

Sesión III:

Aplicaciones de Transporte Óptimo y su formulación dinámica

Daniel López-Castaño
Johns Hopkins University

Geometría en Ciencias de Datos
Universidad Nacional de Colombia
Agosto 18-21

Sinkhorn Distances: Lightspeed Computation of Optimal Transport

Marco Cuturi
Graduate School of Informatics, Kyoto University
mcuturi@i.kyoto-u.ac.jp

Abstract

Optimal transport distances are a fundamental family of distances for probability measures and histograms of features. Despite their appealing theoretical properties, excellent performance in retrieval tasks and intuitive formulation, their computation involves the resolution of a linear program whose cost can quickly become prohibitive whenever the size of the support of these measures or the histograms’ dimension exceeds a few hundred. We propose in this work a new family of optimal transport distances that look at transport problems from a maximum-entropy perspective. We smooth the classic optimal transport problem with an entropic regularization term, and show that the resulting optimum is also a distance which can be computed through Sinkhorn’s matrix scaling algorithm at a speed that is several orders of magnitude faster than that of transport solvers. We also show that this regularized distance improves upon classic optimal transport distances on the MNIST classification problem.

1 Introduction

Choosing a suitable distance to compare probabilities is a key problem in statistical machine learning. When little is known on the probability space on which these probabilities are supported, various information divergences with minimalistic assumptions have been proposed to play that part, among which the Hellinger, χ_2 , total variation or Kullback-Leibler divergences. When the probability space is a metric space, optimal transport distances (Villani, 2009, §6), *a.k.a.* earth mover’s (EMD) in computer vision (Rubner et al., 1997), define a more powerful geometry to compare probabilities.

This power comes, however, with a heavy computational price tag. No matter what the algorithm employed – network simplex or interior point methods – the cost of computing optimal transport distances scales at least in $O(d^3 \log(d))$ when comparing two histograms of dimension d or two point clouds each of size d in a general metric space (Pele and Werman, 2009, §2.1).

Foundations and Trends® in
Machine Learning
11:5-6

Computational Optimal Transport With Applications to Data Science

Gabriel Peyré and Marco Cuturi

now

the essence of knowledge



Q Search

Examples

Getting Started

Tutorials

API

ott.geometry

ott.problems

ott.solvers

ott.initializers

ott.neural

ott.experimental

ott.tools



downloads 952k build failing docs passing coverage 87%

Optimal Transport Tools (OTT)

Introduction

OTT is a [JAX](#) package that bundles a few utilities to compute, and differentiate as needed, optimal transport (OT) problems, taken in a fairly wide sense. For instance, **OTT** can compute the [Wasserstein distance](#) (or [Gromov-Wasserstein distance](#)) between weighted point clouds (or more generally, between two probability measures) in a wide variety of scenarios, but also estimate a [Monge map](#), [Wasserstein barycenter](#), or even solve more general tasks such as differentiable approximations to ranking or clustering.

To achieve this, **OTT** rests on two families of tools:

- the first family consists in *discrete* solvers computing transport between two families of point clouds or histograms using e.g. the [Sinkhorn algorithm](#) [Cuturi, 2013] or low-rank solvers [Scetbon et al., 2019], with further extensions to more advanced scenarios such as the [Gromov-Wasserstein problem](#) [Cuturi, 2011, Peyré et al., 2016];
- the second family consists in *continuous* solvers, whose goal is to output, given two point clouds of n samples, a *function* that is an approximate [Monge map](#), a [transport map](#) that can map each sample to the second. Such functions can be recovered using directly tools above, or more generally, using [entropic map](#) approximations. Such maps can also be parameterized as neural architectures, either as MLP or as gradients of [input convex neural network](#) [Amos et al., 2017], trained with advanced optimization approaches [Amos, 2023, Korotin et al., 2021, Makkuva et al., 2020]. Such functions can also be parameterized as Neural ODEs, where to an input source point is associated the end-representation of an integral with a time-parameterized velocity field. Such velocity fields can be parameterized as neural networks and learned from data, using the framework of flow matching [Albergo et al., 2023], augmented with OT couplings for priors [de Lathauwer et al., 2023].

Wassersteinization

[wos-ur-stahyn-ahy-sey-shuh-n]

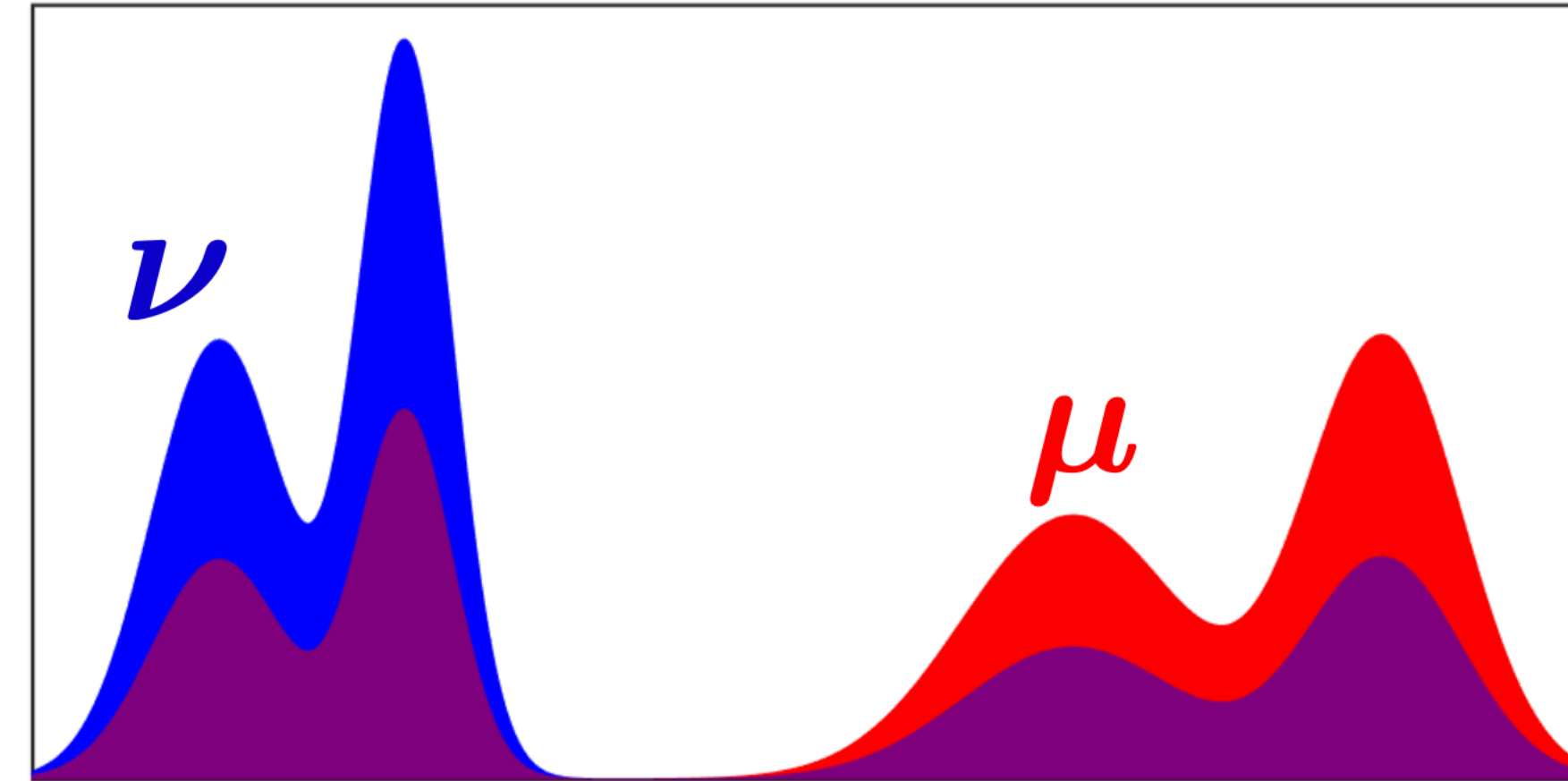
noun

Introduction of optimal transport into an optimization or learning problem.

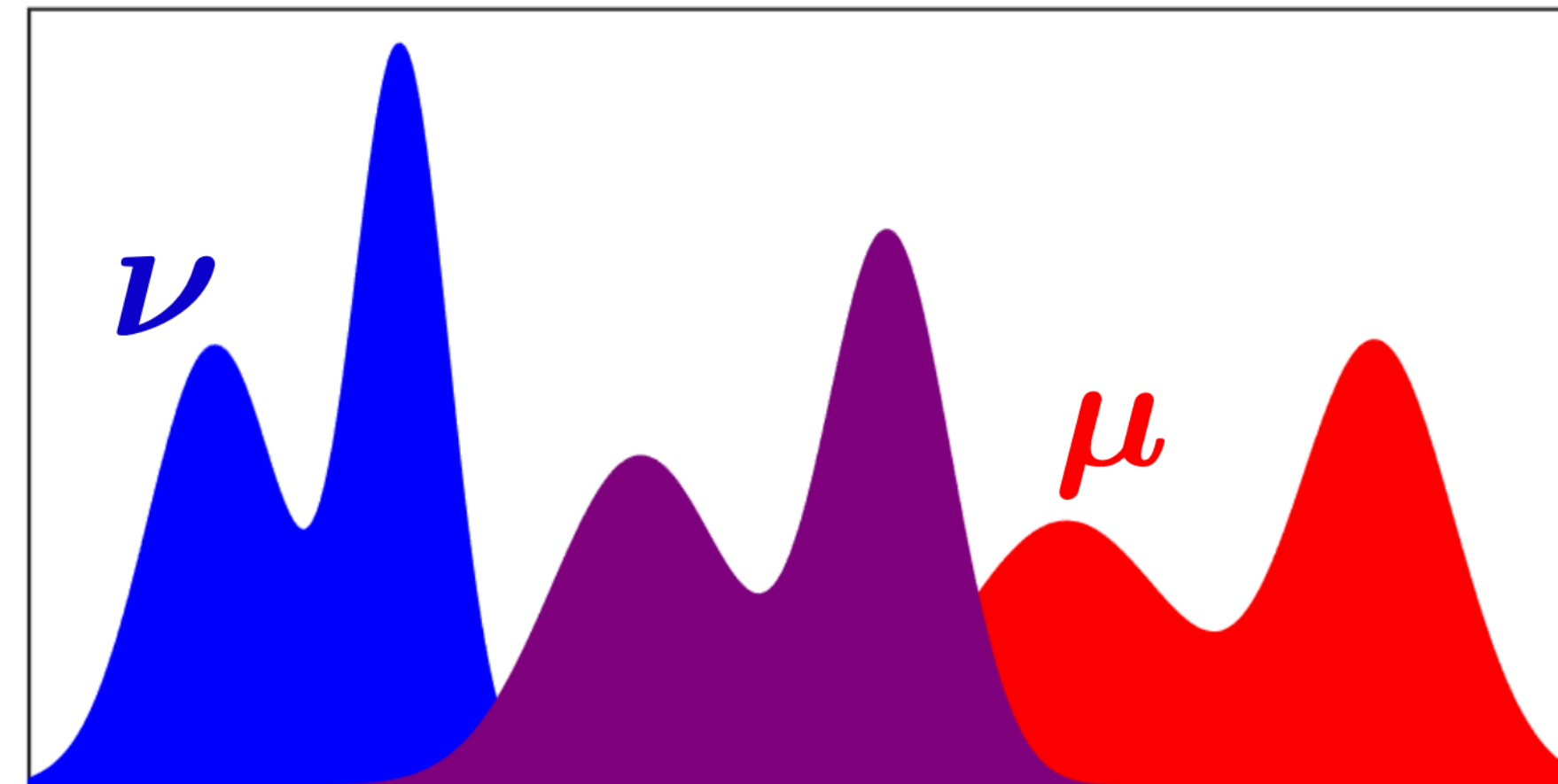
cf. least-squarification, L_1 -ification, deep-netification, kernelization

Wasserstein - Baricentros

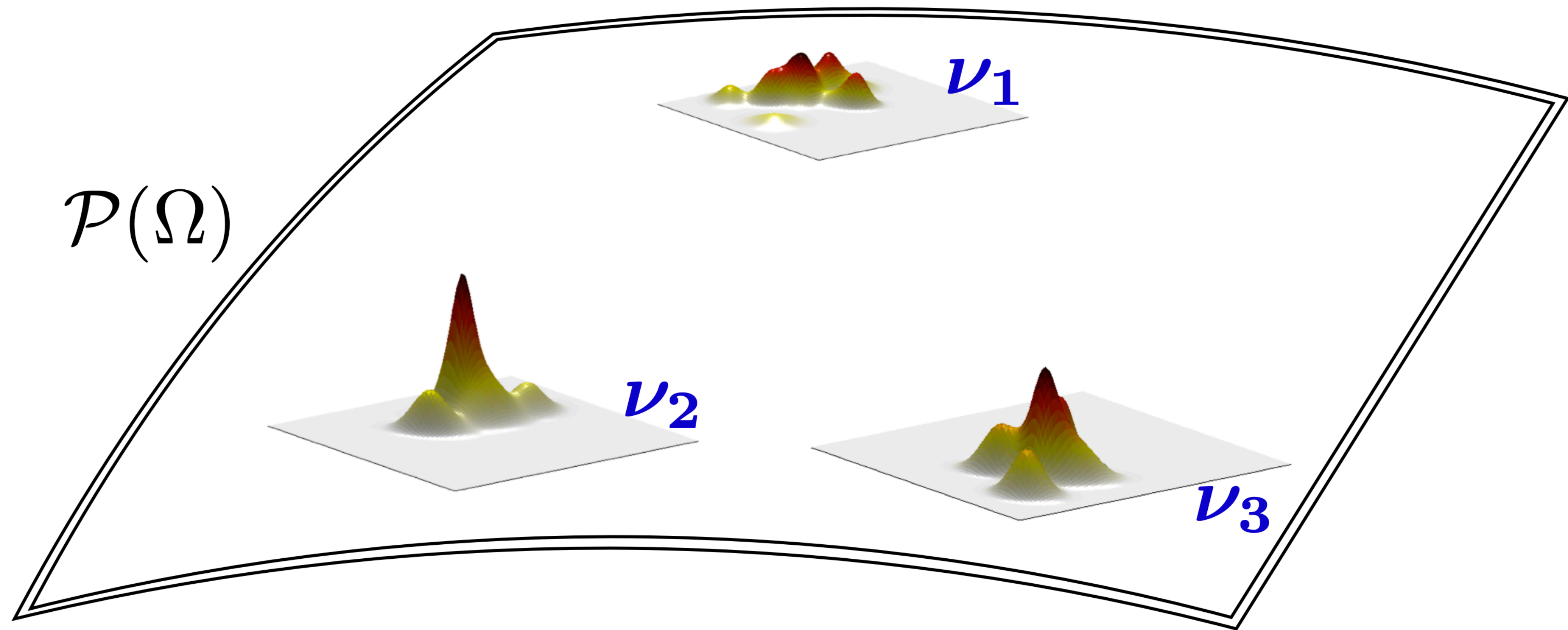
promedio en L_1



promedio en W

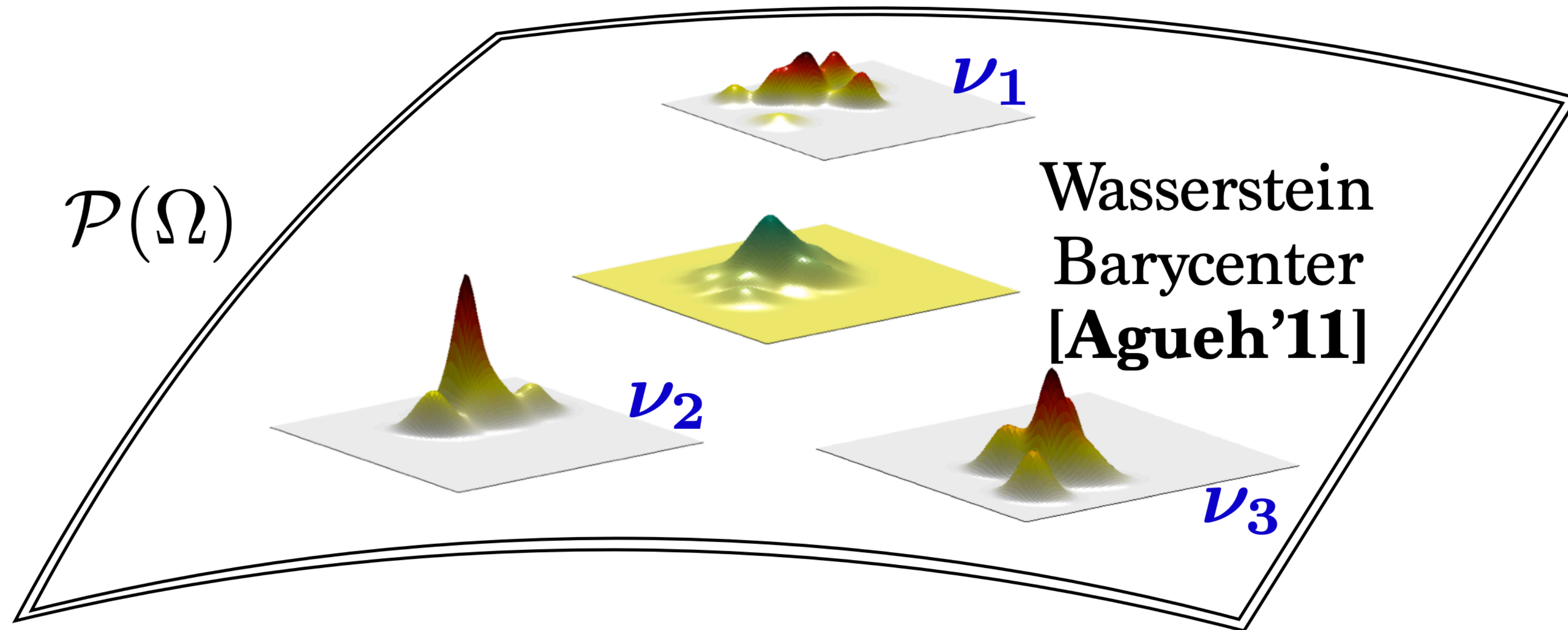


Wasserstein - Baricentros



Wasserstein - Baricentros

$$\min_{\mu \in \mathcal{P}(\Omega)} \sum_{i=1}^N \lambda_i W_p^p(\mu, \nu_i)$$



The Earth Mover's Distance



$$\text{dist}(I_1, I_2) = W_1(\mu, \nu)$$

[Rubner'98]

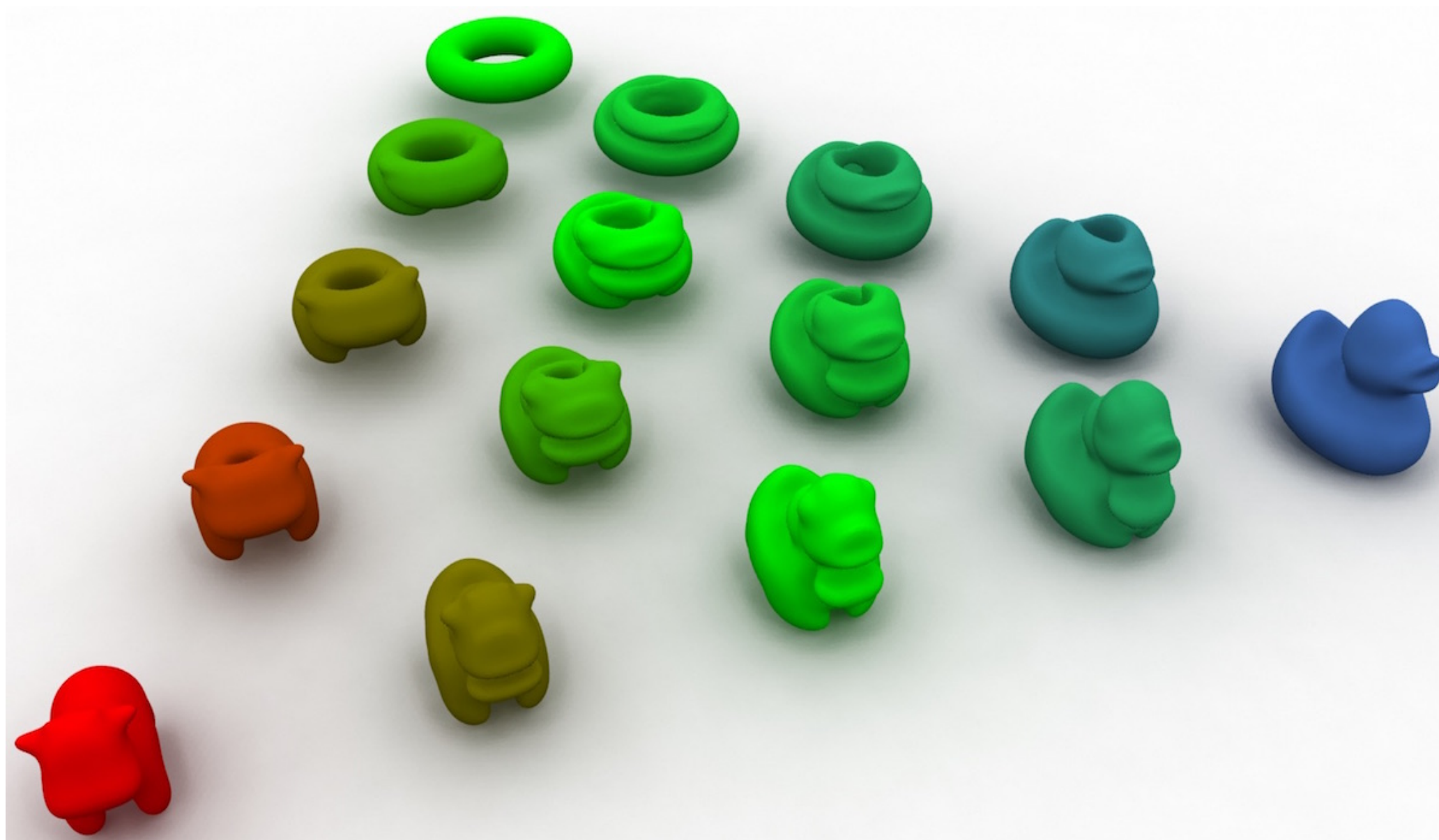
The Word Mover's Distance



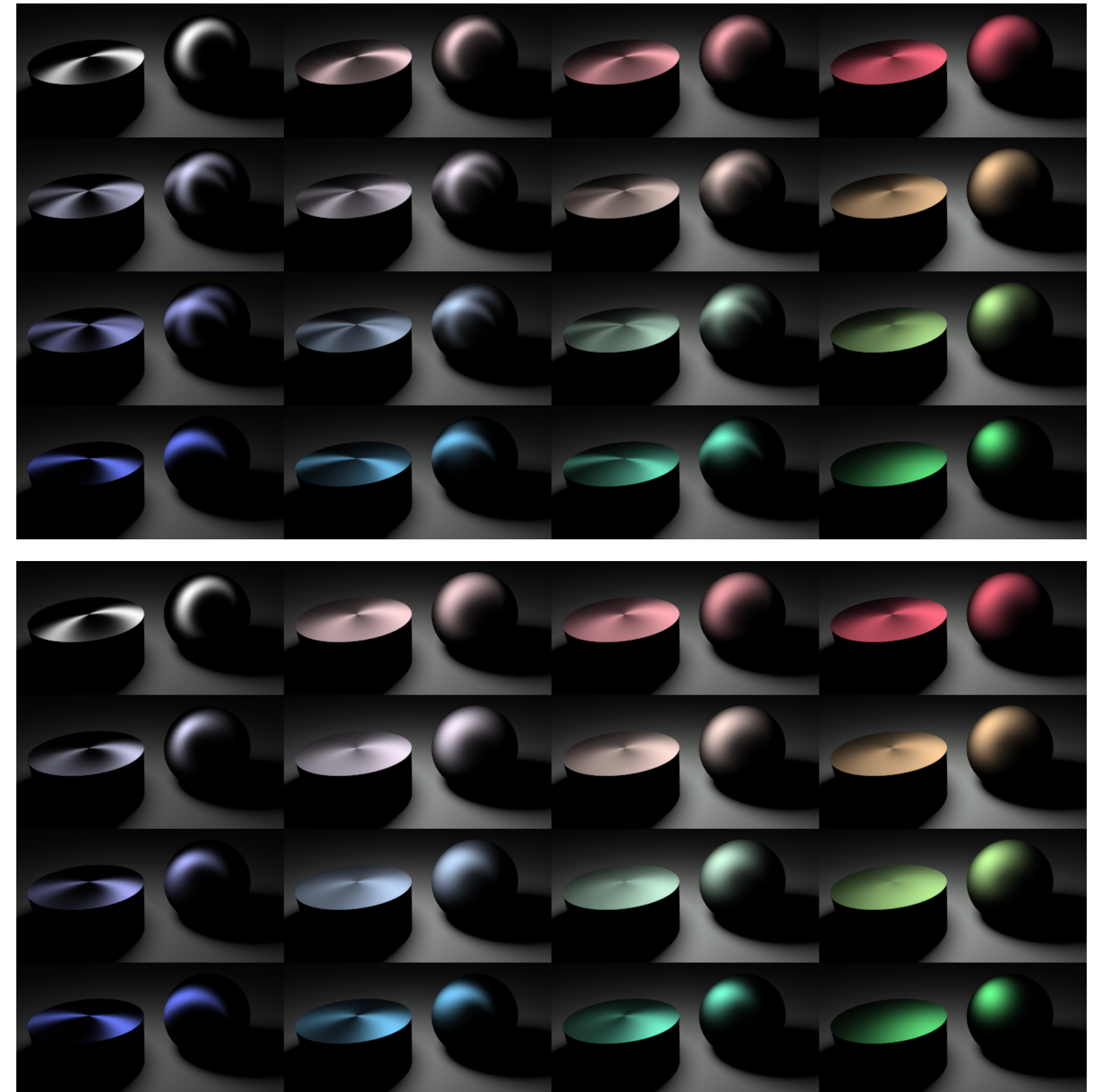
$$\text{dist}(D_1, D_2) = W_2(\mu, \nu)$$

[Kusner'15]

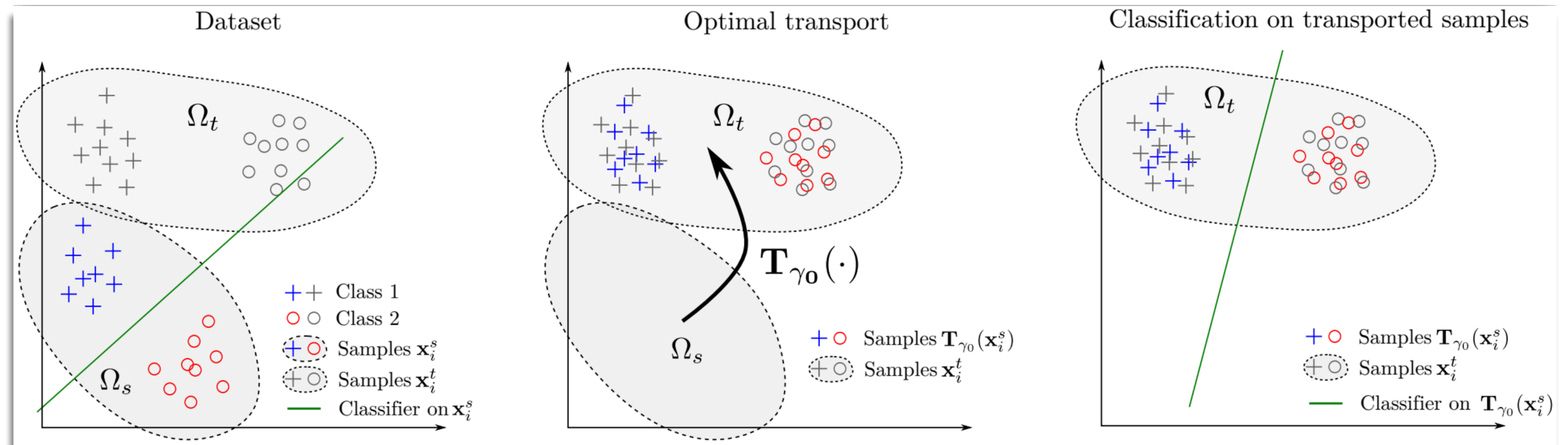
Wasserstein - Baricentros



[Solomon'15]



Adaptación de dominios



1. **Estimamos** el mapa transporte
2. **Transportamos** las muestras con labels al nuevo dominio
3. **Entrenamos** el clasificador en muestras transportadas con labels

Density fitting

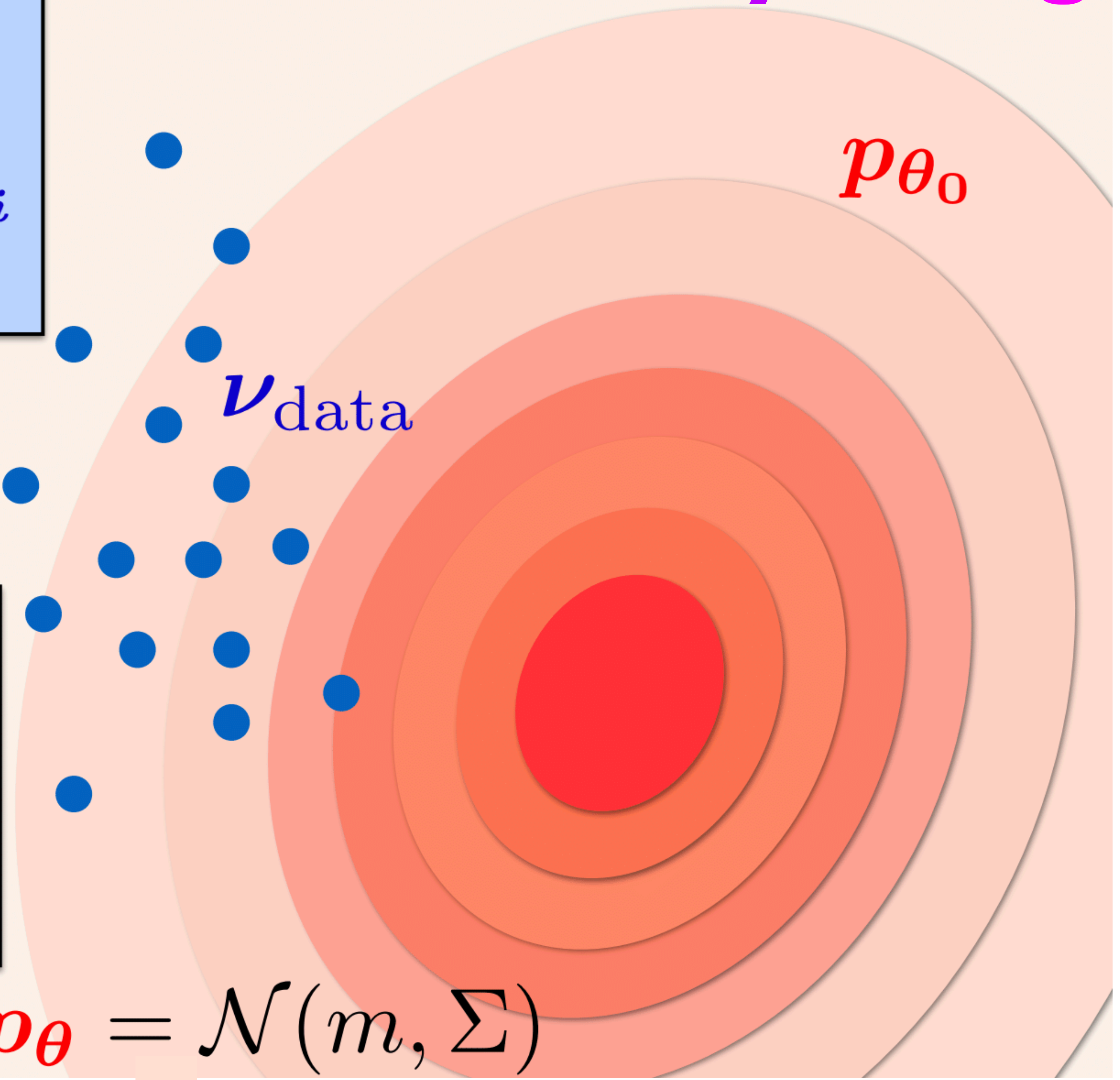
We collect data

$$\nu_{\text{data}} = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$$

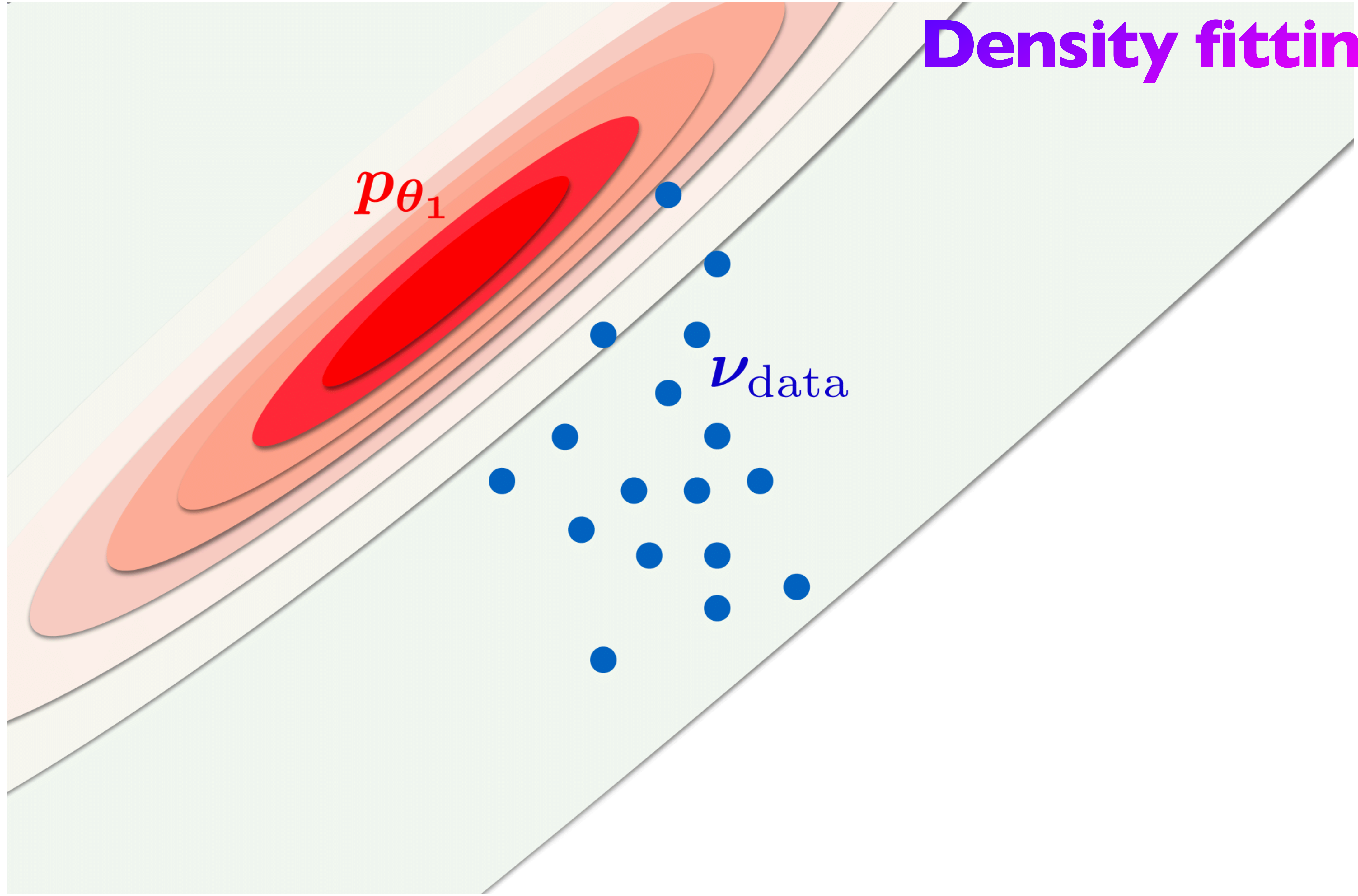
We fit a parametric family of densities

$$\{p_{\theta}, \theta \in \Theta\}$$

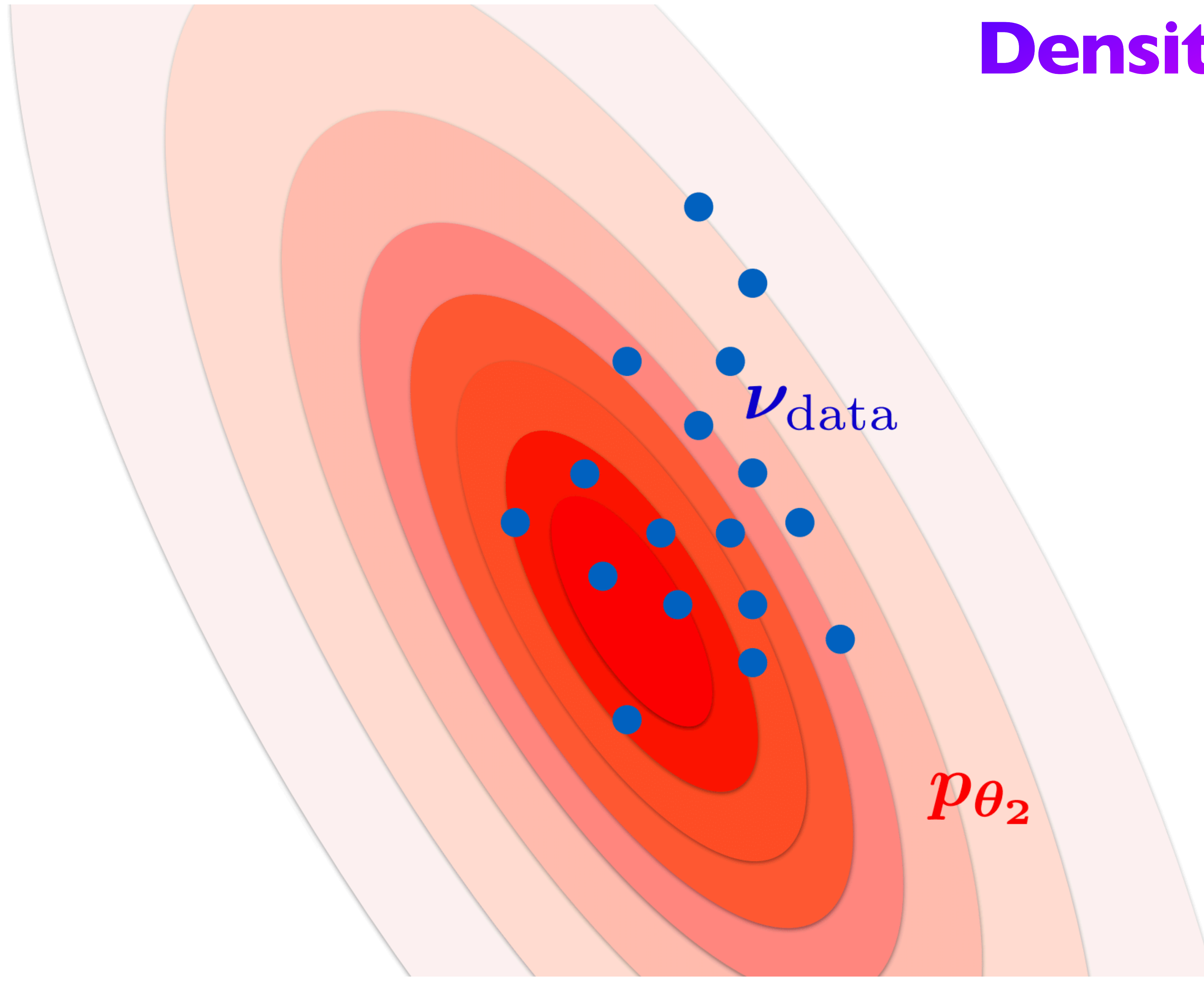
e.g. $\theta = (m, \Sigma); p_{\theta} = \mathcal{N}(m, \Sigma)$



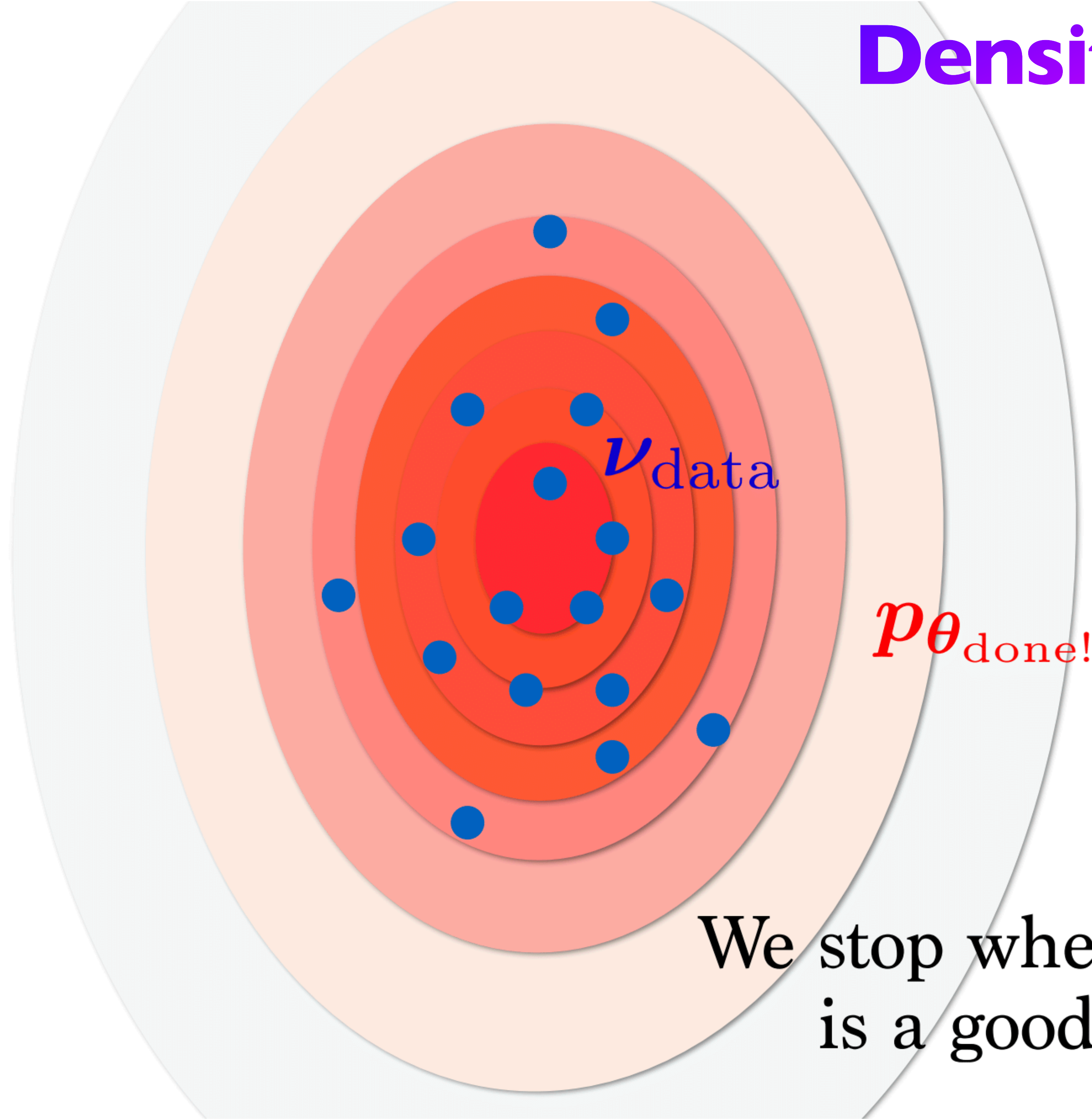
Density fitting



Density fitting

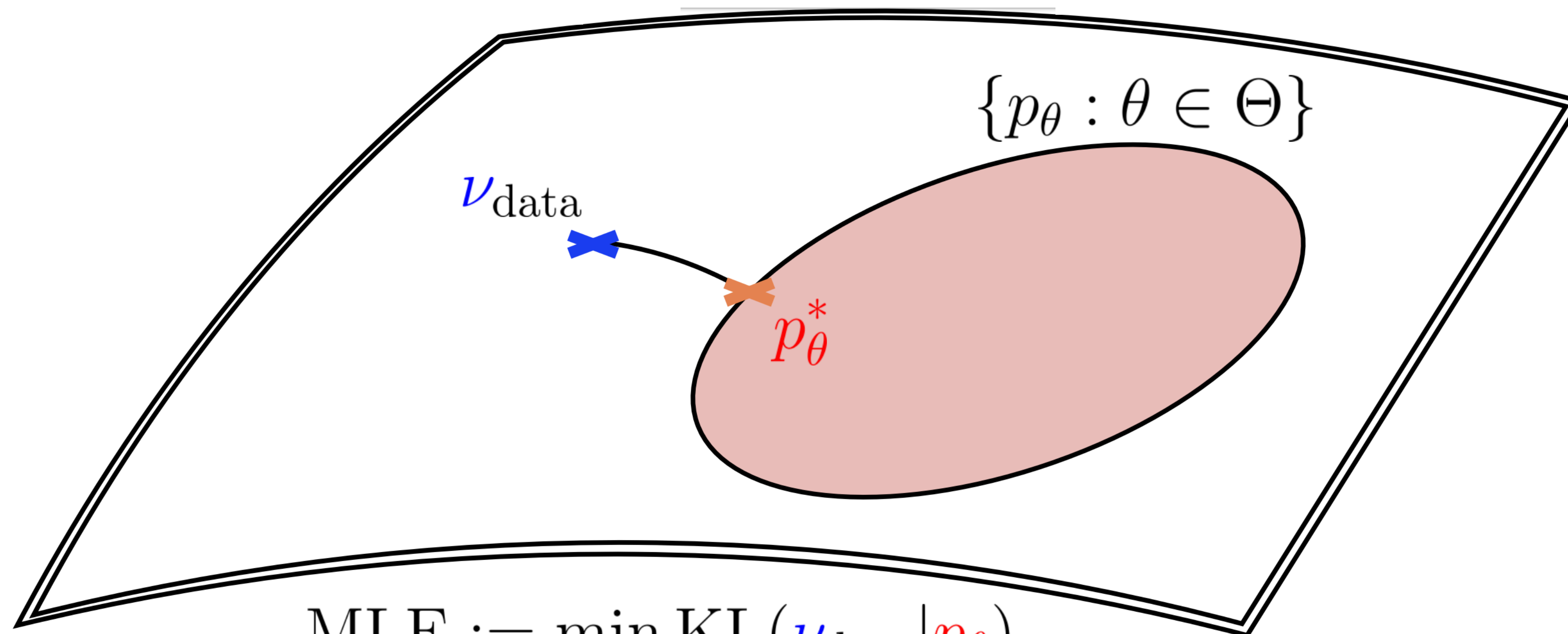


Density fitting



We stop when there
is a good fit.

Density fitting



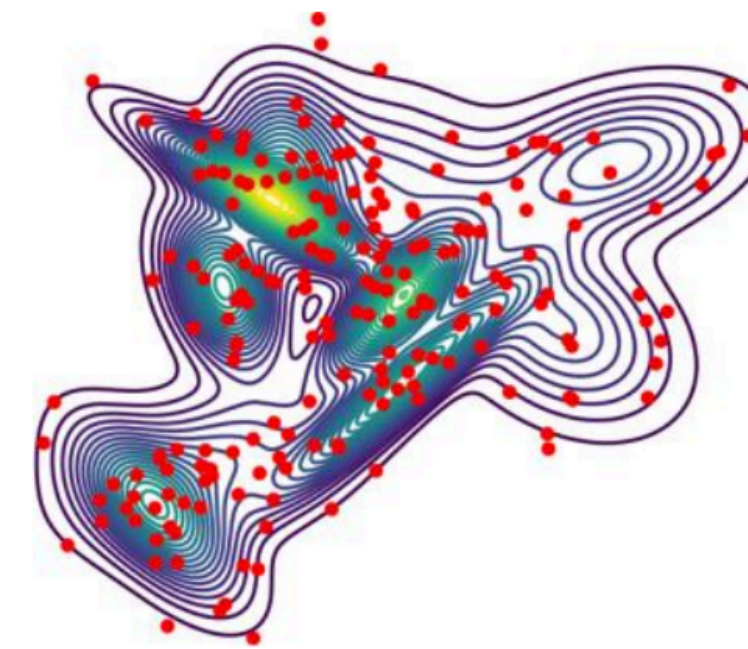
$$\text{MLE} := \min_{\theta \in \Theta} \text{KL}(\nu_{\text{data}} | p_\theta)$$

$$\longrightarrow \text{MKE} := \min_{\theta \in \Theta} \mathcal{W}_2(\nu_{\text{data}}, p_\theta)$$

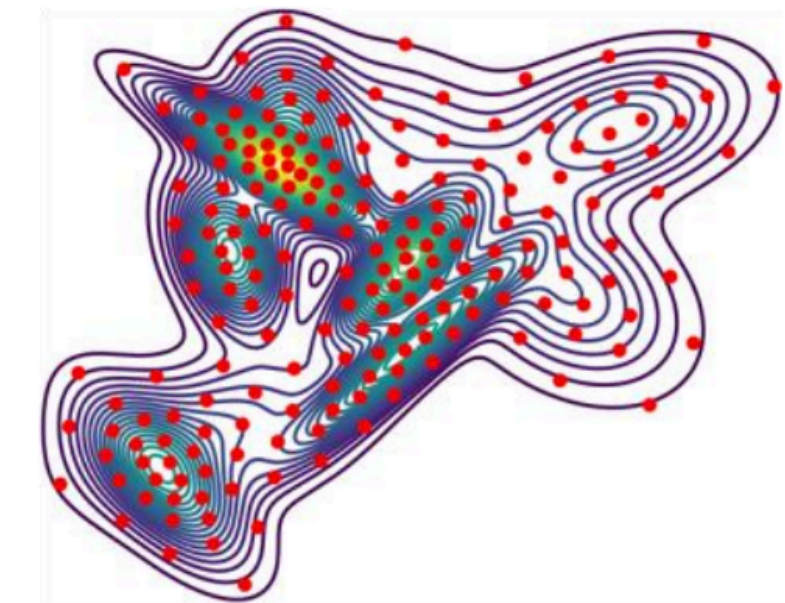
[Bassetti 2006]

Minimum Kantorovich Estimator

Blue Noise and Distribution Approximators

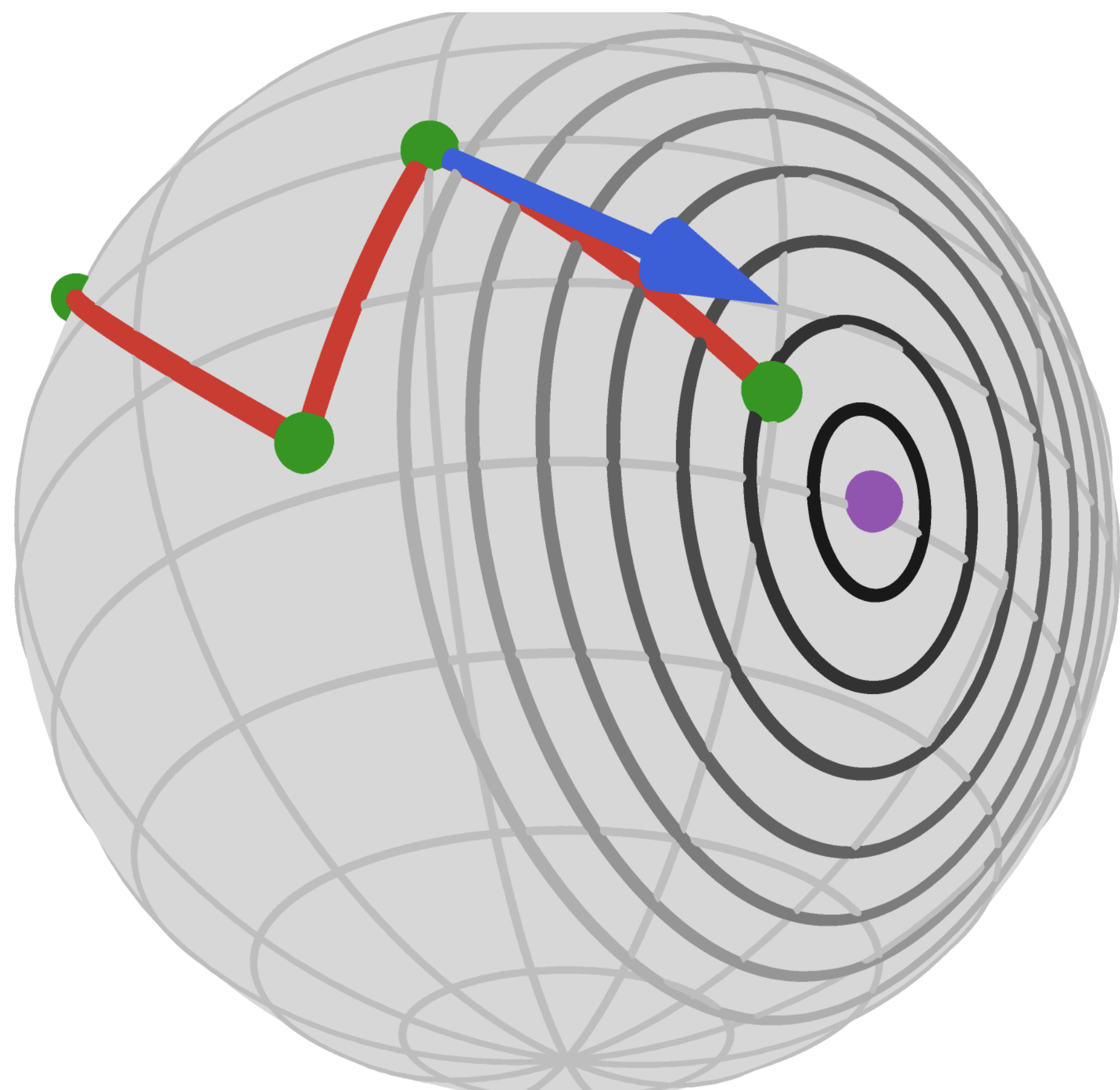


Uniform samples



Transport

$$\min_{x_1, \dots, x_n} W_2^2 \left(\mu, \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \right)$$



Gracias!



SESIÓN III.A KANTOROVICH DUAL

1. Problema Relajado de Kantorovich

$$\inf_{\gamma} \left\{ \int_{X \times Y} c(x,y) d\gamma \mid \gamma \geq 0 \text{ y } \gamma \in \Pi(\mu, \nu) \text{ donde} \right.$$

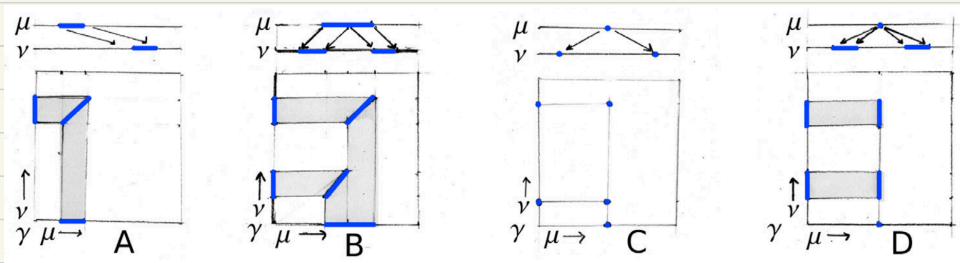
$$\Pi(\mu, \nu) = \{ \gamma \in \mathcal{P}(X \times Y) \mid P_X \# \gamma = \mu ; P_Y \# \gamma = \nu \}$$

OBSERVACION: Si $(Id \times T) \# \mu$ es en $\Pi(\mu, \nu)$, entonces T "pushes" μ en ν .

Demostración. $(Id \times T) \in \Pi(\mu, \nu)$ entonces $(P_Y) \# (Id \times T) \# \mu = \nu$
lo que es equivalente a $(P_Y \circ (Id \times T)) \# \mu = \nu$ y finalmente
 $T \# \mu = \nu$. ■

Note que si el plan de transporte γ tiene la forma $\gamma = (Id \times T) \# \mu$

$$\min_{X \times Y} \left\{ \int_{X \times Y} c(x,y) d((Id \times T) \# \mu) \right\} = \min_X \left\{ \int_X c(x, T(x)) d\mu \right\} \quad (\text{Monge})$$



Q: Preguntas naturales:

- En qué condiciones se garantiza que existe un plan de transporte?
- Cuando este admite un mapa de transporte?

→ Método directo de cálculo de variaciones

SKETCH: $C \subseteq X$, (X, d) espacio métrico. Meta: probar $\min_{x \in C} F(x)$ existe.

Paso 0: Probar que $\inf_{x \in C} F(x) < +\infty$
Problema es tratable.

Paso 1: Tomar una sucesión minimizadora

i.e. escoger $\{x_n\}_{n=1}^{\infty} \subseteq C$ t.q. $\lim_{n \rightarrow \infty} F(x_n) = \inf_C F(x)$

Paso 2: Probar que C es compacto

i.e. $\exists$ subsecisión $\{x_{n_k}\}_{k=1}^{\infty} \subseteq C$ t.q. $x_{n_k} \xrightarrow{k \rightarrow \infty} x_{\infty} \in C$

Paso 3: Probar que F es semicontinua por debajo,

$\therefore \inf F(x) = \lim_{n \rightarrow \infty} F(x_n) = \lim_{k \rightarrow \infty} F(x_{n_k}) = \dots = \liminf_{k \rightarrow \infty} F(x_{n_k}) \geq F(x_{\infty})$

$\therefore x_{\infty}$ es un minimizador de F en C

El reto es encontrar la topología (métrica) adecuada para C :

- suficientemente débil para ganar compacidad
- suficientemente fuerte para que F sea semicontinua por debajo.

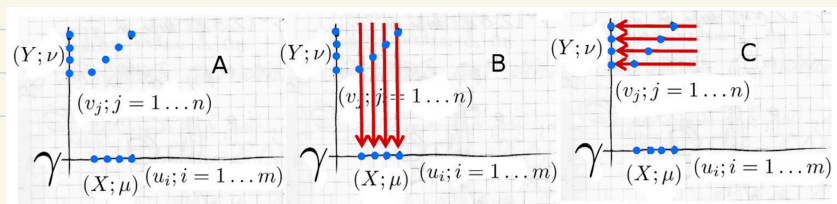
TEOREMA: Dados $(X, d_X), (Y, d_Y)$ espacios polacos, $\forall C: X \times Y \rightarrow \mathbb{R} \cup \{\infty\}$ semicontinua por debajo y acotada por debajo, $\Pi(\mu, \nu)$ es compacto en la topología angosta, y existe $\delta_{\mu, \nu} \in \Pi(\mu, \nu)$ que satisface

$$\delta_{\mu, \nu} = \argmin_{\Pi(\mu, \nu)} \int C(x, y) d\delta(x, y)$$

2. Construyendo el dual de Kantorovich en el escenario discreto

Reemplazamos medidas μ y ν por vectores $U = (u_i)_{i=1, \dots, m}$ $V = (v_j)_{j=1, \dots, n}$
 $\therefore$ plan de transporte es una matriz $\delta = [\delta_{ij}]$

$$\min_{\delta} \langle C, \delta \rangle \quad \text{sujeto a} \quad \begin{cases} P_X \delta = U \\ P_Y \delta = V \\ \delta_{ij} \geq 0 \quad \forall i, j \end{cases}$$



Introducámonos arbitrariamente

$$L(\varphi, \psi) = \langle C, x \rangle - \langle \varphi, P_x x - u \rangle - \langle \psi, P_y x - v \rangle$$

$\uparrow$ $\uparrow$
 $\varphi \in \mathbb{R}^m$ $\psi \in \mathbb{R}^n$

Note que

$$\sup_{\varphi, \psi} [L(\varphi, \psi)] = \begin{cases} \langle C, x \rangle & \text{si } P_x x = u \text{ y } P_y x = v \\ +\infty & \text{o/w.} \end{cases}$$

Ahora considere

$$\inf_{x \geq 0} \left[\sup_{\varphi, \psi} [L(\varphi, \psi)] \right] = \inf_{\substack{x \geq 0 \\ P_x x = u \\ P_y x = v}} \langle C, x \rangle$$

Reemplazando L,

$$\begin{aligned} \inf_{x \geq 0} \left[\sup_{\varphi, \psi} \left(\langle C, x \rangle - \langle \varphi, P_x x - u \rangle - \langle \psi, P_y x - v \rangle \right) \right] &= \sup_{\varphi, \psi} \left[\inf_{x \geq 0} \left(\langle C, x \rangle - \langle \varphi, P_x x - u \rangle - \langle \psi, P_y x - v \rangle \right) \right] \\ &= \sup_{\varphi, \psi} \left[\inf_{x \geq 0} \left(\langle x, C - P_x^t \varphi - P_y^t \psi \rangle + \langle \varphi, u \rangle + \langle \psi, v \rangle \right) \right] \\ &= \sup_{\substack{\varphi, \psi \\ P_x^t \varphi + P_y^t \psi \leq C}} [\langle \varphi, u \rangle + \langle \psi, v \rangle] \end{aligned}$$

generalización de minimax (Von Neumann)

$$\therefore \sup_{\varphi, \psi} [\langle \varphi, u \rangle + \langle \psi, v \rangle] \text{ sujeto a } \varphi_i + \psi_j \leq C_{ij} \quad \forall i, j.$$

OBSERVACION: • El problema primal dependia de $n+m$ variables, el problema dual depende de $n+m$ parámetros.

3. El dual de Kantorovich en el escenario continuo

$$(DK) \quad \sup_{\varphi, \psi} \left[\int_X \varphi d\mu + \int_Y \psi d\nu \right] \text{ sujeto a } \begin{cases} \varphi(x) + \psi(y) \leq C(x, y) \\ \forall (x, y) \in X \times Y \end{cases}$$

El dual de Kantorovich es aún materia de estudio porque el conjunto de restricciones no es compacto. Podremos hablar más de su estructura en la noción de C -transformada.

DEFINICIÓN: • Para cualquier función f en Y con valores en $\mathbb{R} \cup \{-\infty\}$ y no idénticamente $-\infty$, definimos su C -transformada como

$$f^C(x) = \inf_{y \in Y} [C(x, y) - f(y)] \quad x \in X;$$

- Si ϕ es una función t.g. existe f con $\phi = f^C$; entonces ϕ se dice C -cóncava;
- $\Psi^C(X)$ denota el conjunto de las funciones C -cóncavas en X .

OBSERVACIÓN: Si el par (ϕ, ψ) es admisible para (DK) entonces (ψ^C, ψ) también lo es.

Demostración. $\left\{ \begin{array}{l} \forall x, y \in X \times Y, \phi(x) + \psi(y) \leq C(x, y) \\ \psi^C(x) = \inf_{y \in Y} [C(x, y) - \psi(y)] \end{array} \right.$

$$\begin{aligned} \psi^C(x) + \psi(y) &= \inf_{y \in Y} [C(x, y) - \psi(y)] + \psi(y) \\ &\leq C(x, y) - \psi(y) + \psi(y) = C(x, y) \end{aligned}$$

OBSERVACIÓN: Si el par (ϕ, ψ) es admisible para (DK) , uno obtiene un mejor par reemplazando ϕ por ψ^C .

Demostración. $\left\{ \begin{array}{l} \forall y \in Y \quad \psi^C(x) = \inf_{y \in Y} [C(x, y) - \psi(y)] \\ \phi(x) \leq C(x, y) - \psi(y) \end{array} \right\} \Rightarrow \psi^C(x) \leq \phi(x)$

¿Por qué hacemos esto?

$$\inf_K(K) = \sup_{\phi \in \Psi^C(X)} \int_X \phi d\mu + \int_Y \psi^C d\nu = \sup_{\psi \in \Psi(Y)} \int_X \psi^C d\mu + \int_Y \psi d\nu$$

y el conjunto de funciones admisibles $\Psi^C(X)$ es compacto!!
↳ existencia

4. Del dual de Kantorovich al mapa de transporte óptimo:

C-SUPERDIFERENCIAL:

Queremos, cuando exista, el mapa $T: X \rightarrow Y$ asociado al plan de transporte $(Id_X \times T) \# \mu$ que minimiza el funcional del problema de Monge

TEOREMA: Sea ϕ una función C^1 -convexa. Para todo $(x, y) \in \mathcal{D}^c \phi$, tenemos

$$\nabla \phi(x) - \nabla_x C(x, y) = 0,$$

donde $\mathcal{D}^c \phi = \{(x, y) : \phi(z) \leq \phi(x) + C(x, y) - C(x, y), \forall z \in X\}$
denota el C -superdiferencial de ϕ .

Por definición de la C -transformada, si $(x, y) \in \mathcal{D}^c \phi$, entonces $\phi^c(y) = C(x, y) - \phi(x)$. Entonces, el C -superdiferencial se puede caracterizar por el conjunto de puntos $(x, y) \in X \times Y$ tal que $\phi(x) + \phi^c(y) = C(x, y)$.

Q: ¿Cuál es la relación entre el C -superdiferencial y el mapa transporte?

Considere un punto $(x, y) \in \mathcal{D}^c \phi$, entonces, $\forall z \in X$
 $C(x, y) - \phi(x) \leq C(z, y) - \phi(z)$.

Calculamos la derivada en x en la dirección w (arbitraria)

$$\lim_{t \rightarrow 0^+} \frac{\phi(x+tw) - \phi(x)}{t} \leq \lim_{t \rightarrow 0^+} \frac{C(x+tw, y) - C(x, y)}{t}$$

$$\therefore \nabla \phi(x) \cdot w \leq \nabla_x C(x, y) \cdot w.$$

El mismo razonamiento en la dirección $-w$ nos da

$$\nabla \phi(x) \cdot w = \nabla_x C(x, y) \cdot w \quad \forall w \in X.$$

Q: ¿Qué pasa en el caso de L^2 ? $C(x, y) = \frac{1}{2} \|x - y\|^2$

$\forall (x, y) \in \mathcal{D}^c \phi$, $\nabla \phi(x) + (y - x) = 0 \therefore$ cuando el mapa de transporte T existe, está dado por

$$T(x) = x - \nabla \phi(x) = \nabla \left(\underbrace{\frac{\|x\|^2}{2}}_{\text{convexa}} - \phi(x) \right)$$

OBSERVACIÓN: Si $C(x, y) = \frac{1}{2} \|x - y\|^2$ y $\phi \in \Psi^c(X)$ entonces

$\bar{\phi}(x) = \frac{\|x\|^2}{2} - \phi(x)$ es una función convexa. (equivalencia si $X = Y = \mathbb{R}^n$)

Demostración. $\Phi(x) = \psi^c(x)$

$$= \inf_y \left[\frac{\|x - y\|^2}{2} - \psi(y) \right]$$

$$= \inf_y \left[\frac{\|x\|^2}{2} - x \cdot y + \frac{\|y\|^2}{2} - \psi(y) \right]$$

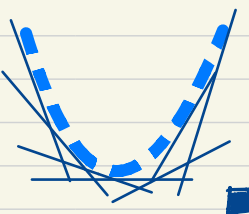
Entonces,

$$-\bar{\Phi}(x) = \Phi(x) - \frac{\|x\|^2}{2} = \inf_y \left[-x \cdot y + \left(\frac{\|y\|^2}{2} - \psi(y) \right) \right]$$

o, equivalentemente

$$\bar{\Phi}(x) = \sup_y \left[x \cdot y - \left(\frac{\|y\|^2}{2} - \psi(y) \right) \right]$$

La función $x \mapsto x \cdot y - \left(\frac{\|y\|^2}{2} - \psi(y) \right)$ es afín en x , entonces el gráfico de $\bar{\Phi}$ es el "upper envelope" de una familia de hiperplanos, por lo tanto $\bar{\Phi}$ es una función convexa.



OBSERVACIÓN: Considere las trayectorias de dos partículas parametrizadas por $t \in [0, 1]$, $t \mapsto (1-t)x_1 + tT(x_1)$
 $t \mapsto (1-t)x_2 + tT(x_2)$.

Si $x_1 \neq x_2$ y $0 < t < 1$ entonces no hay colisión entre partículas

Demostración. Por contradicción, suponga que hay una colisión, entonces existe $t \in (0, 1)$ y $x_1 \neq x_2$ tal que

$$(1-t)x_1 + tT(x_1) = (1-t)x_2 + tT(x_2)$$

$$(1-t)x_1 + t\nabla\bar{\Phi}(x_1) = (1-t)x_2 + t\nabla\bar{\Phi}(x_2)$$

$$(1-t)(x_1 - x_2) + t(\nabla\bar{\Phi}(x_1) - \nabla\bar{\Phi}(x_2)) = 0$$

$\therefore$

$$(1-t)\|x_1 - x_2\|^2 + t(\nabla\bar{\Phi}(x_1) - \nabla\bar{\Phi}(x_2)) \cdot (x_1 - x_2) = 0$$

($\rightarrow \leftarrow$)

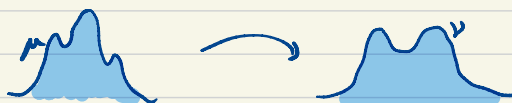
SESION III.B: WASSERSTEIN GRADIENT FLOWS

1. Nociones básicas de transporte óptimo

DEF: Dadas dos medidas de probabilidad μ, ν sobre $\mathbb{R}^d$ with Segundo momento finito (notado $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$), la distancia Wasserstein-2 entre μ y ν está definida como

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int \|x - y\|^2 \gamma(dx, dy)$$

donde $\Pi(\mu, \nu)$ es el conjunto de couplings de μ y ν .



HECHOS:

- Para $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, un coupling óptimo γ^* siempre existe (no necesariamente es único).
- $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ es un espacio métrico
- $W_1(\mu_n, \mu) \rightarrow 0 \iff \mu_n \rightarrow \mu$ débilmente y $\int \|x\|^2 d\mu_n \rightarrow \int \|x\|^2 d\mu$
- $x \mapsto \delta_x$ es un embedding isométrico (i.e. $W_2(\delta_x, \delta_y) = \|x - y\|$)

TEOREMA FUNDAMENTAL DEL TRANSPORTE ÓPTIMO

Suponga que μ tiene una densidad con respecto a la medida de Lebesgue (notado $\mu \in \mathcal{P}_{2,ac}$). Entonces, el plan de transporte óptimo es inducido por un mapa de transporte:

$$W_2^2(\mu, \nu) = \int \|x - y\|^2 \gamma^*(dx, dy) = \int \|x - T^*(x)\|^2 \mu(dx)$$

$$\nu = T^* \# \mu \iff \text{si } x \sim \mu \text{ entonces } T^*(x) \sim \nu$$

Más aún, T^* está caracterizado como el único $(\mu - c.t.p)$ gradiente de una función convexa que pushforward μ en ν . Llamamos T^* el mapa de Brenier

EJEMPLOS:

$$\bullet T \# \delta_x = \delta_{T(x)}$$

• Si $\mu = N(m_1, \Sigma_1)$ y $\nu = N(m_2, \Sigma_2)$ entonces

$$T^*(x) = m_2 + \Sigma_1^{-1/2} (\Sigma_1^{-1/2} \Sigma_2^{-1/2} \Sigma_1^{1/2})^T \Sigma_1^{-1/2} (x - m_1) \quad (\text{ejercicio})$$

2. Curvas de medidas:

Curva en $\mathbb{R}^d$



$\dot{x}$ es el vector tangente a la curva

curva en $\mathcal{P}_{2loc}(\mathbb{R}^d)$



¿qué es $\dot{\mu}_t$?

Piense en (x_t) como la trayectoria de una partícula

x_t = posición en el tiempo t

$\dot{x}_t$ = velocidad en el tiempo t

Piense en (μ_t) como una colección de partículas

$\mu_t(x)$ = densidad de masa en x en el tiempo t .

$v_t(x)$ = velocidad de x en el tiempo t .

v_t es el campo vectorial de velocidades.

perspectiva "LAGRANGIANA"

perspectiva "EULERIANA"

Junbando las dos perspectivas, piense ahora en

x_t como una variable aleatoria $x_t \sim \mu_t$

Entonces, $\dot{x}_t = v_t(x_t)$

PROPOSICIÓN

Si $x_0 \sim \mu_0$ y $\dot{x}_t = v_t(x_t)$ entonces $\mu_t = \text{leg}(x_t)$ y

satisface

$$\partial_t \mu_t + \text{div}(\mu_t v_t) = 0 \quad (\text{ecuación de continuidad})$$

Demostración. La forma débil del POE es

$$\partial_t \int \phi d\mu_t = \int \langle \nabla \phi, v_t \rangle d\mu_t$$

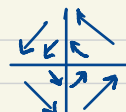
$$\int \varphi \partial_t \mu_t$$

Pero $\partial_t \int \varphi d\mu_t = \partial_t \mathbb{E}(\varphi(x_t)) = \mathbb{E}(\nabla \varphi(x_t), \dot{x}_t) = \mathbb{E}(\langle \nabla \varphi(x_t), v_t(x_t) \rangle)$
 $= \int \langle \nabla \varphi, v_t \rangle d\mu_t = - \int \varphi \operatorname{div}(\mu_t v_t) \quad \left(\text{integration by parts} \right)$

Q: Podemos pensar en v_t como el vector tangente a μ_t ?

No hay unicidad: $\mu_t = N(0, Id) \quad \forall t \in \mathbb{R}$

$v_t = 0$ vs. $v_t = \text{rotation}$



Criterio de selección: Escoger v_t que minimice la energía cinética

$$v_t = \operatorname{argmin}_{v: \mathbb{R}^d \rightarrow \mathbb{R}^d} \left\{ \frac{1}{2} \int \|v\|^2 d\mu_t : \operatorname{div}(\mu_t v_t) = -\partial_t \mu_t \right\} \quad (*)$$

Proposición: Existe un único vector v_t que satisface (*). Está caracterizado por " $v_t = \nabla \varphi_t$ " para algún $\varphi_t: \mathbb{R}^d \rightarrow \mathbb{R}$
 (Mas preciso: $v_t = \lim_{k \rightarrow \infty} \nabla \varphi_t^{(k)}$, donde $\varphi_t^{(k)} \in C_c^\infty(\mathbb{R}^d)$)

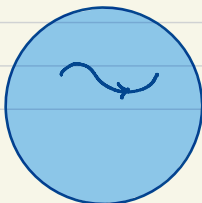
Demostración. Unicidad se sigue de (*) estrictamente convexo. Note que la restricción en (*) es

$$\int \langle \nabla \varphi, v_t \rangle d\mu_t = \partial_t \int \varphi d\mu_t, \quad \forall \varphi \in C_c^\infty(\mathbb{R}^d)$$

Si reemplazamos v_t por su proyección en $\overline{\{\nabla \varphi: \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu_t)}$ aún satisface la restricción pero tiene una norma menor.

Hodge decomposition.

3. Elementos de Geometría Riemanniana:



Una variedad Riemanniana M tiene asociada, para cada punto $p \in M$, un espacio vectorial — el espacio tangente $T_p M$ —, equipado con un producto interno $\langle \cdot, \cdot \rangle_p$.

El producto interno determina distancias:

$$d(p_0, p_1) := \inf \left\{ \int_0^1 \| \dot{p}_t \|^2 dt : \text{curvas } p_t: [0,1] \rightarrow M \text{ que conectan } p_0 \text{ y } p_1 \right\}$$

Un minimizador $(p_t)_{t \in [0,1]}$ de este problema variacional es una geodésica (const. speed) o camino más corto.

La aplicación exponencial $\exp_p: T_p M \rightarrow M$ mapea $v \in T_p M$ al punto sobre la geodésica que arranca en p con velocidad v por unidad de tiempo.

La aplicación logarítmica $\log: M \rightarrow T_p M$ es la inversa de la exponencial.

Dado $f: M \rightarrow \mathbb{R}$, el gradiente $\nabla f(p)$ en $p \in M$ es el elemento de $T_p M$ $\dagger q$ para todas las curvas $(p_t)_{t \in \mathbb{R}}$ con $p_0 = p$,

$$\partial_t f(p_t) \big|_{t=0} = \langle \nabla f(p), \dot{p}_0 \rangle_p.$$

4. Cálculo de Otto:

Definimos el espacio tangente

$$T_{\mu} \mathcal{P}_{2,ac}(\mathbb{R}^d) := \overline{\{ \nabla \psi : \psi \in C_c^\infty(\mathbb{R}^d) \}}^{L^2(\mu)}$$

equipado con la norma $L^2(\mu)$: $\| \nabla \psi \|^2 := \int \| \nabla \psi \|^2 d\mu$

Q: Cuáles son las geodésicas?

TEOREMA: La fórmula de Benamou-Brenier satisface:

$$W_2^2(\mu_0, \mu_1) = \inf \left\{ \int_0^1 \| v_t \|^2 dt : (q_t, v_t) \text{ resuelven la ecuación de continuidad} \right\}$$

La geodésica que conecta μ_0 y μ_1 se describe como

$$\mu_t = ((1-t) \text{id} + t \nabla \phi) \# \mu_0, \quad \nabla \phi = \text{mapa de Brenier}$$

Demostración. Sea (μ, ν) que solucionan la ecuación de continuidad. Entonces,

$$\text{Si } \dot{x}_t = V_t(x_t), \quad x_t \sim \mu_t \quad \text{Jensen ineq.}$$

$$\int_0^1 \|\dot{\mu}_t\|_{\mu_t}^2 dt = \int_0^1 \mathbb{E} \|\dot{x}_t\|^2 dt \geq \mathbb{E} \left\| \int_0^1 \dot{x}_t dt \right\|^2 = \mathbb{E} \|x_1 - x_0\|^2 \geq W_2^2(\mu_0, \mu_1) \quad \square$$

Hemos obtenido así una estructura Riemanniana asociada a la distancia Wasserstein-2.

$$\exp_{\mu}(\nabla \varphi) = (\text{id} + \nabla \varphi) \# \mu$$

$$\log_{\mu}(x) = \nabla \varphi_{\mu \rightarrow \nu} - \text{id}$$

Ahora, sea $\mathcal{F}: \mathcal{P}_{2ac}(\mathbb{R}^d) \rightarrow \mathbb{R}$ un funcional. ¿Qué es el gradiente?

DEFINICIÓN: La primera variación de $\mathcal{F}$ en μ es una función $\delta \mathcal{F}(\mu): \mathbb{R}^d \rightarrow \mathbb{R}$ tal que para todo X con $\int dX = 0$ y $\mu + \varepsilon X \in \mathcal{P}(\mathbb{R}^d)$ para un $\varepsilon > 0$ suficientemente pequeño,

$$\lim_{\varepsilon \rightarrow 0} \frac{\mathcal{F}(\mu + \varepsilon X) - \mathcal{F}(\mu)}{\varepsilon} = \int \delta \mathcal{F}(\mu) dX. \quad (\text{definido sobre constantes})$$

Cómo calcular $\delta \mathcal{F}(\mu)$: sea $(\mu_t)_{t \in \mathbb{R}}$ una curva, $\mu_0 = \mu$.
 Calcule $\partial_t \mathcal{F}(\mu_t)$ y escríbalo de la forma

$$\partial_t \mathcal{F}(\mu_t) \Big|_{t=0} = \int (\dots) \partial_t \mu_t \Big|_{t=0} \Rightarrow (\dots) = \delta \mathcal{F}(\mu)$$

TEOREMA: El gradiente de Wasserstein $\nabla_{W_2} \mathcal{F}(\mu)$ de $\mathcal{F}$ en μ es el campo vectorial

$$\nabla_{W_2} \mathcal{F}(\mu) = \nabla \delta \mathcal{F}(\mu).$$

Demostración. Por definición, $\nabla_{W_2} \mathcal{F}(\mu)$ satisface:

$$\bullet \text{ para } \partial_t \mu_t + \text{div}(\mu_t \nabla \varphi_t) = 0$$

$$\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{W_2} \mathcal{F}(\mu_t), \nabla \varphi_t \rangle_{\mu_t}.$$

- $\nabla_{\mu_t} F(\mu_t) \in T_{\mu} P_{2ac}(\mathbb{R}^d) \Leftrightarrow \nabla_{\mu_t} F(\mu_t)$ is an campo vectorial gradient

Note que,

$$\begin{aligned} \partial_t F(\mu_t) &= \int \delta F(\mu_t) \partial_t \mu_t = - \int \delta F(\mu_t) \operatorname{div}(\mu_t \nabla \varphi_t) \\ &= \int \langle \nabla \delta F(\mu_t), \nabla \varphi_t \rangle d\mu_t = \langle \nabla \delta F(\mu_t), \nabla \varphi_t \rangle_{\mu_t} \quad \blacksquare \end{aligned}$$

COROLARIO: Et Wasserstein gradient flow de F satisfait

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \delta F(\mu_t)) \Leftrightarrow \dot{x}_t = -\nabla \delta F(\mu_t)(x_t)$$