} - \boldsymbol{D}^p)/\gamma} d\boldsymbol{\mu} d\boldsymbol{\nu}$$

REGULARIZED DUAL

$\nabla_{\boldsymbol{\psi}} = 0$ $\gamma > 0$

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\varphi}^{\boldsymbol{D},\gamma} d\boldsymbol{\nu}.$$

$$\boldsymbol{\varphi}^{\boldsymbol{D},\gamma}(\boldsymbol{y}) = -\gamma \log \int e^{\frac{\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})^p}{\gamma}} d\boldsymbol{\mu}(\boldsymbol{x})$$

REGULARIZED SEMI-DUAL

Regularized Semidual Wasserstein

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\varphi}^{\boldsymbol{D},\gamma} d\boldsymbol{\nu}.$$

$$\boldsymbol{\varphi}^{\boldsymbol{D},\gamma}(\boldsymbol{y}) = -\gamma \log \int e^{\frac{\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})^p}{\gamma}} d\boldsymbol{\mu}(\boldsymbol{x})$$

REGULARIZED SEMI-DUAL

substituting

$$\sup_{\boldsymbol{\varphi}} \int_{\boldsymbol{y}} \left[\int_{\boldsymbol{x}} \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{x}) - \gamma \log \int_{\boldsymbol{x}} e^{\frac{\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})^p}{\gamma}} d\boldsymbol{\mu}(\boldsymbol{x}) \right] d\boldsymbol{\nu}(\boldsymbol{y}).$$

REGULARIZED SEMI-DUAL

Semi-discrete case: Stochastic Opt.

$$\sup_{\boldsymbol{\varphi}} \int_{\boldsymbol{y}} \left[\int_{\boldsymbol{x}} \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{x}) - \gamma \log \int_{\boldsymbol{x}} e^{\frac{\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})^p}{\gamma}} d\boldsymbol{\mu}(\boldsymbol{x}) \right] d\boldsymbol{\nu}(\boldsymbol{y}).$$

REGULARIZED SEMI-DUAL

Semi-discrete case: Stochastic Opt.

$$\sup_{\boldsymbol{\varphi}} \int_{\boldsymbol{y}} \left[\int_{\boldsymbol{x}} \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{x}) - \gamma \log \int_{\boldsymbol{x}} e^{\frac{\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})^p}{\gamma}} d\boldsymbol{\mu}(\boldsymbol{x}) \right] d\boldsymbol{\nu}(\boldsymbol{y}).$$

REGULARIZED SEMI-DUAL

What if $\boldsymbol{\mu}$ is a discrete measure?

$\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a_i} \delta_{\boldsymbol{x_i}}$

$\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu})$ is now just a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$!

Semi-discrete case: Stochastic Opt.

$$\sup_{\varphi} \int_{y} \left[\int_{x} \varphi(x) d\mu(x) - \gamma \log \int_{x} e^{\frac{\varphi(x) - D(x,y)^p}{\gamma}} d\mu(x) \right] d\nu(y).$$

REGULARIZED SEMI-DUAL

What if μ is a discrete measure?

$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$

$\varphi \in L_1(\mu)$ is now just a vector $\alpha \in \mathbb{R}^n$!

$$\sup_{\alpha \in \mathbb{R}^n} \int_{y} \left[\sum_{i=1}^{n} \alpha_i a_i - \gamma \log \sum_{i=1}^{n} e^{\frac{\alpha_i - D(x_i,y)^p}{\gamma}} a_i \right] d\nu(y)$$

$$= \sup_{\alpha \in \mathbb{R}^n} \mathbb{E}_{\nu}[f(\alpha, y)]$$

STOCHASTIC REGULARIZED SEMI-DUAL

(in Discrete Setting)

$$\sup_{\boldsymbol{\varphi}} \int_{\boldsymbol{y}} \left[\int_{\boldsymbol{x}} \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{x}) - \gamma \log \int_{\boldsymbol{x}} e^{\frac{\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})^p}{\gamma}} d\boldsymbol{\mu}(\boldsymbol{x}) \right] d\boldsymbol{\nu}(\boldsymbol{y}).$$

REGULARIZED SEMI-DUAL

(in Discrete Setting)

$$\sup_{\boldsymbol{\varphi}} \int_{\boldsymbol{y}} \left[\int_{\boldsymbol{x}} \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{x}) - \gamma \log \int_{\boldsymbol{x}} e^{\frac{\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})^p}{\gamma}} d\boldsymbol{\mu}(\boldsymbol{x}) \right] d\boldsymbol{\nu}(\boldsymbol{y}).$$

REGULARIZED SEMI-DUAL

What if $\boldsymbol{\nu}$ is **also** a discrete measure?

$\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a_i} \delta_{\boldsymbol{x_i}}$

$\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b_j} \delta_{\boldsymbol{y_j}}$

(in Discrete Setting)

$$\sup_{\varphi} \int_{y} \left[\int_{x} \varphi(x) d\mu(x) - \gamma \log \int_{x} e^{\frac{\varphi(x) - D(x,y)^p}{\gamma}} d\mu(x) \right] d\nu(y).$$

REGULARIZED SEMI-DUAL

What if ν is **also** a discrete measure?

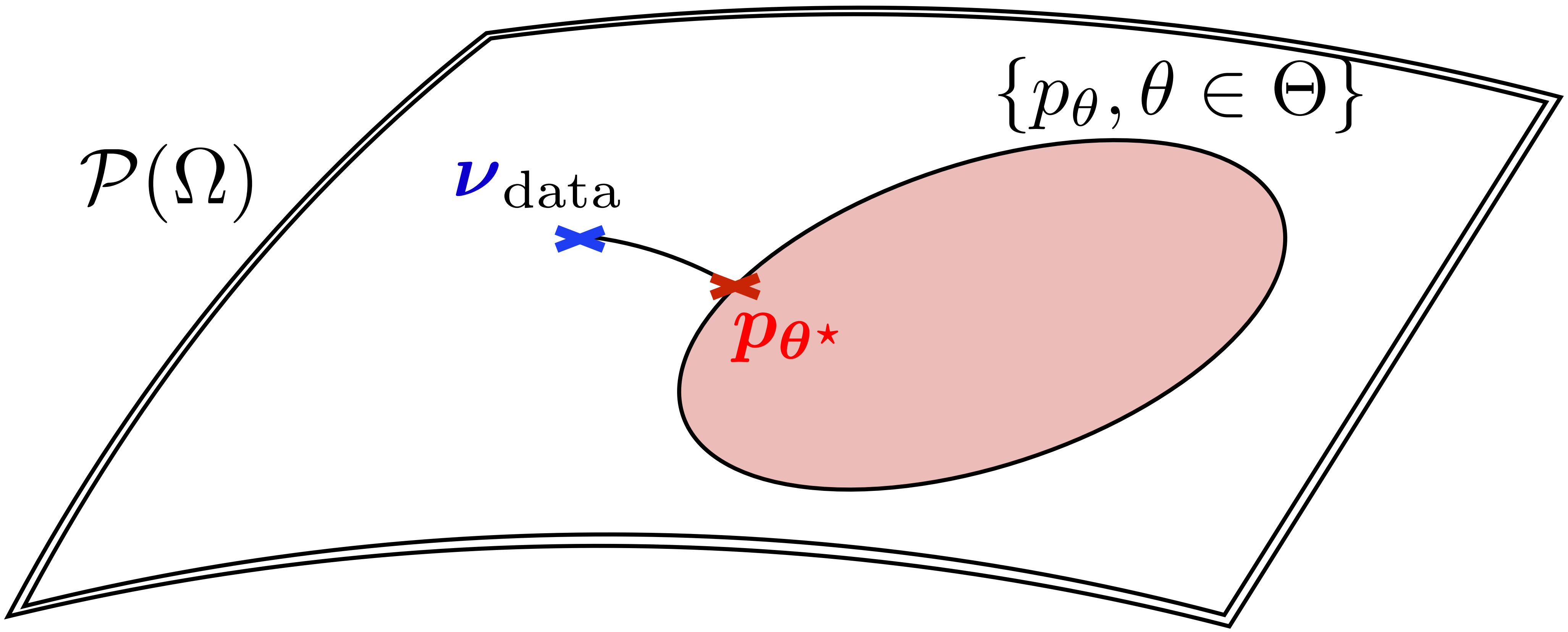
$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ $\qquad$ $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$

$$\sup_{\alpha \in \mathbb{R}^n} \int_{y} \left[\sum_{i=1}^{n} \alpha_i a_i - \gamma \log \sum_{i=1}^{n} e^{\frac{\alpha_i - D(x_i, y)^p}{\gamma}} a_i \right] d\nu(y)$$

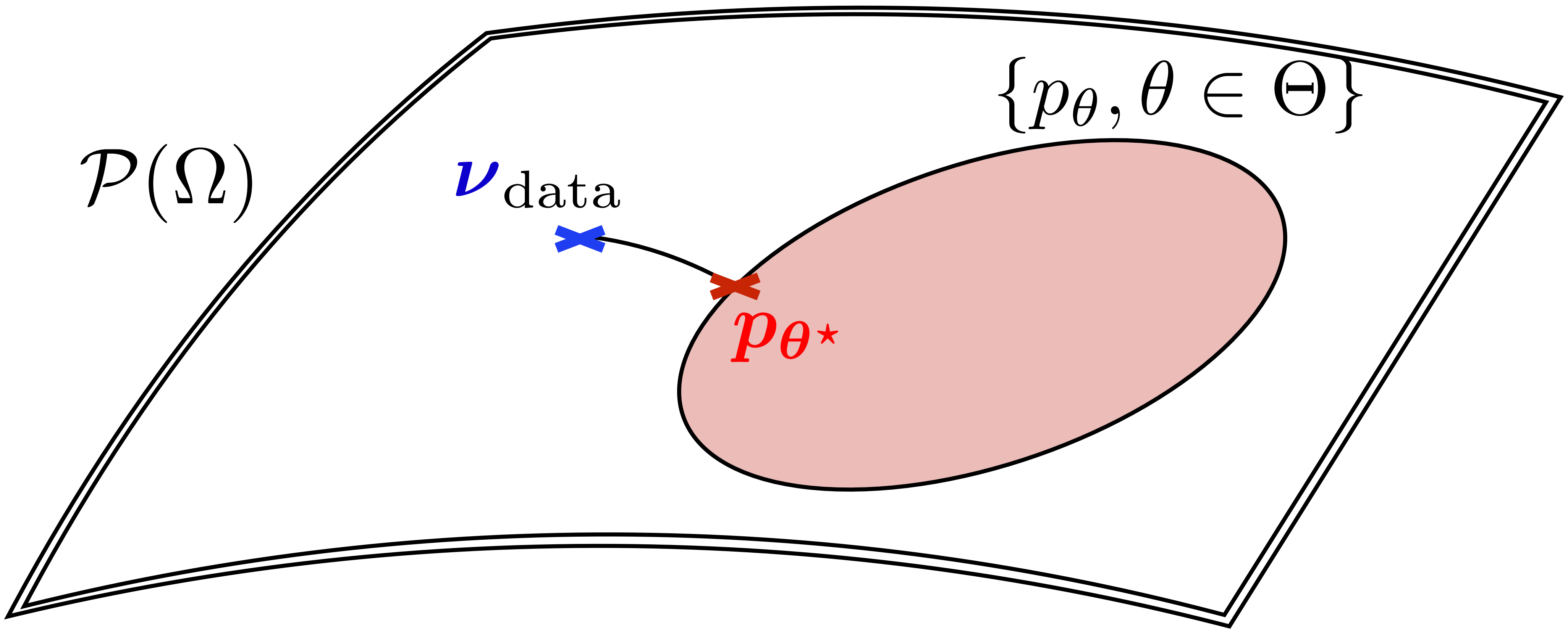
$$\sup_{\alpha \in \mathbb{R}^n} \alpha^T a - \gamma b^T \log K^T (a \odot e^{\frac{\alpha}{\gamma}})$$

REGULARIZED SEMI-DUAL

Minimum Kantorovich Estimators



Minimum Kantorovich Estimators



$$\min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_\theta \|)$$

MLE

$$\min_{\theta \in \Theta} W(\nu_{\text{data}}, p_\theta)$$

MKE

[Bassetti'06]

In a discrete setting

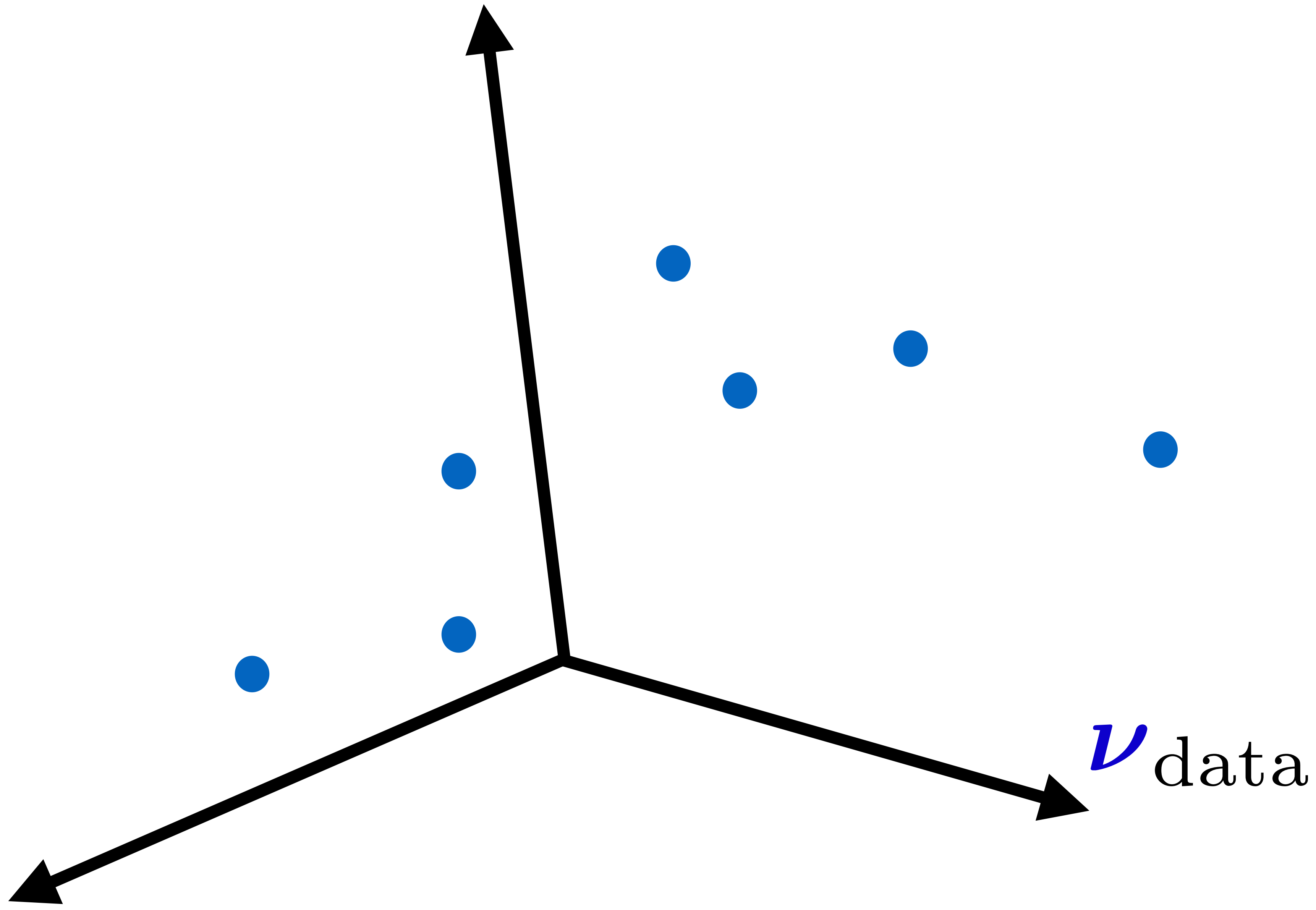
- Suppose Ω is a discrete, finite space.

$$W_\gamma(p_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\text{data}}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \langle \boldsymbol{\alpha}, p_{\boldsymbol{\theta}} \rangle + \langle \boldsymbol{\beta}, \boldsymbol{\nu}_{\text{data}} \rangle - \gamma \langle e^{\boldsymbol{\alpha}/\gamma}, \; K e^{\boldsymbol{\beta}/\gamma} \rangle$$

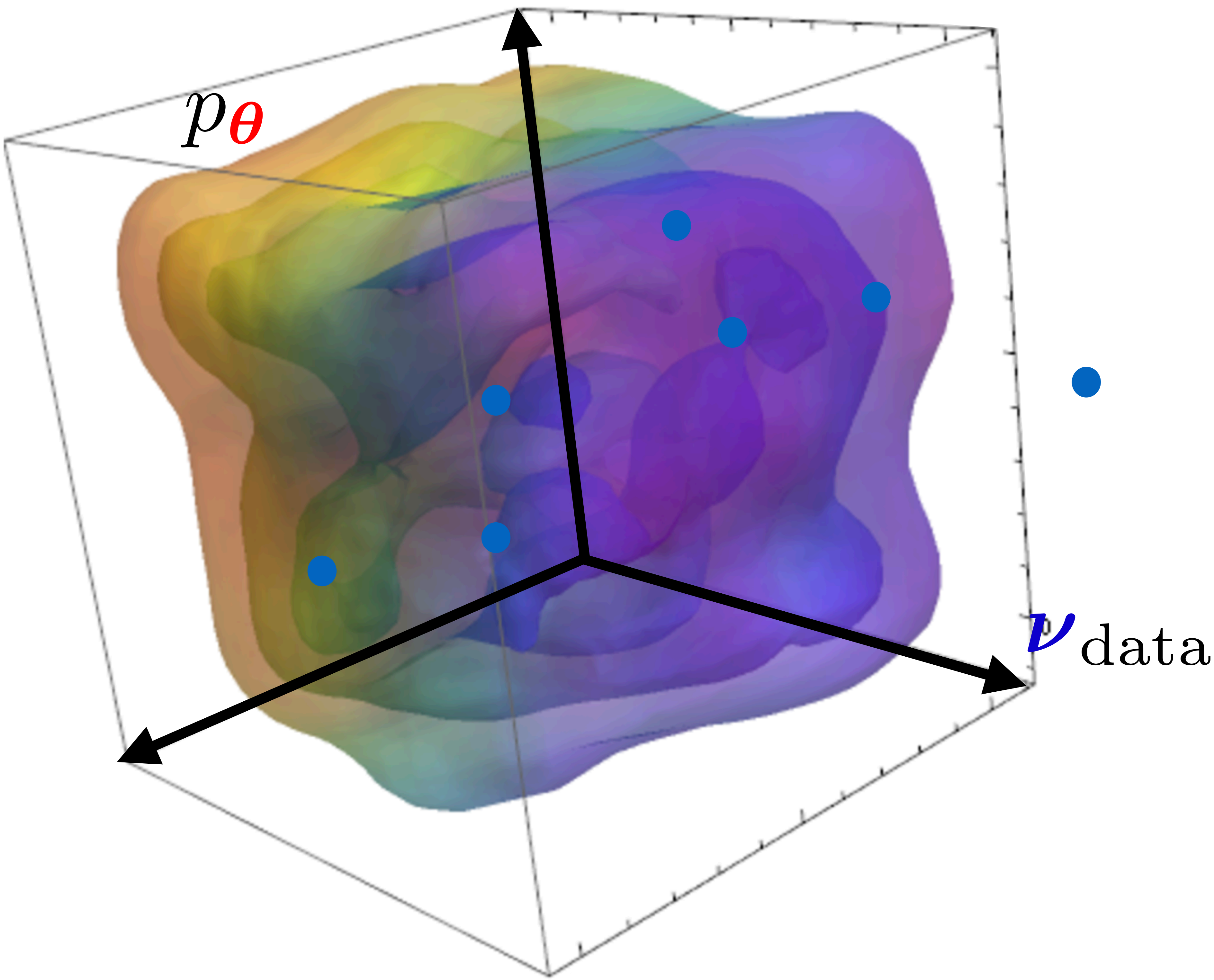
$$\nabla_\theta W_\gamma = \left(\frac{\partial p_\theta}{\partial \theta}\right)^T \boldsymbol{\alpha}^\star$$

- Used for discrete models with very large state spaces in **[MMC'16]**.
- Considered for restricted Boltzmann machines, using stochastic approximation & regularization.

In a continuous observation setting

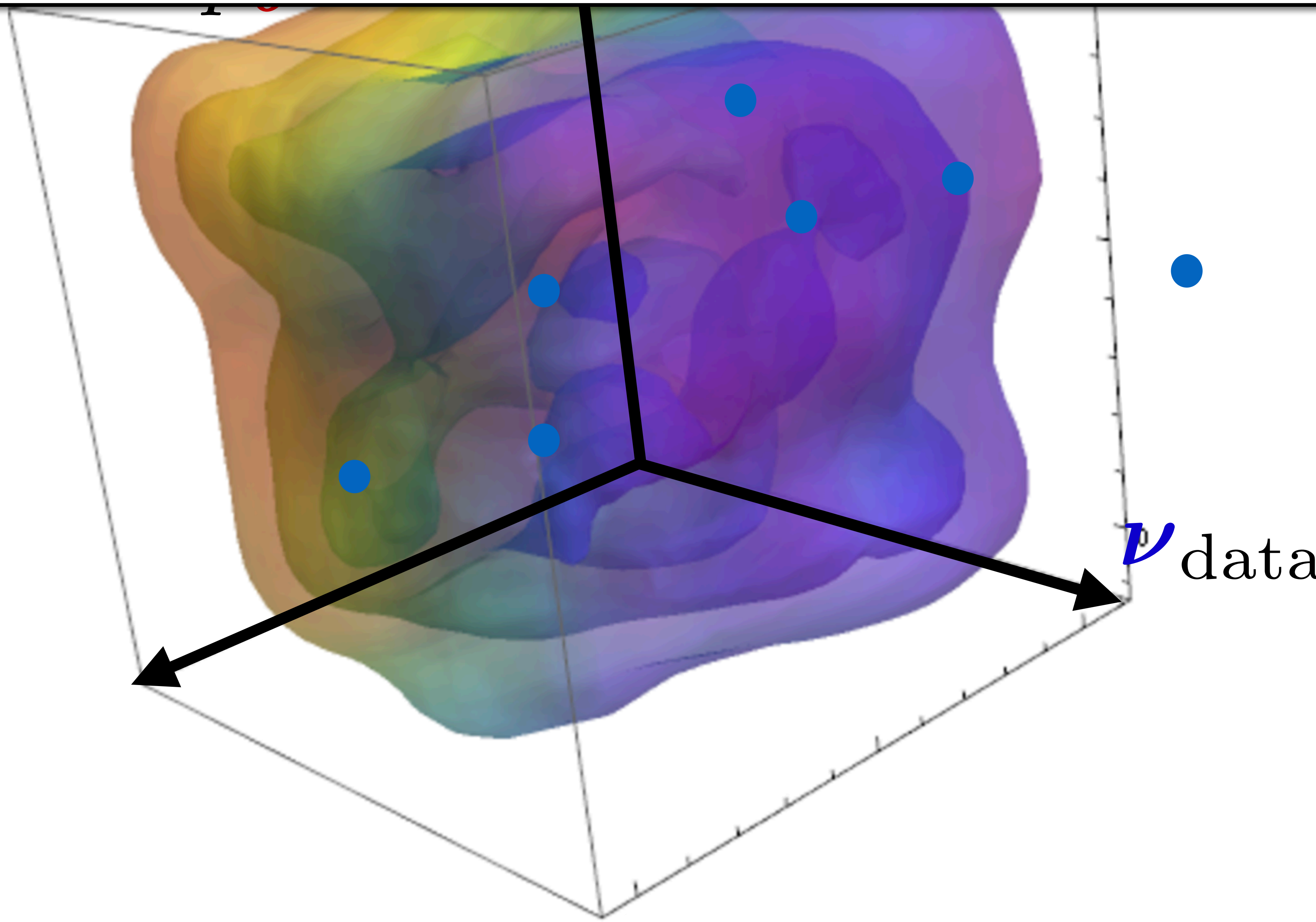


In a continuous observation setting



In a continuous observation setting

$$W_\gamma(p_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\text{data}}) = \max_{\boldsymbol{f},\boldsymbol{b}} \int_\Omega \boldsymbol{f}\, dp_{\boldsymbol{\theta}} + \boldsymbol{b}^T \mathbf{1}_m - \gamma \langle e^{\boldsymbol{f}/\gamma},\ K e^{\boldsymbol{b}/\gamma} \rangle_{p_{\boldsymbol{\theta}}}$$

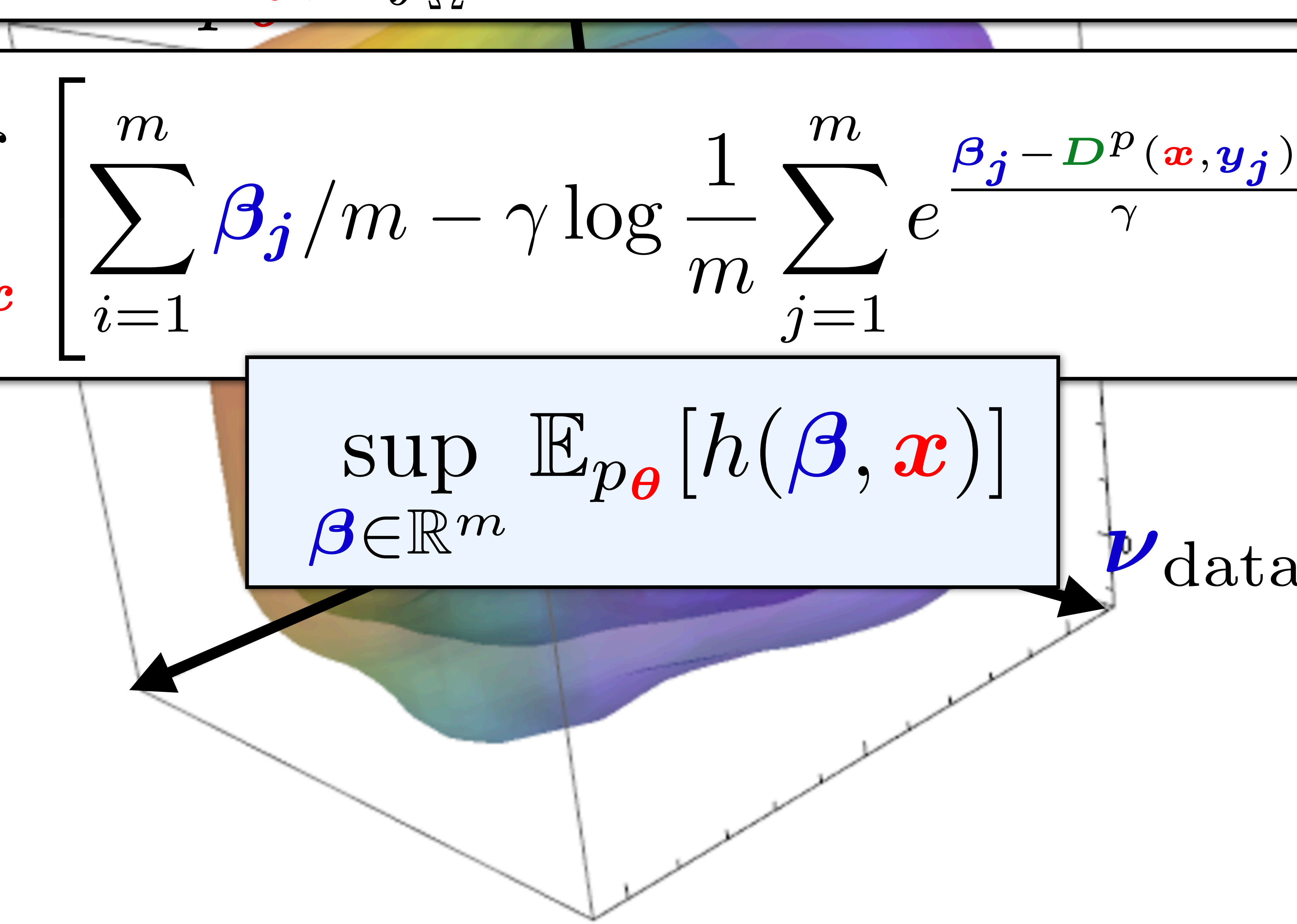


In a continuous observation setting

$$W_\gamma(p_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\text{data}}) = \max_{\boldsymbol{f},\boldsymbol{b}} \int_\Omega \boldsymbol{f}\, dp_{\boldsymbol{\theta}} + \boldsymbol{b}^T \mathbf{1}_m - \gamma \langle e^{\boldsymbol{f}/\gamma},\ K e^{\boldsymbol{b}/\gamma} \rangle_{p_{\boldsymbol{\theta}}}$$

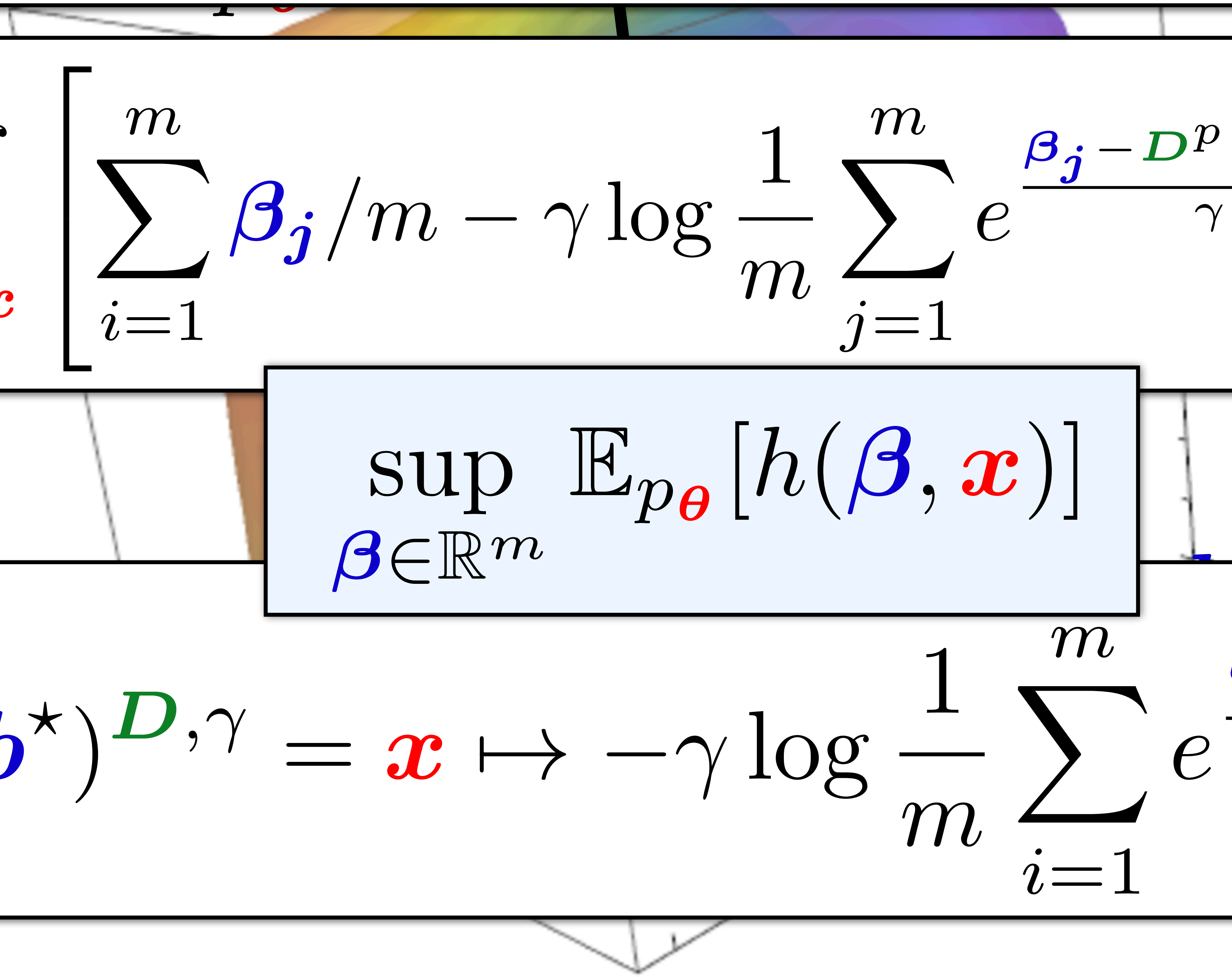
$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^m} \int_{\boldsymbol{x}} \left[\sum_{i=1}^m \boldsymbol{\beta_j}/m - \gamma \log \frac{1}{m} \sum_{j=1}^m e^{\frac{\boldsymbol{\beta_j} - D^p(\boldsymbol{x}, \boldsymbol{y_j})}{\gamma}} \right] p_{\boldsymbol{\theta}}(\boldsymbol{x})$$

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^m} \mathbb{E}_{p_{\boldsymbol{\theta}}}[h(\boldsymbol{\beta}, \boldsymbol{x})]$$



In a continuous observation setting

$$W_\gamma(p_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\text{data}}) = \max_{\boldsymbol{f},\boldsymbol{b}} \int_\Omega \boldsymbol{f}\, dp_{\boldsymbol{\theta}} + \boldsymbol{b}^T \mathbf{1}_m - \gamma \langle e^{\boldsymbol{f}/\gamma},\; K e^{\boldsymbol{b}/\gamma} \rangle_{p_{\boldsymbol{\theta}}}$$



$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^m} \int_{\boldsymbol{x}} \left[\sum_{i=1}^m \boldsymbol{\beta_j}/m - \gamma \log \frac{1}{m} \sum_{j=1}^m e^{\frac{\boldsymbol{\beta_j} - \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y_j})}{\gamma}} \right] p_{\boldsymbol{\theta}}(\boldsymbol{x})$$

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^m} \mathbb{E}_{p_{\boldsymbol{\theta}}}[h(\boldsymbol{\beta}, \boldsymbol{x})]$$

$$\boldsymbol{f}^\star = (\boldsymbol{b}^\star)^{\boldsymbol{D},\gamma} = \boldsymbol{x} \mapsto -\gamma \log \frac{1}{m} \sum_{i=1}^m e^{\frac{\boldsymbol{b}_j^\star - \boldsymbol{D}^p(\boldsymbol{y_j}, \boldsymbol{x})}{\gamma}}$$

In a continuous observation setting

$$W_\gamma(p_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\text{data}}) = \max_{\boldsymbol{f},\boldsymbol{b}} \int_\Omega \boldsymbol{f}\, dp_{\boldsymbol{\theta}} + \boldsymbol{b}^T \mathbf{1}_m - \gamma \langle e^{\boldsymbol{f}/\gamma},\ K e^{\boldsymbol{b}/\gamma} \rangle_{p_{\boldsymbol{\theta}}}$$

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^m} \int_{\boldsymbol{x}} \left[\sum_{i=1}^m \boldsymbol{\beta_j}/m - \gamma \log \frac{1}{m} \sum_{j=1}^m e^{\frac{\boldsymbol{\beta_j} - \boldsymbol{D}^p(\boldsymbol{x},\boldsymbol{y_j})}{\gamma}} \right] p_{\boldsymbol{\theta}}(\boldsymbol{x})$$

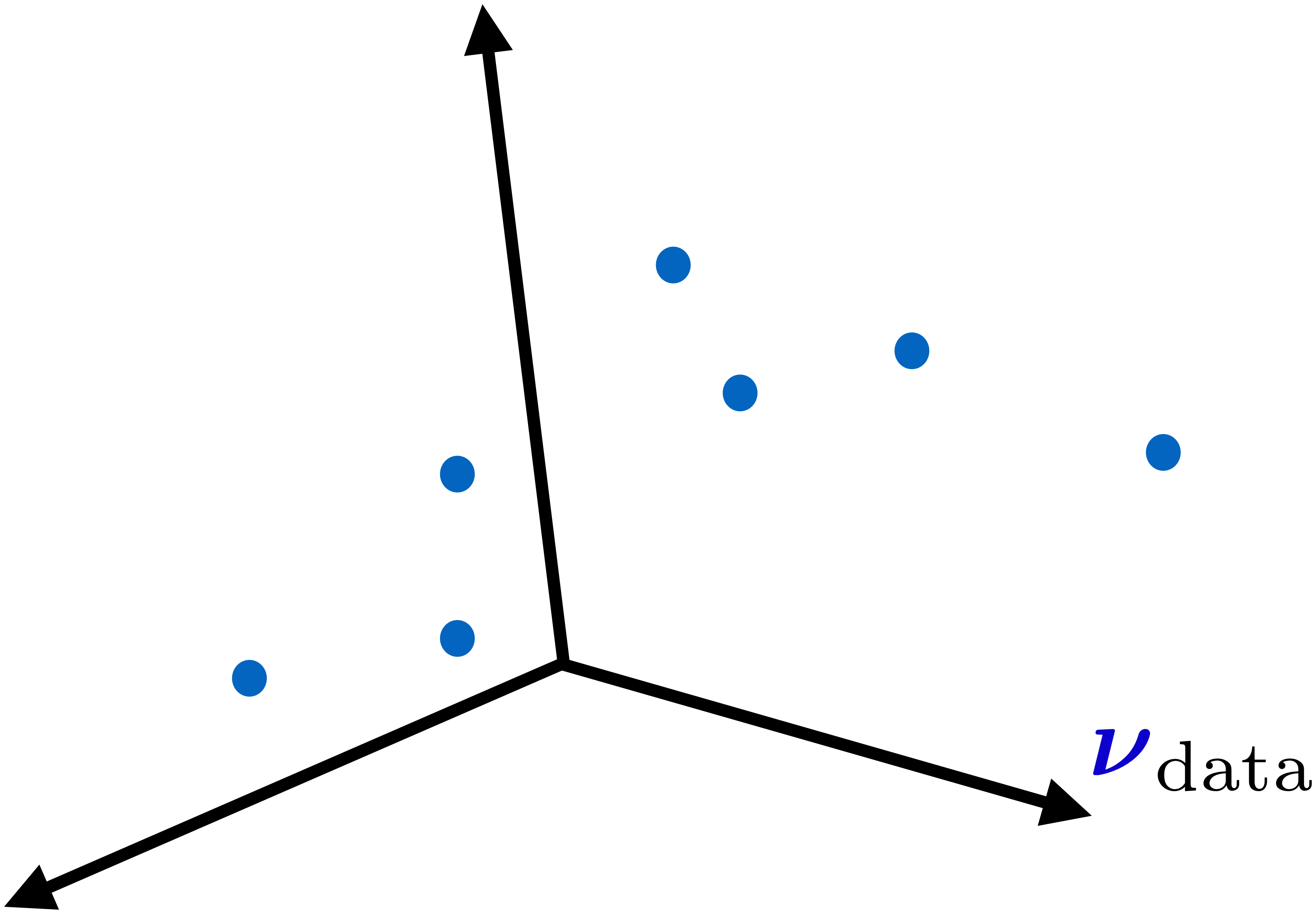
$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^m} \mathbb{E}_{p_{\boldsymbol{\theta}}}[h(\boldsymbol{\beta}, \boldsymbol{x})]$$

$$\boldsymbol{f}^\star = (\boldsymbol{b}^\star)^{\boldsymbol{D},\gamma} = \boldsymbol{x} \mapsto -\gamma \log \frac{1}{m} \sum_{i=1}^m e^{\frac{\boldsymbol{b}_j^\star - \boldsymbol{D}^p(\boldsymbol{y_j},\boldsymbol{x})}{\gamma}}$$

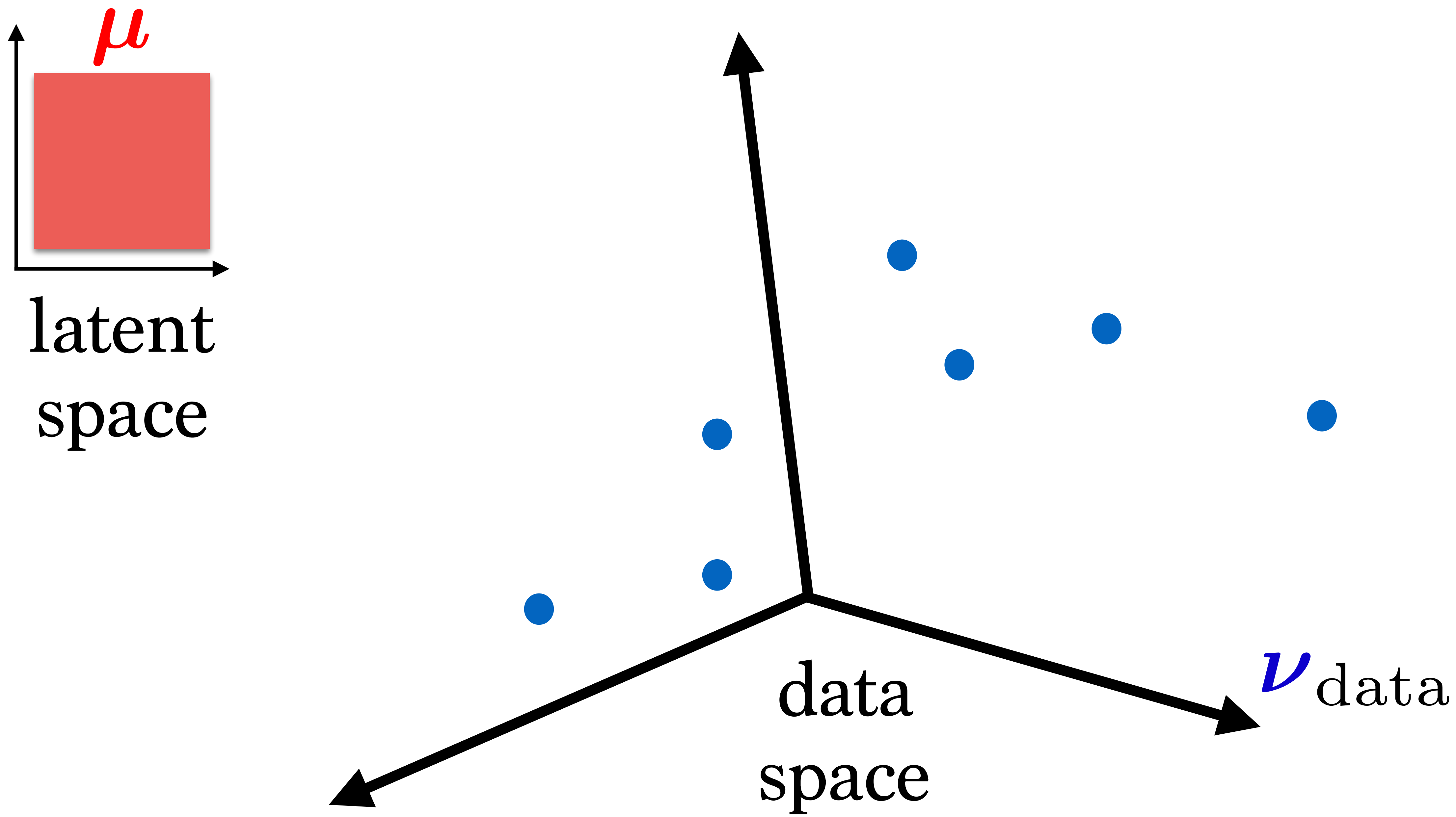
$$\nabla_\theta W_\gamma = \left(\frac{\partial p_\theta}{\partial \theta}\right)^* \boldsymbol{f}^\star$$

[GCPB'16]

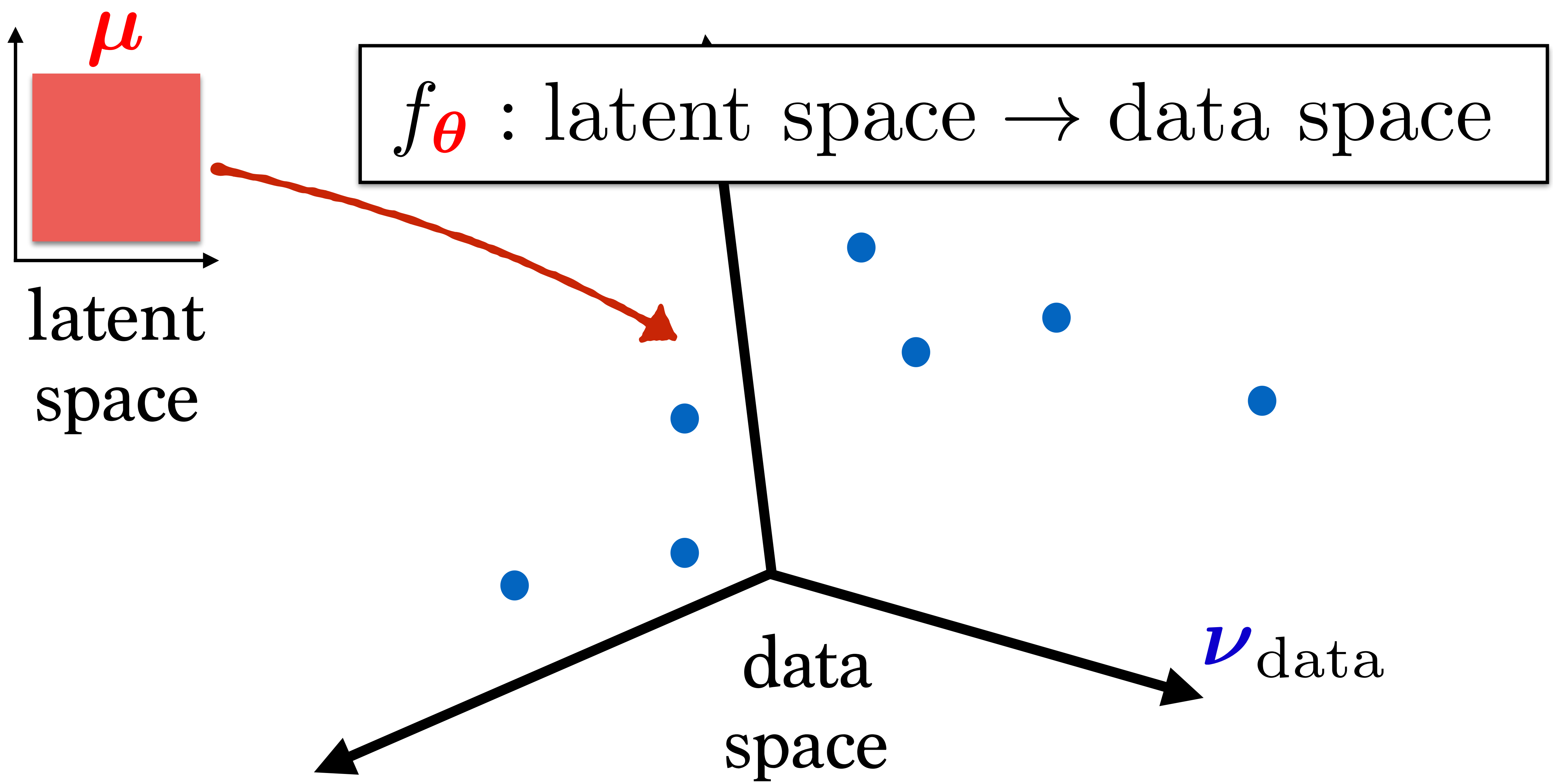
In a generative model setting



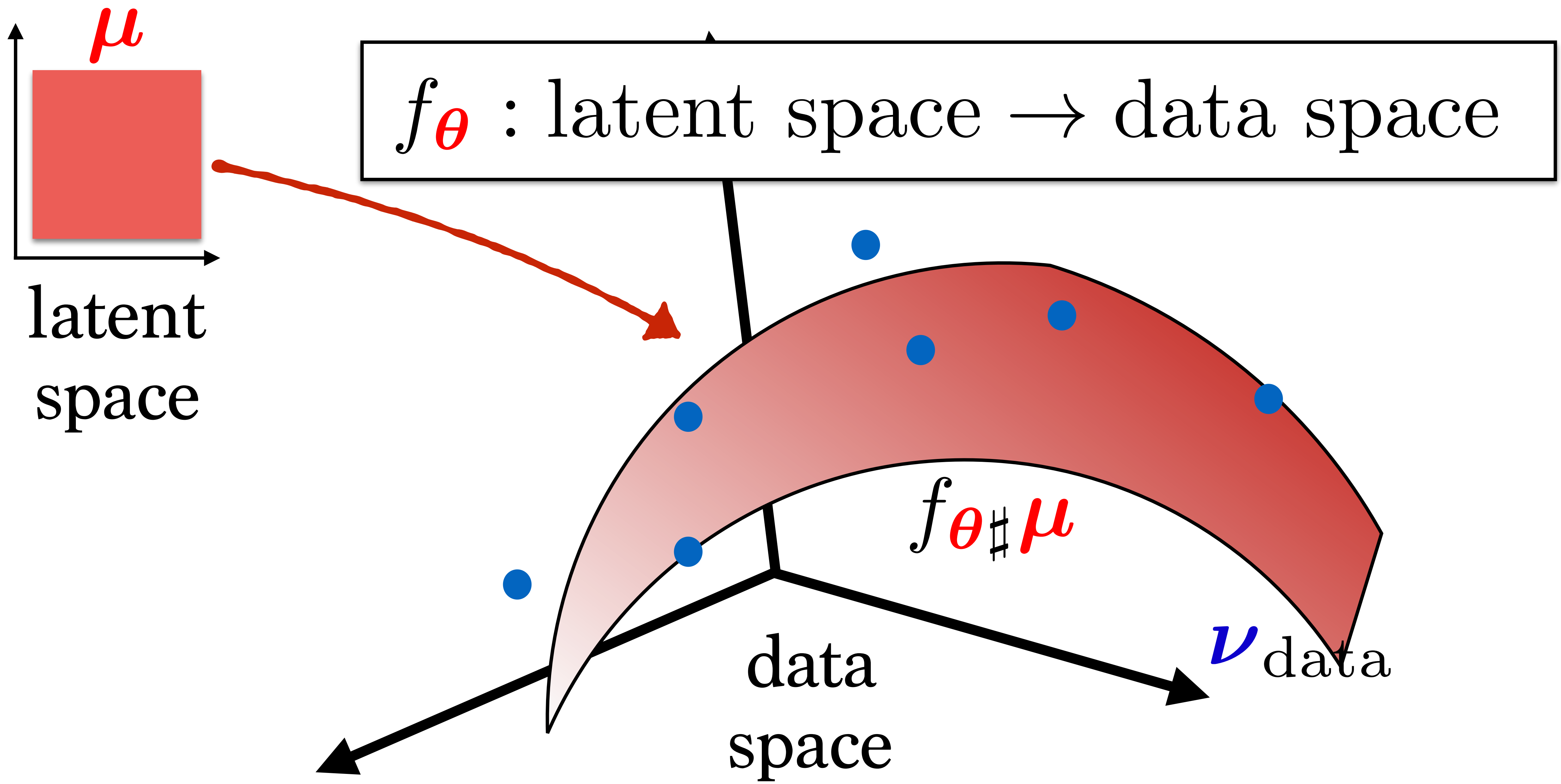
In a generative model setting



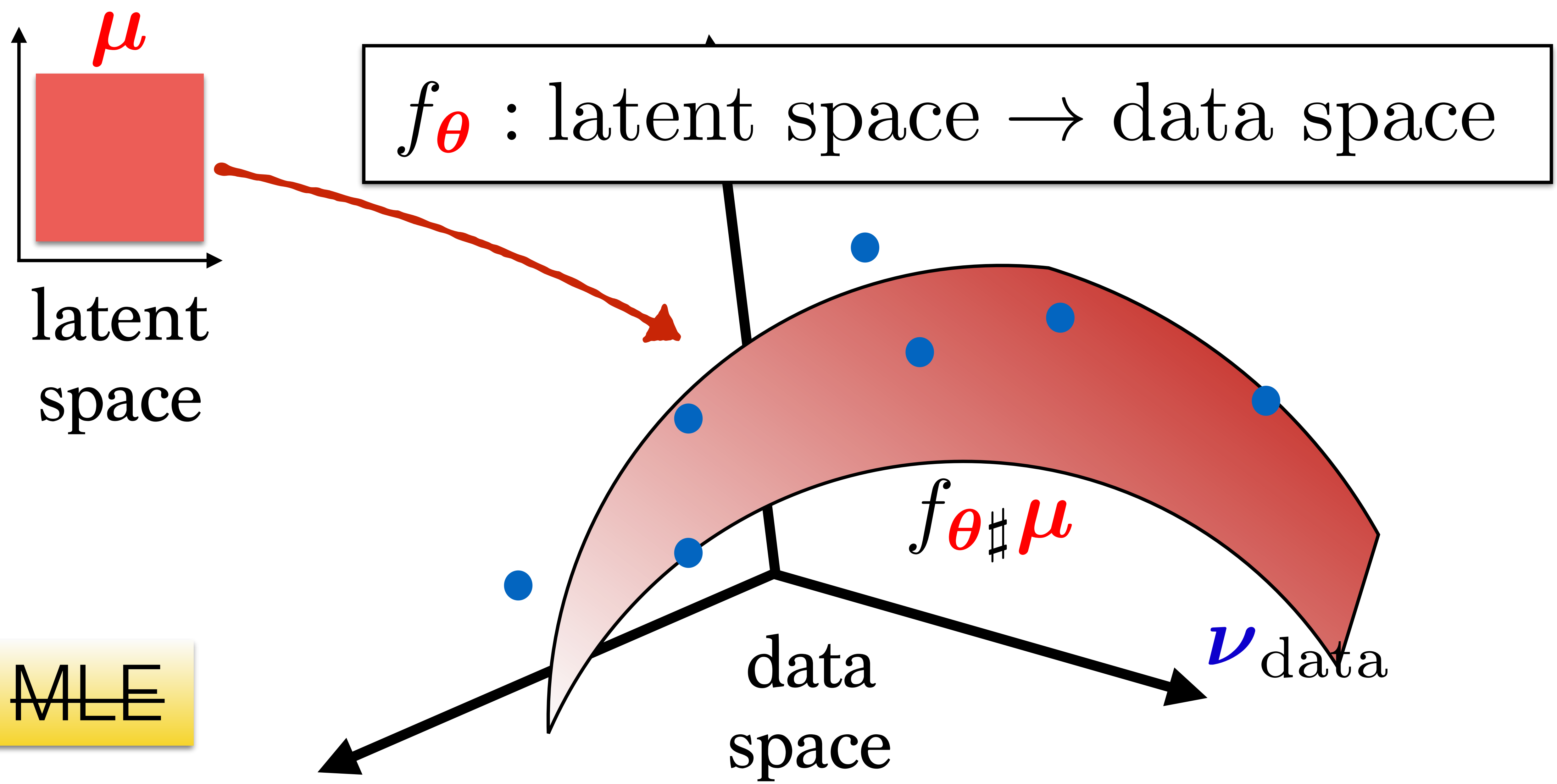
In a generative model setting



In a generative model setting



In a generative model setting



~~MLE~~

$$\min_{\boldsymbol{\theta}\in\Theta} W(\boldsymbol{\nu}_{\text{data}}, f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu})$$

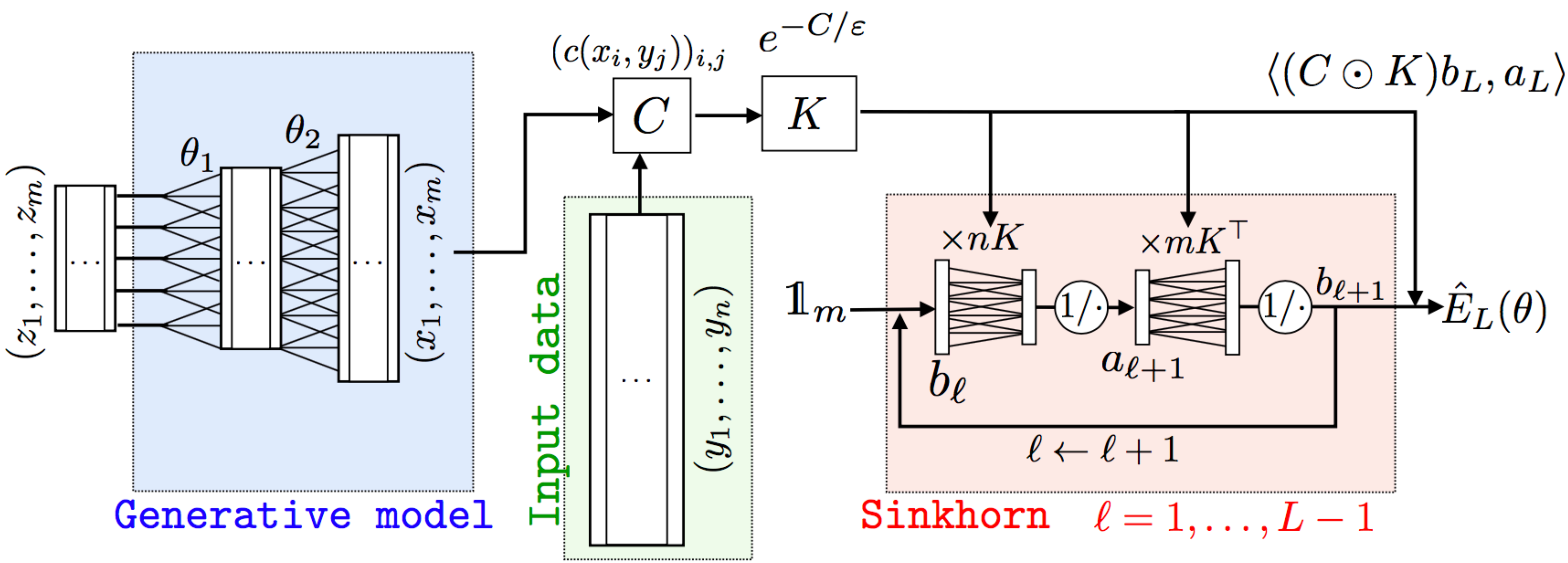
GM-MKE
W-GAN

[ACB'17]

[BGTSS'17]

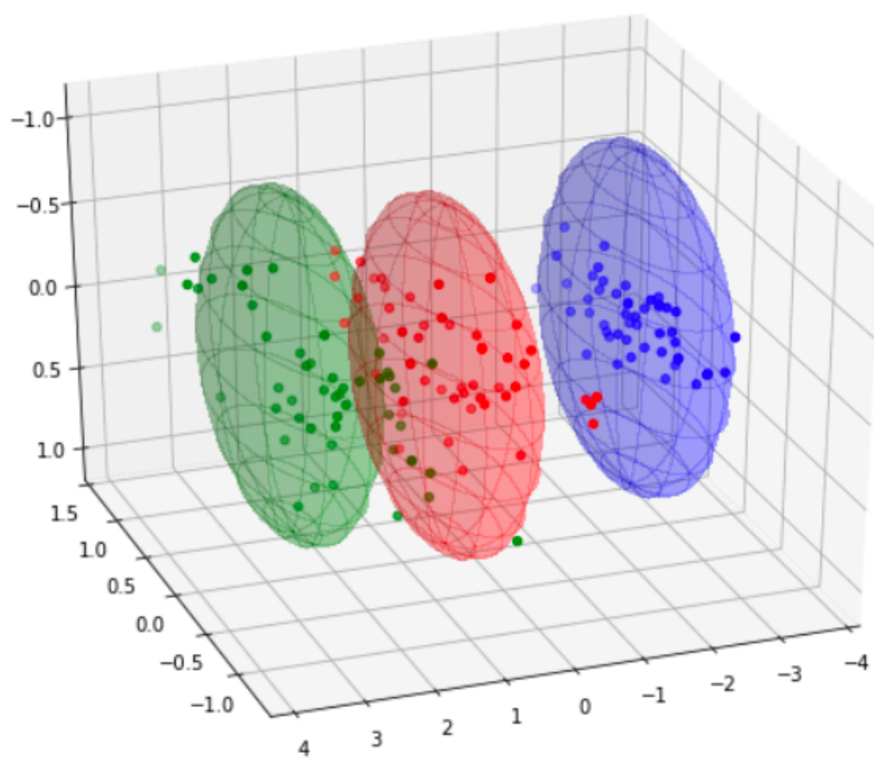
Our algorithmic proposal

Approximate regularized W loss by W_L.

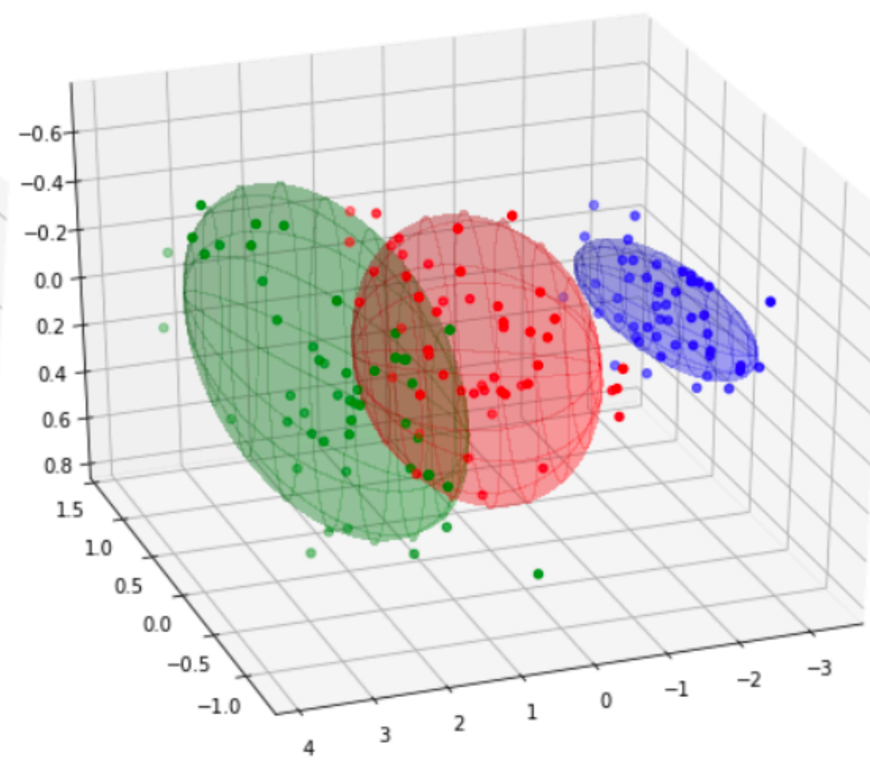


Example: Fitting Ellipses

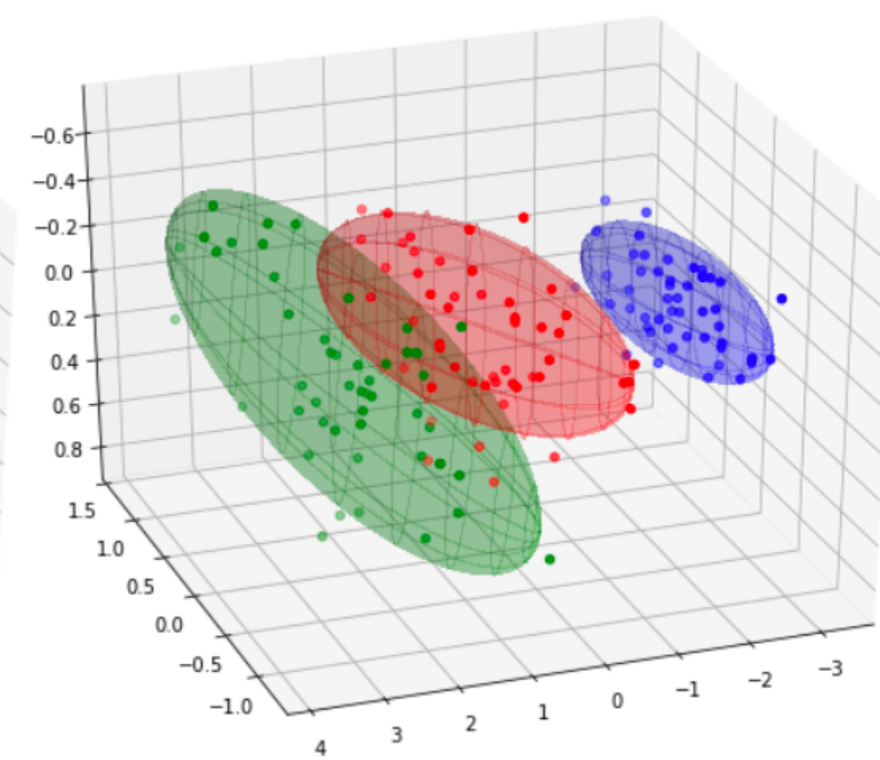
- k-means problem can be seen as a MKE when the model = atomic measures with k atoms.
- We generalize by estimating uniform ellipsoid measures that approximate clouds of points.



(a) Initialization (unit balls, kmeans centers)



(b) After 3 gradient steps



(c) At convergence (15 steps)

Example: MNIST, Learning $f_{\boldsymbol{\theta}}$