Generative Models and Optimal Transport

Marco Cuturi

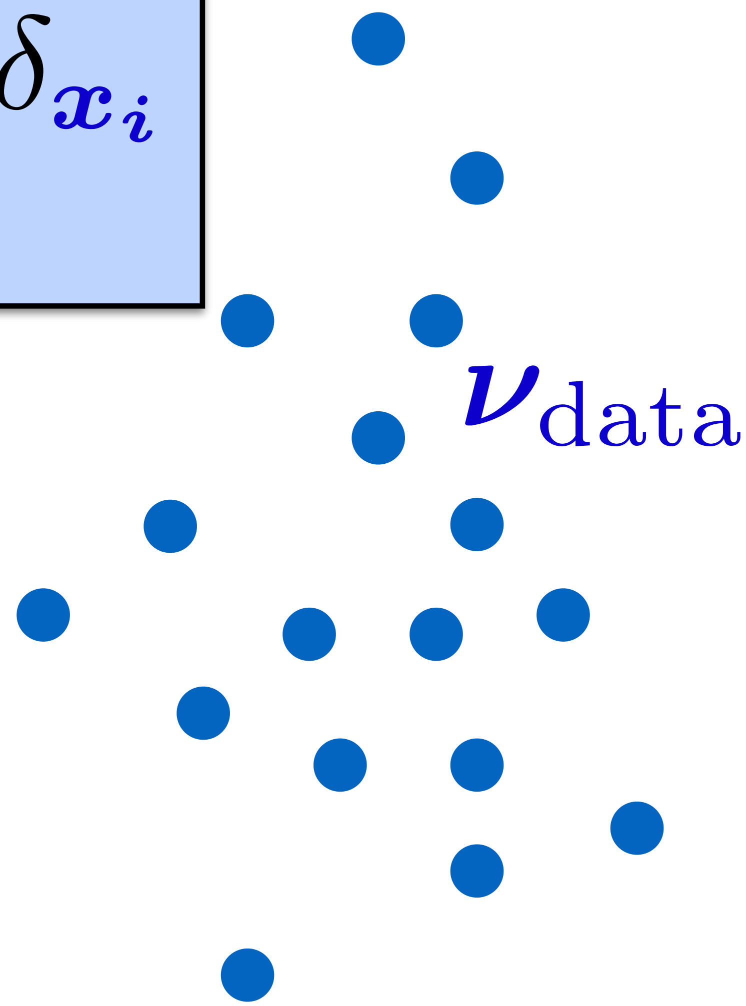
ENSAE ParisTech
École nationale de la statistique et de l'administration économique
université PARIS-SACLAY

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

Statistics 0.1 : Density Fitting

We collect data

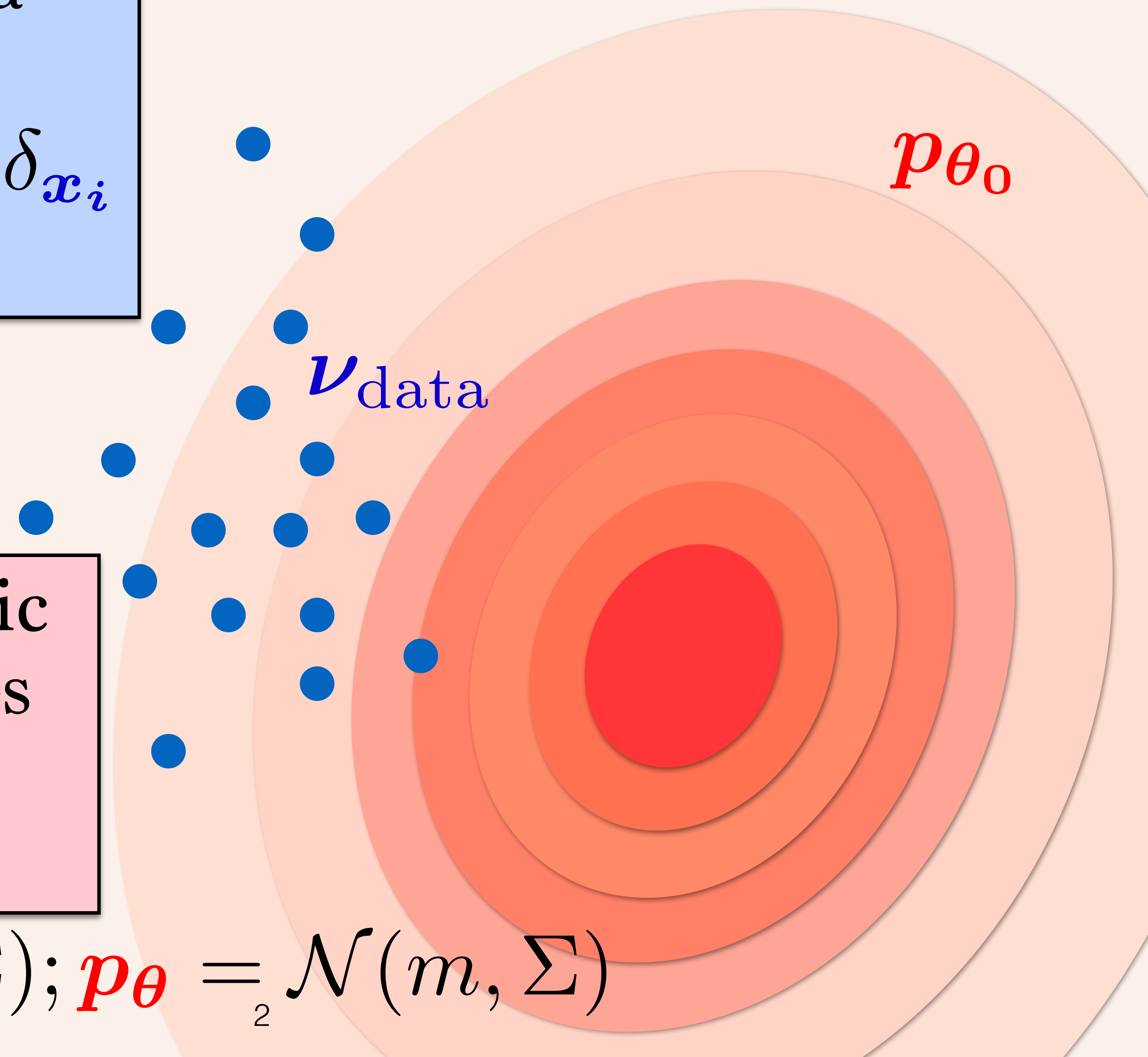
$$\boldsymbol{\nu}_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{\boldsymbol{x_i}}$$



Statistics 0.1 : Density Fitting

We collect data

$$\boldsymbol{\nu}_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{x_i}}$$

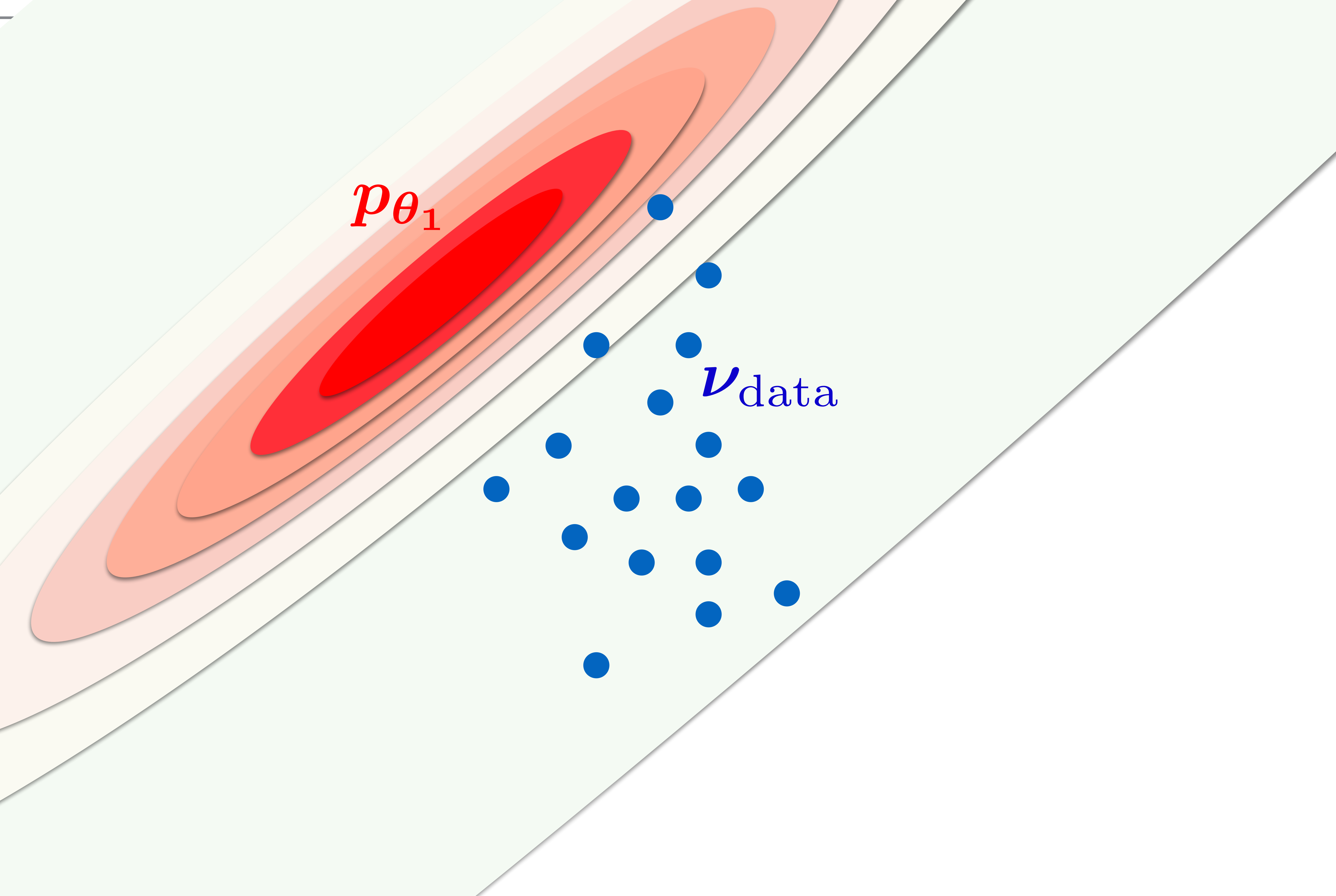


We fit a parametric family of densities

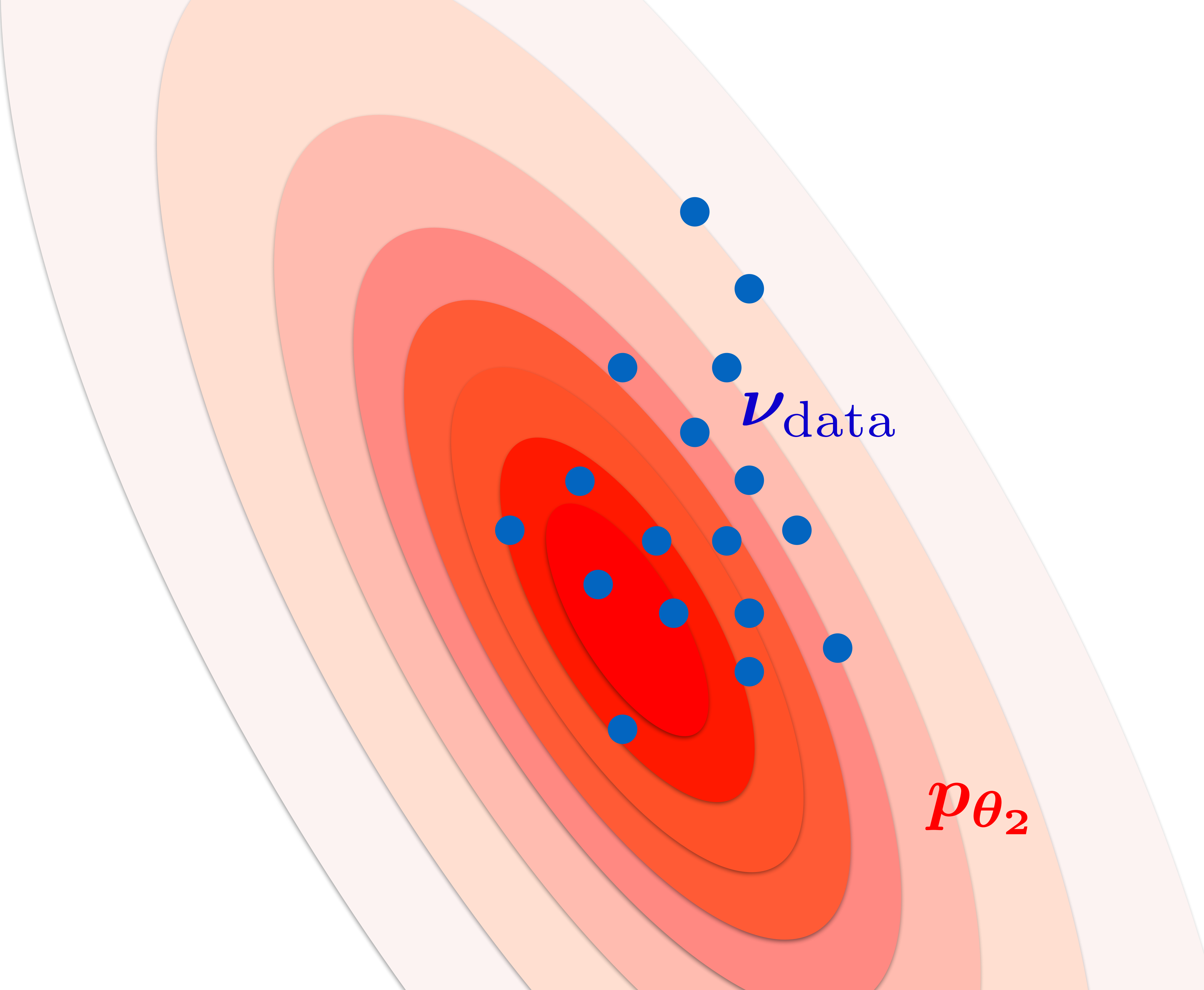
$$\{\boldsymbol{p_\theta}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

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

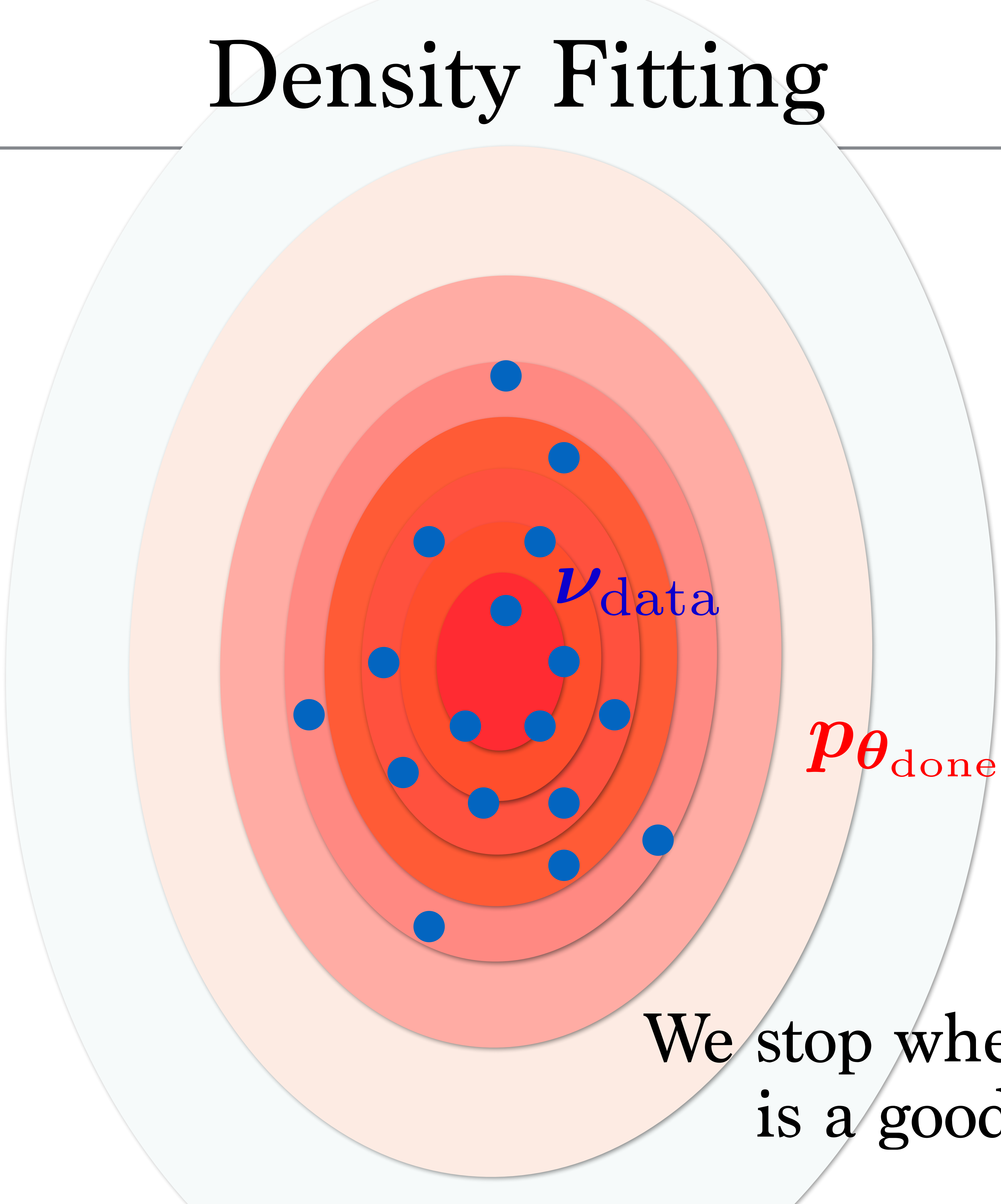
Density Fitting



Density Fitting

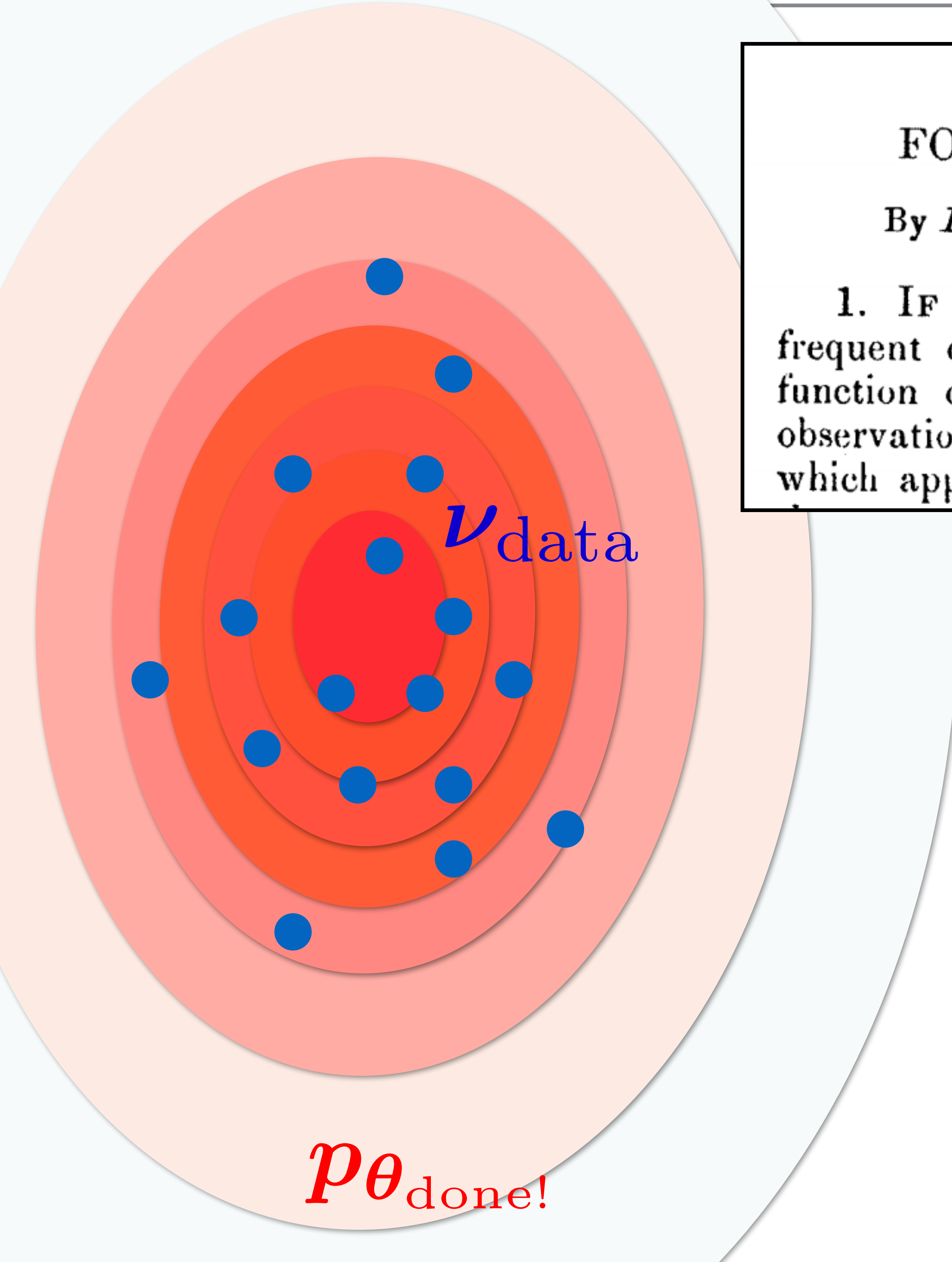


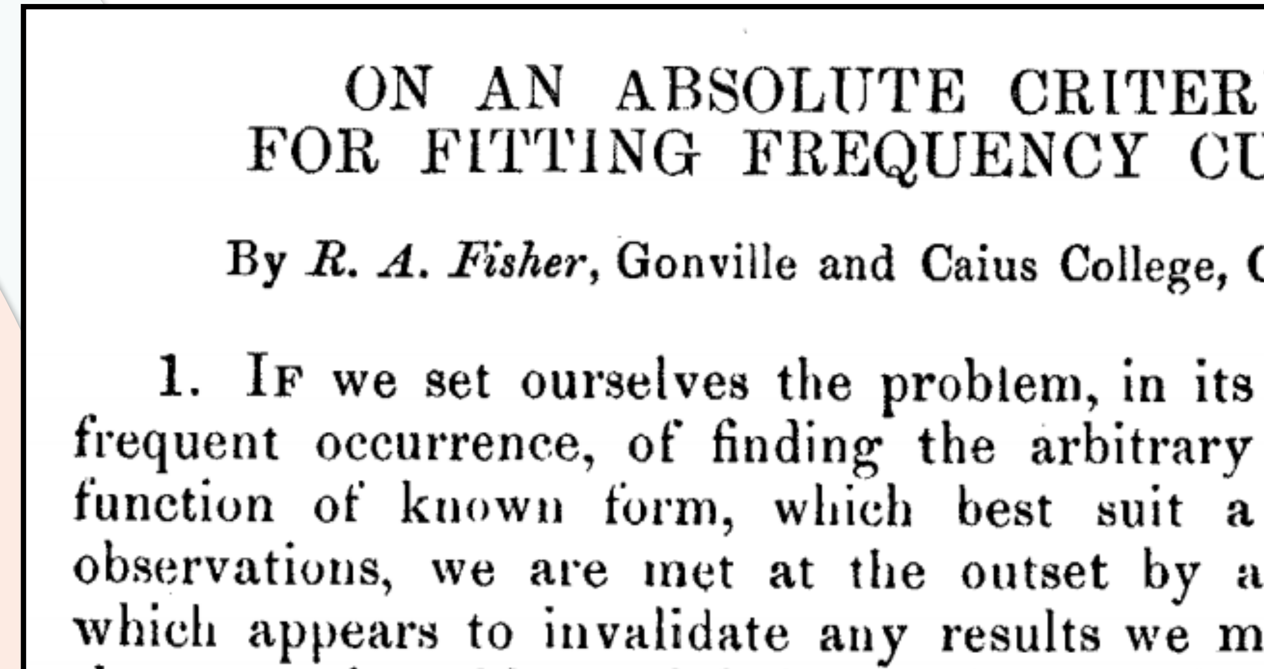
Density Fitting



We stop when there is a good fit.

Maximum Likelihood Estimation





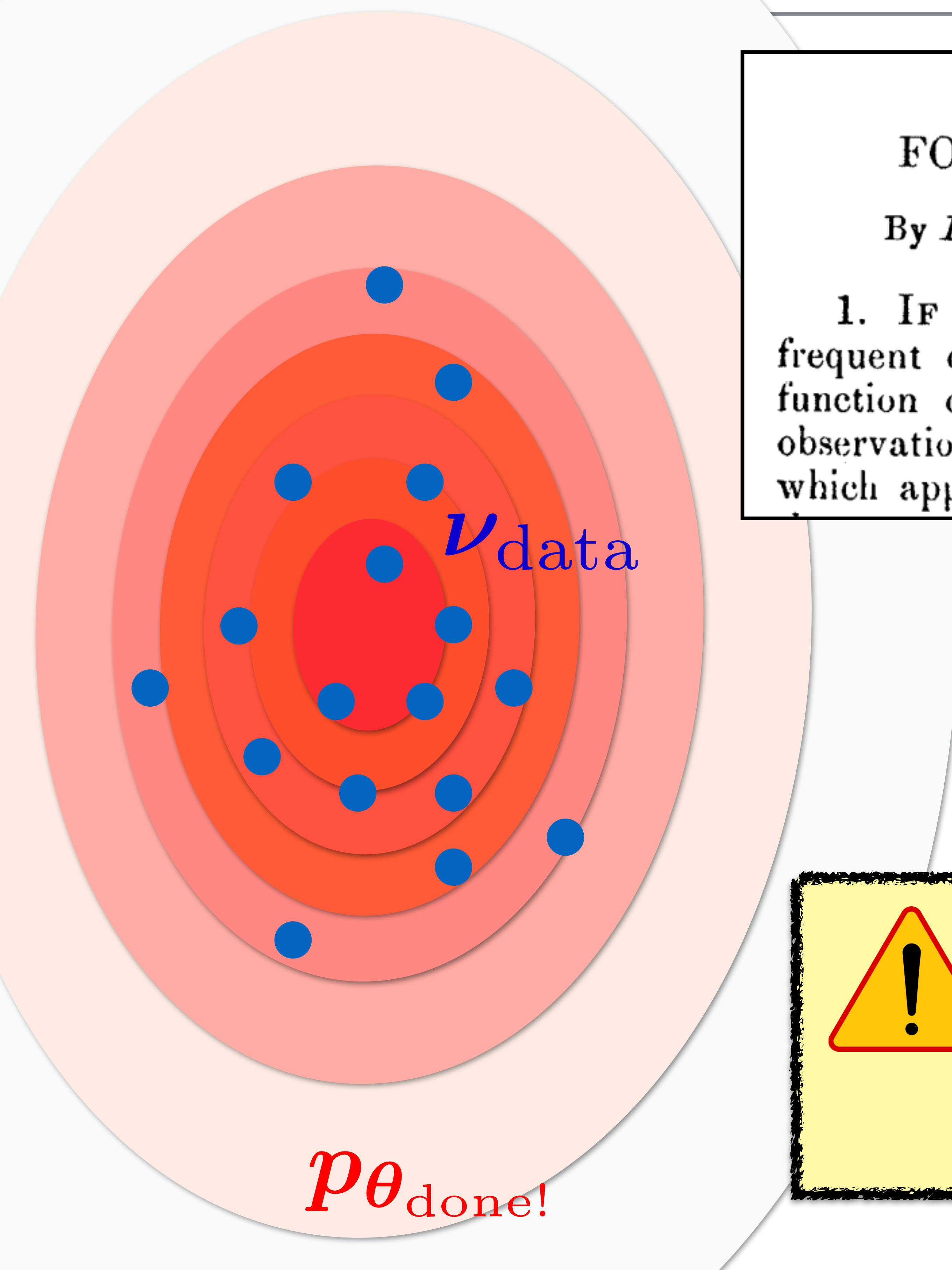
ON AN ABSOLUTE CRITERION FOR FITTING FREQUENCY CURVES.

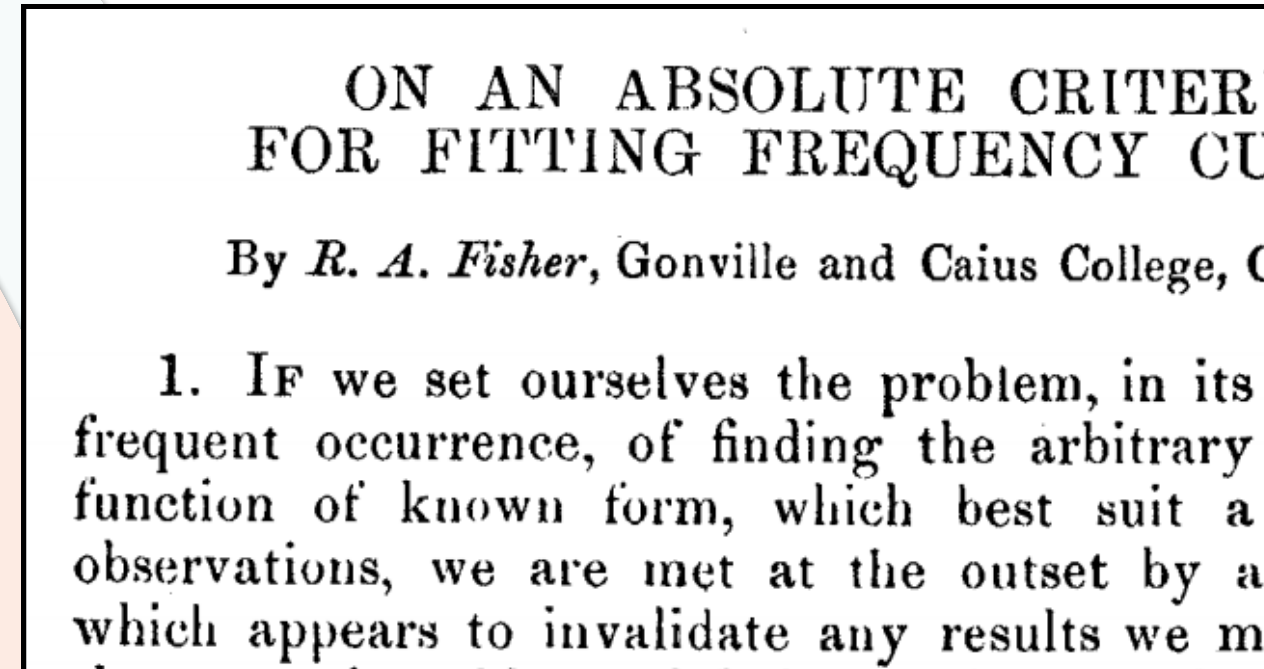
By *R. A. Fisher*, Gonville and Caius College, Cambridge.

1. If we set ourselves the problem, in its
frequent occurrence, of finding the arbitrary
function of known form, which best suit a
observations, we are met at the outset by an
which appears to invalidate any results we ma

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

Maximum Likelihood Estimation





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
frequent occurrence, of finding the arbitrary
function of known form, which best suit a
observations, we are met at the outset by an
which appears to invalidate any results we ma

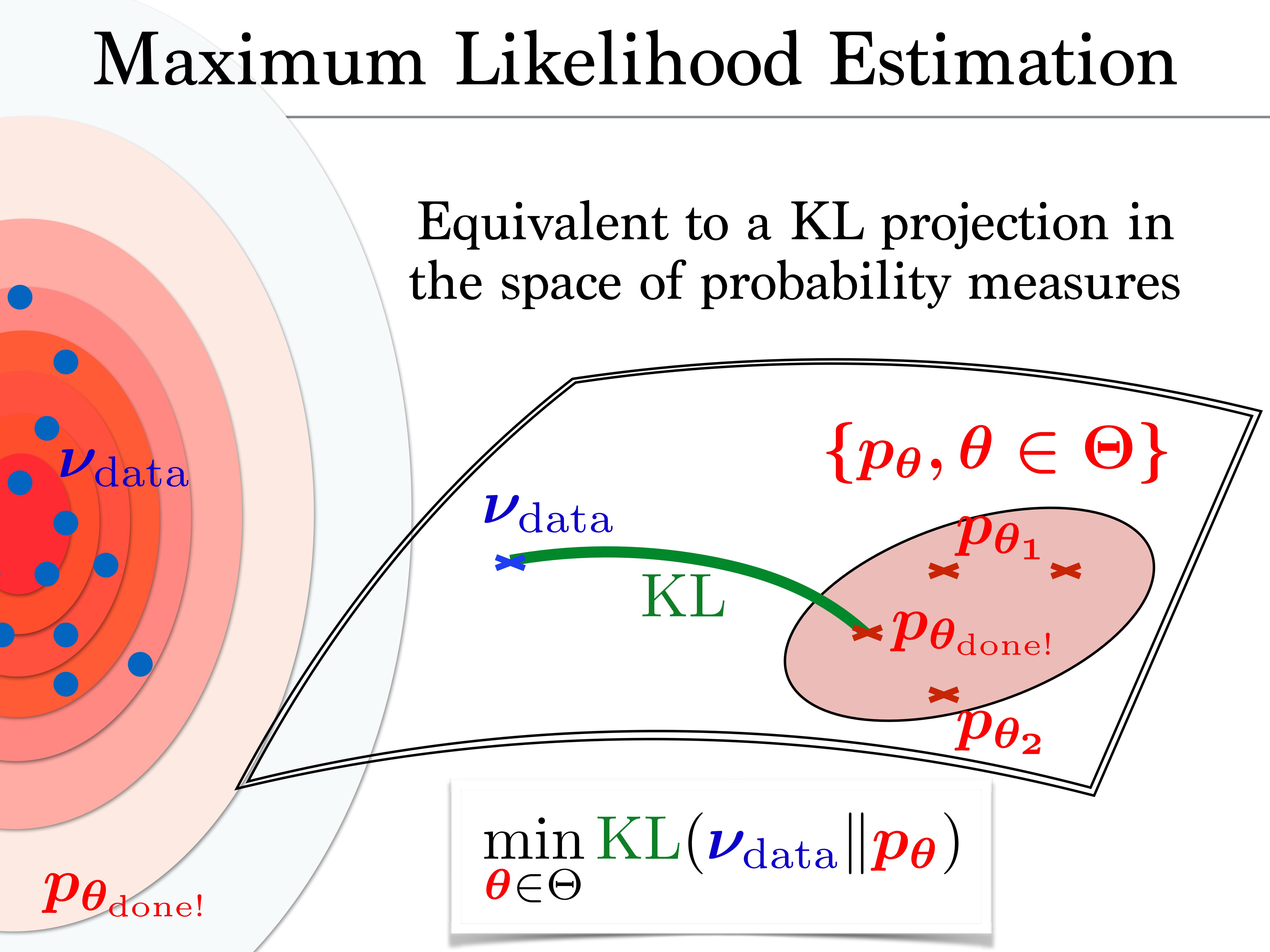
$$\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log p_{\theta}(x_i)$$

$$\log 0 = -\infty$$

$p_{\theta}(x_i)$ **must** be > 0

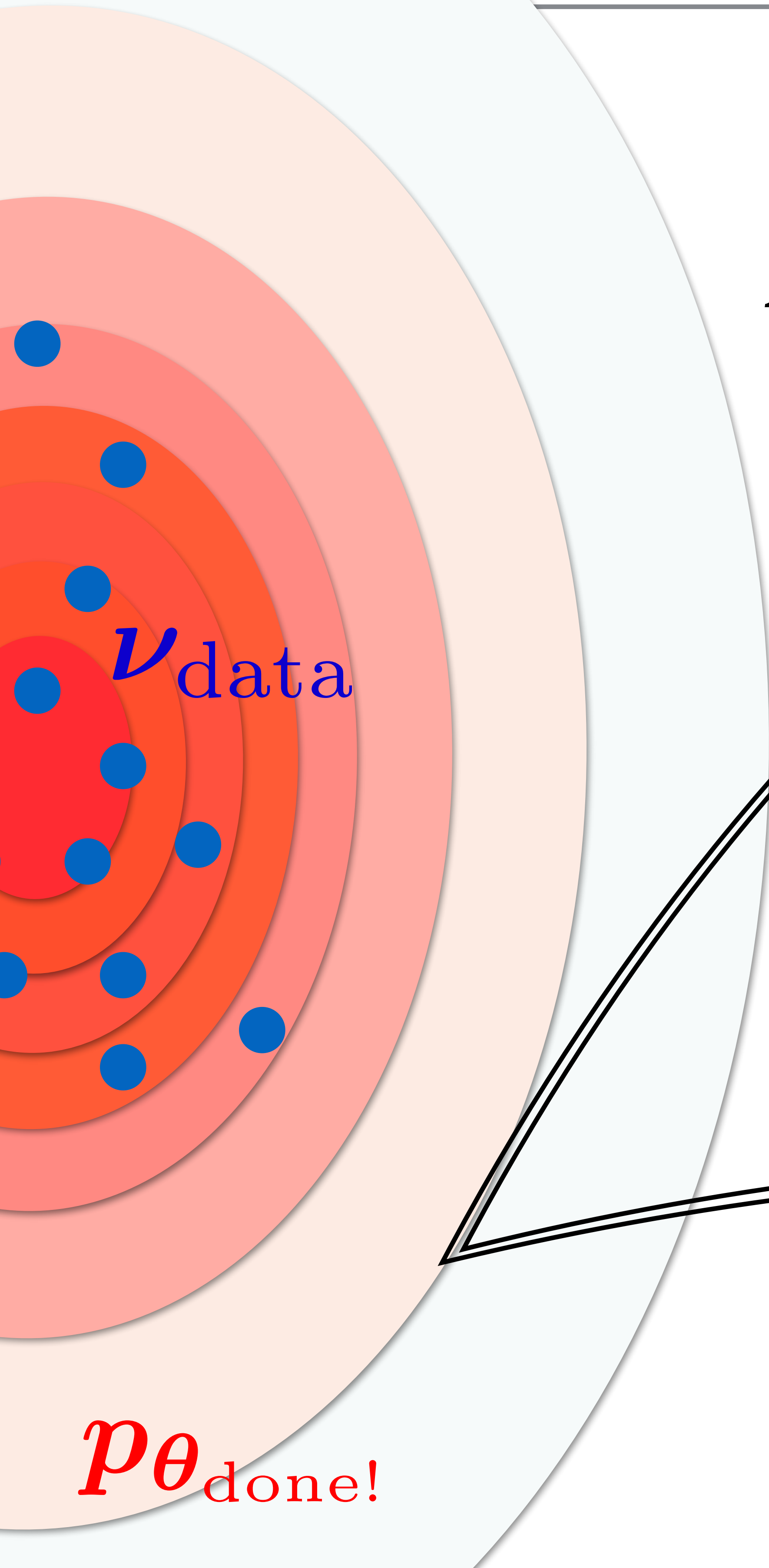
Maximum Likelihood Estimation

Equivalent to a KL projection in the space of probability measures

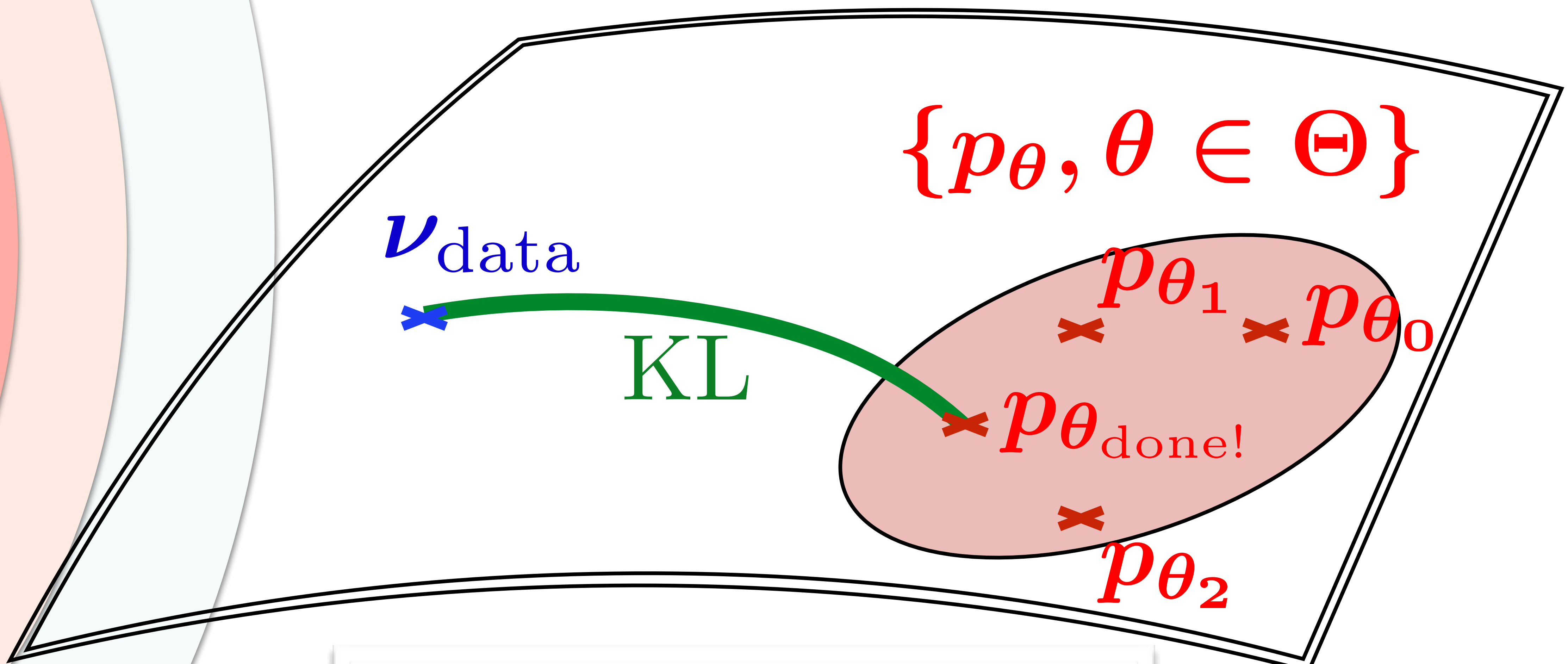


$$\min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_\theta)$$

Maximum Likelihood Estimation

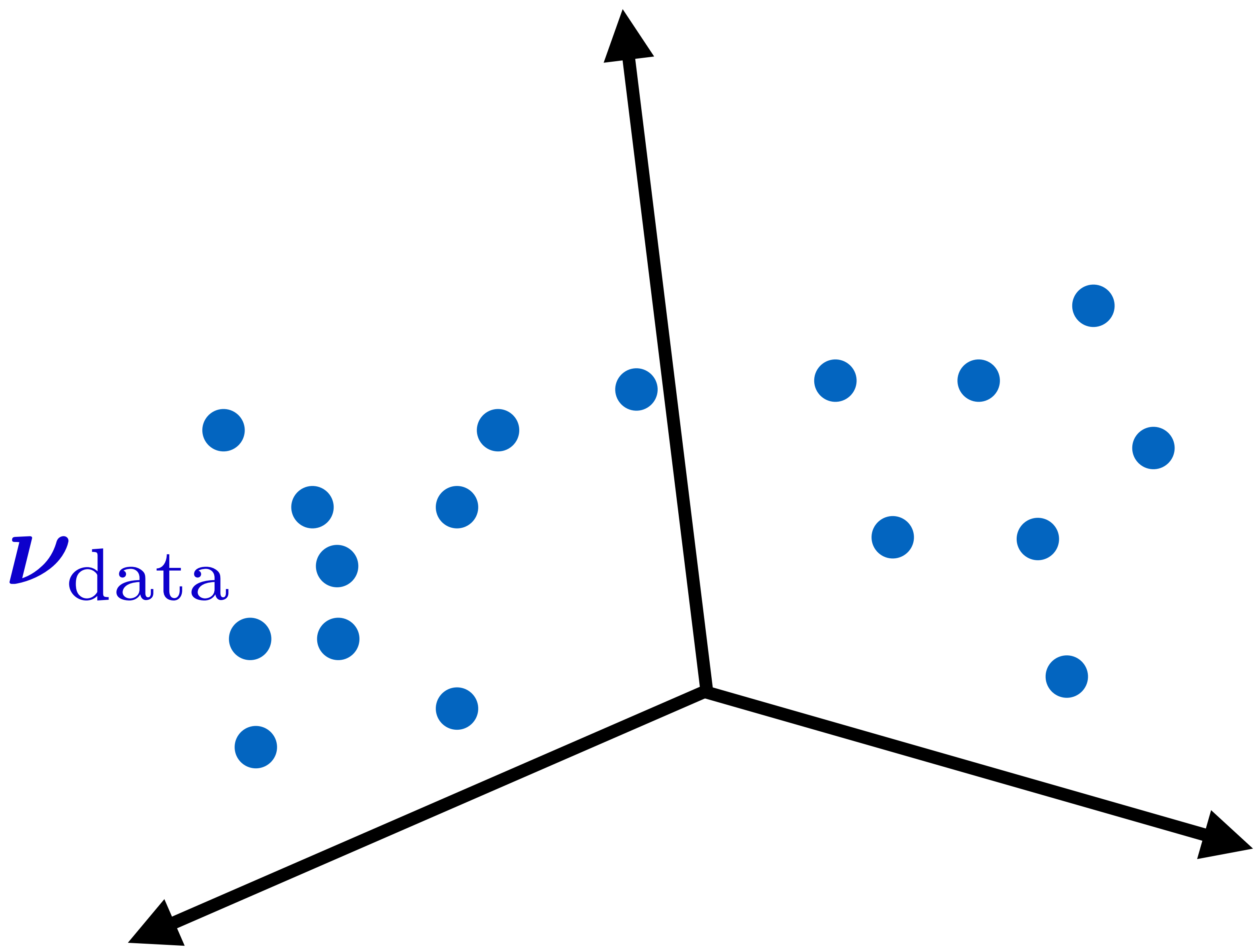


Equivalent to a KL projection in the space of probability measures



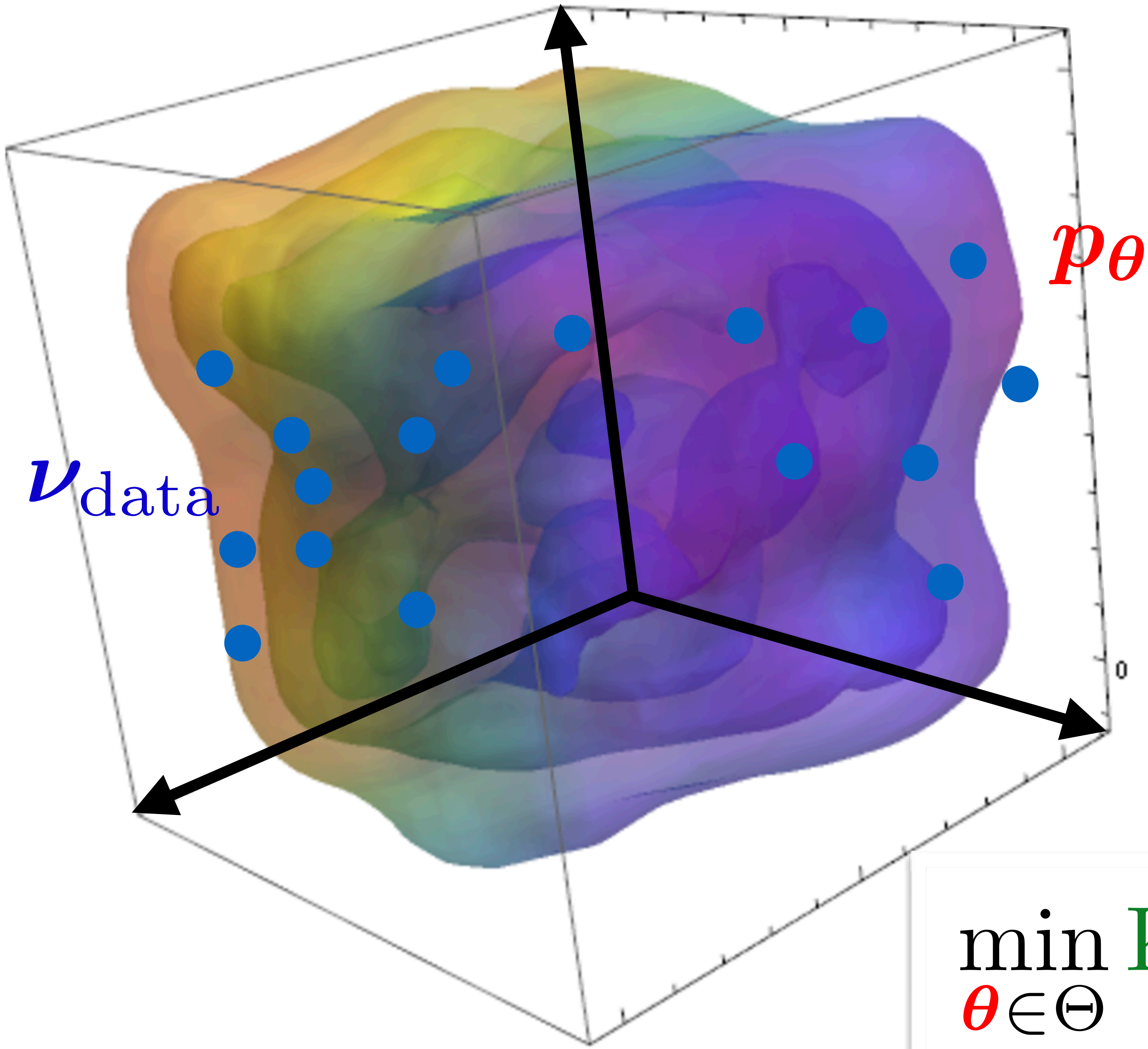
$$\min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_\theta)$$

In higher dimensional spaces...



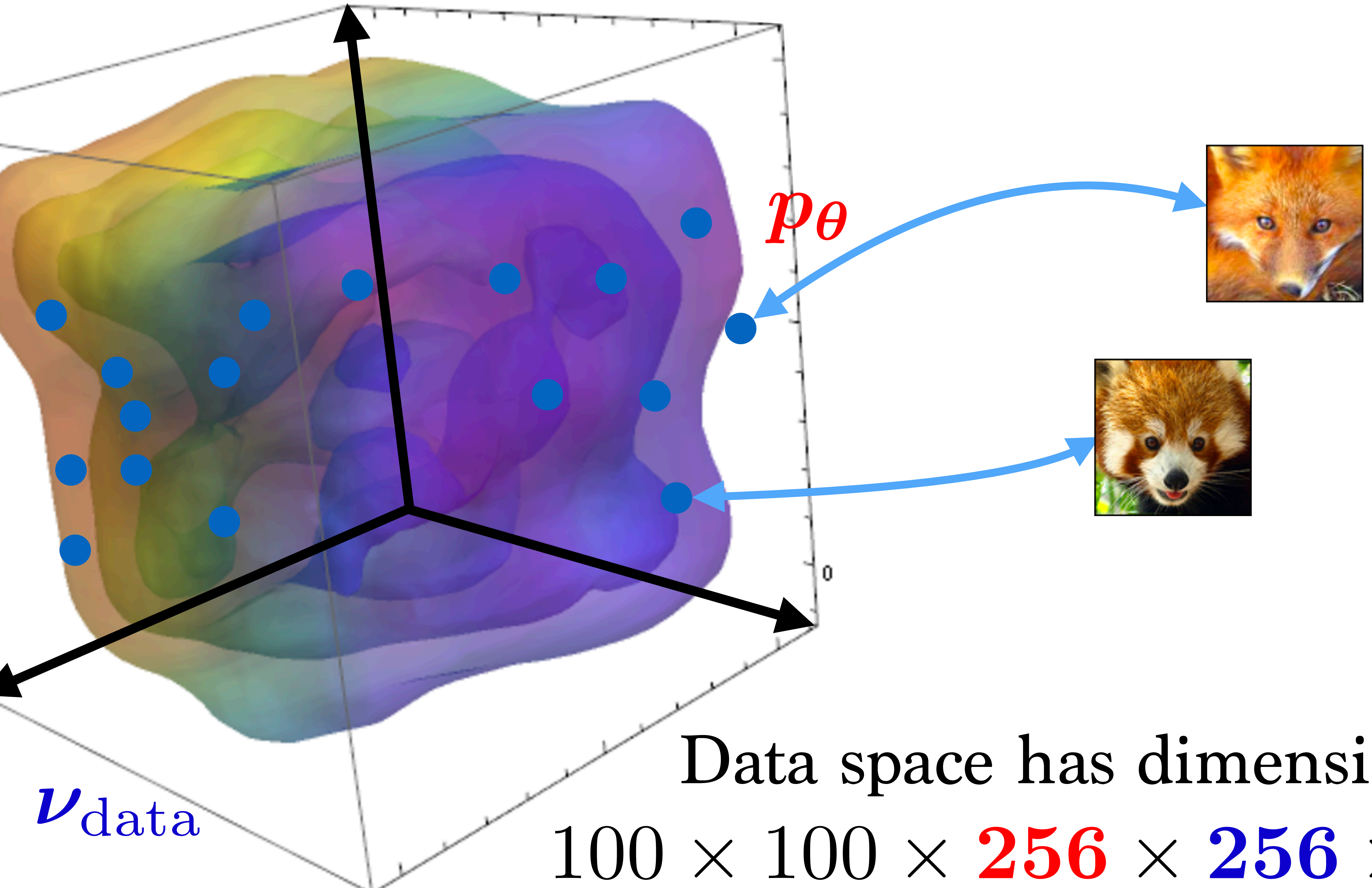
$$\min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_{\theta})$$

In higher dimensional spaces...



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

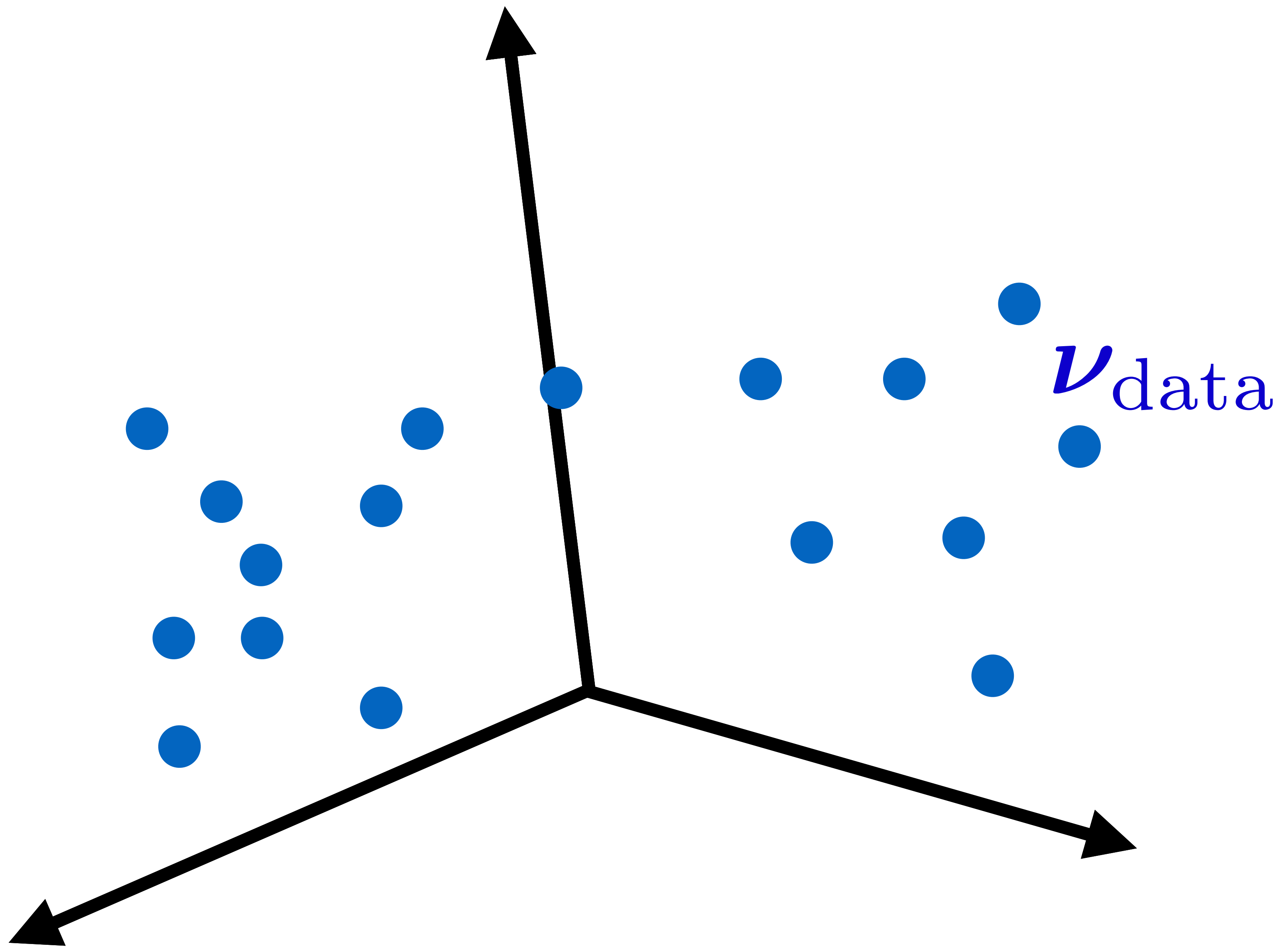
In higher dimensional spaces...



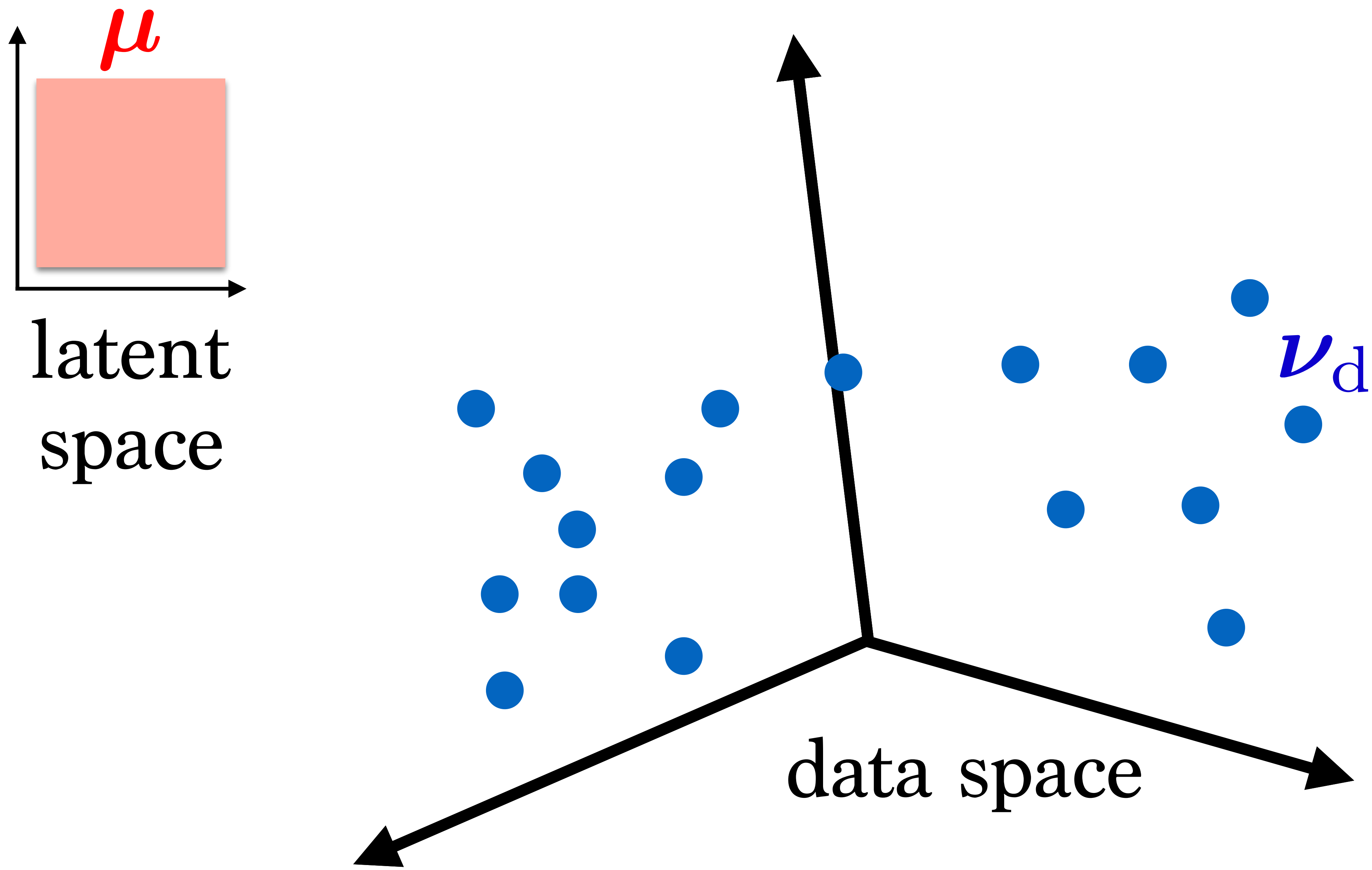
Data space has dimension

$$100 \times 100 \times \mathbf{256} \times \mathbf{256} \times \mathbf{256} \approx 167 \times 10^9$$

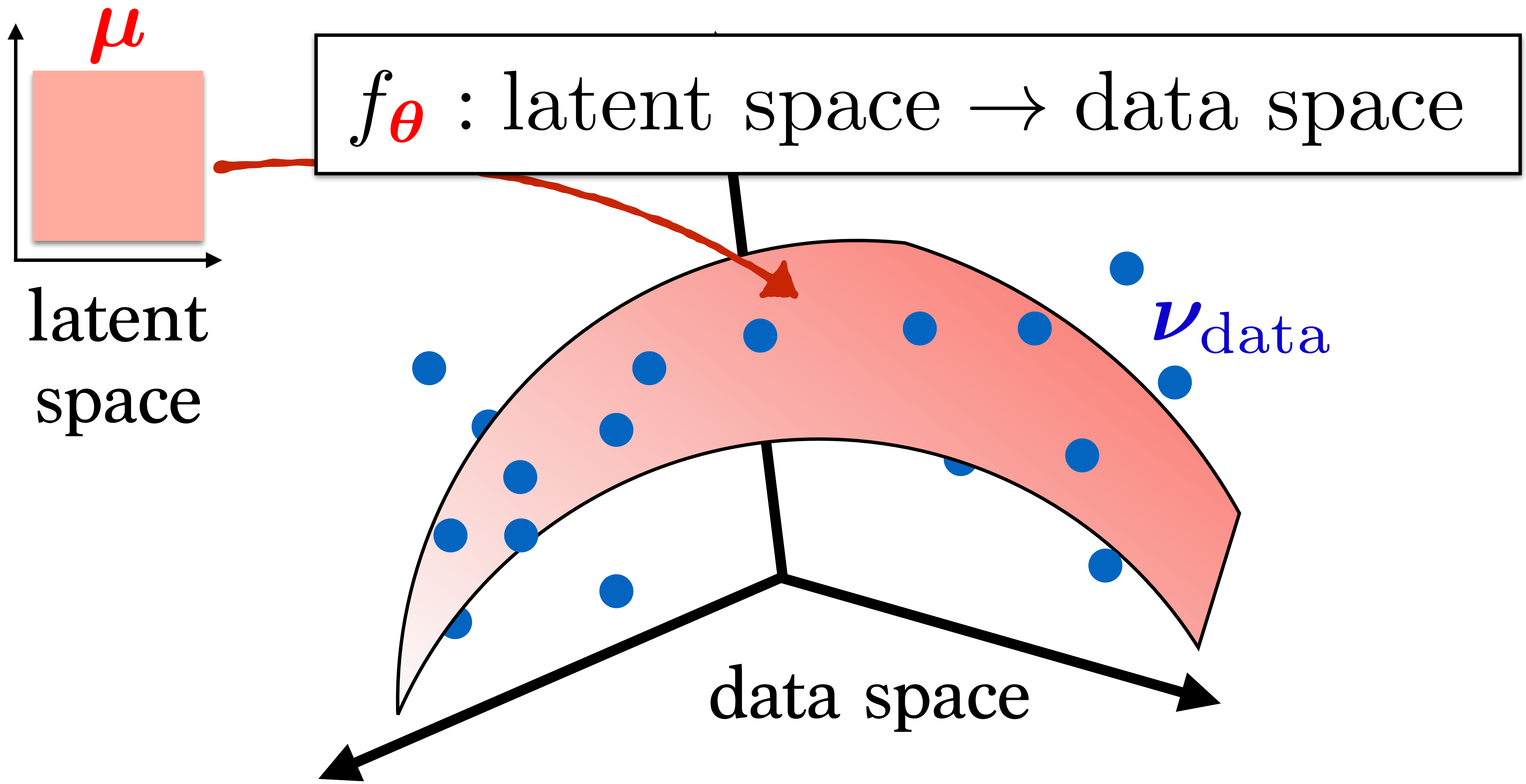
Generative Models



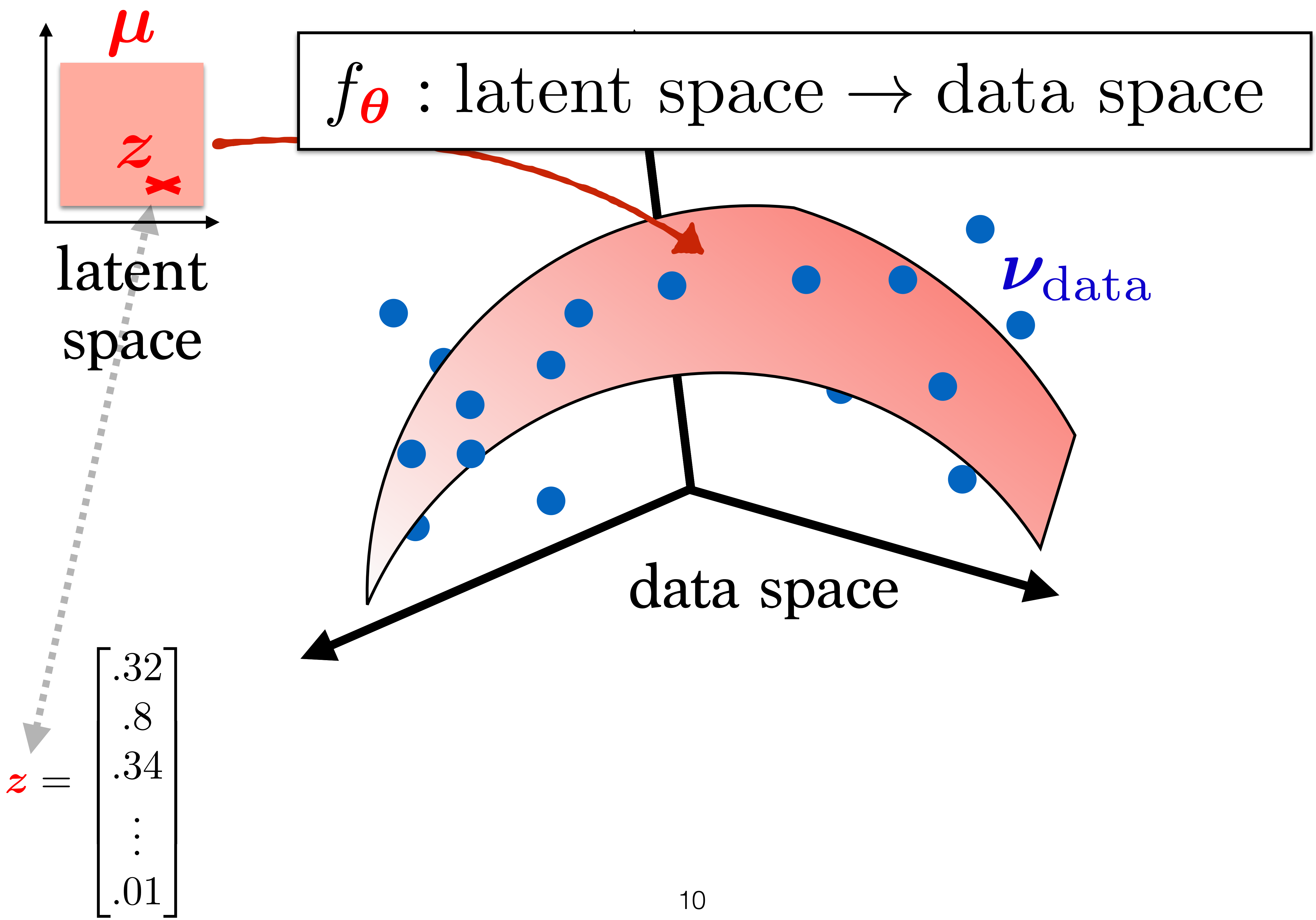
Generative Models



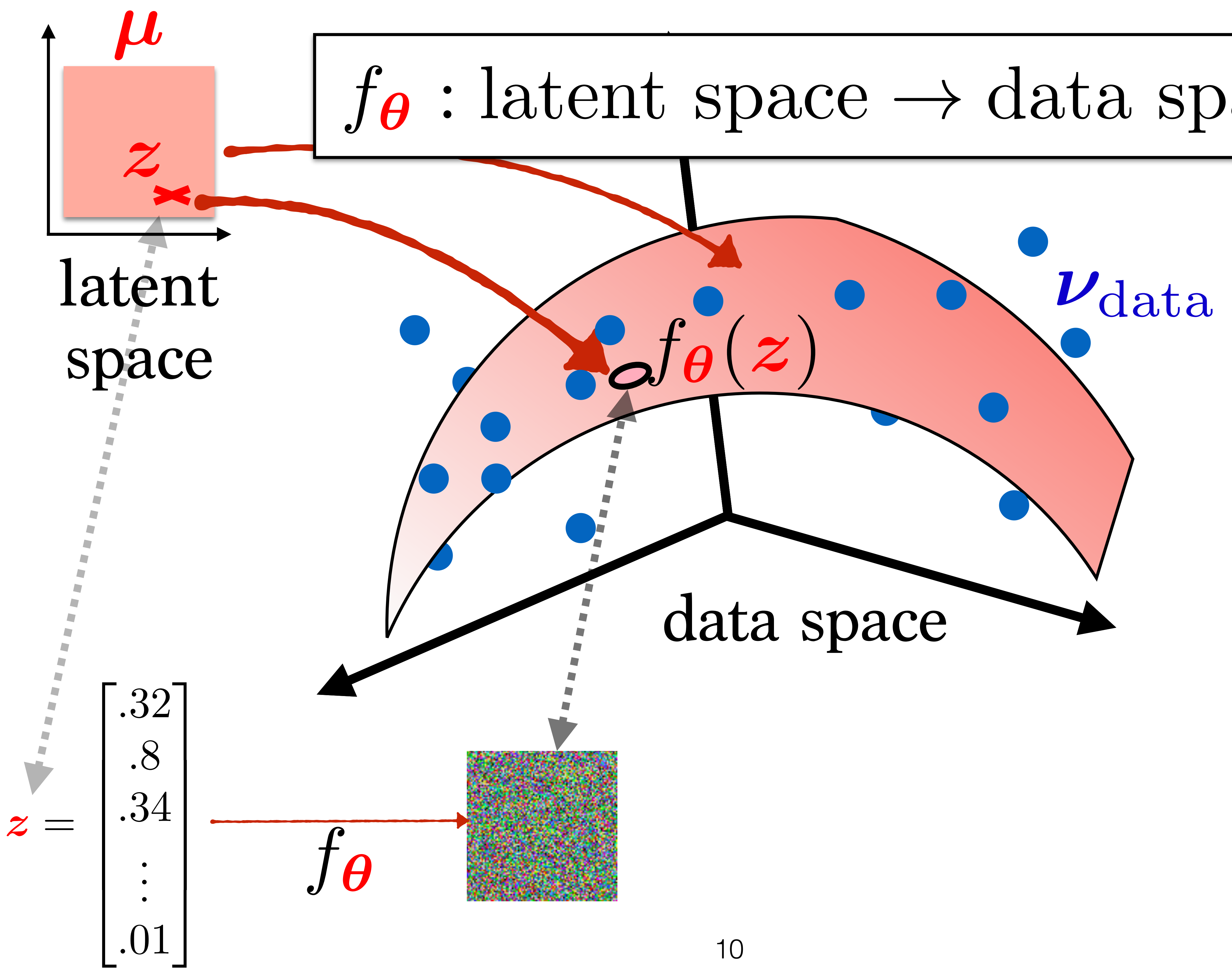
Generative Models



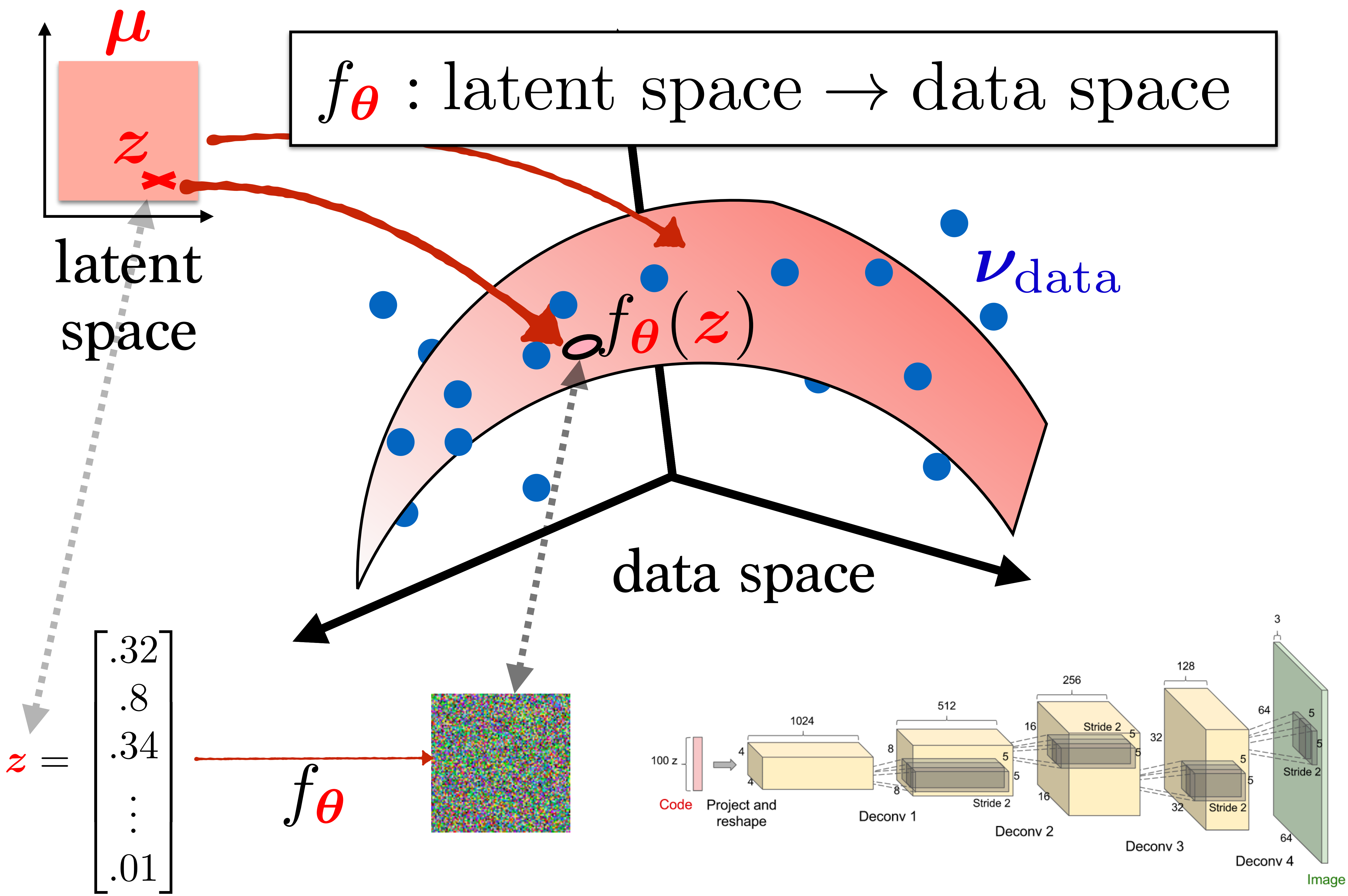
Generative Models



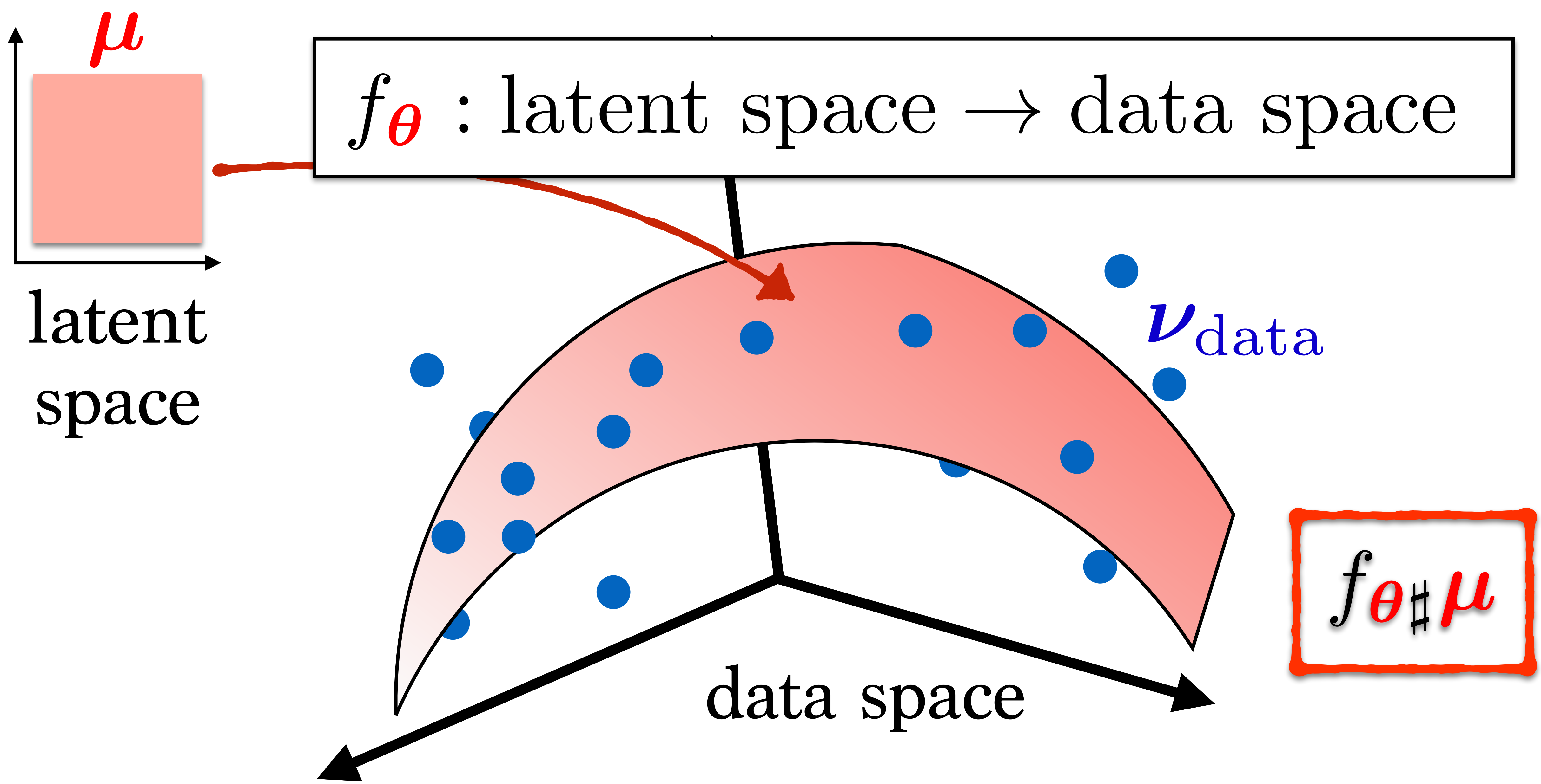
Generative Models



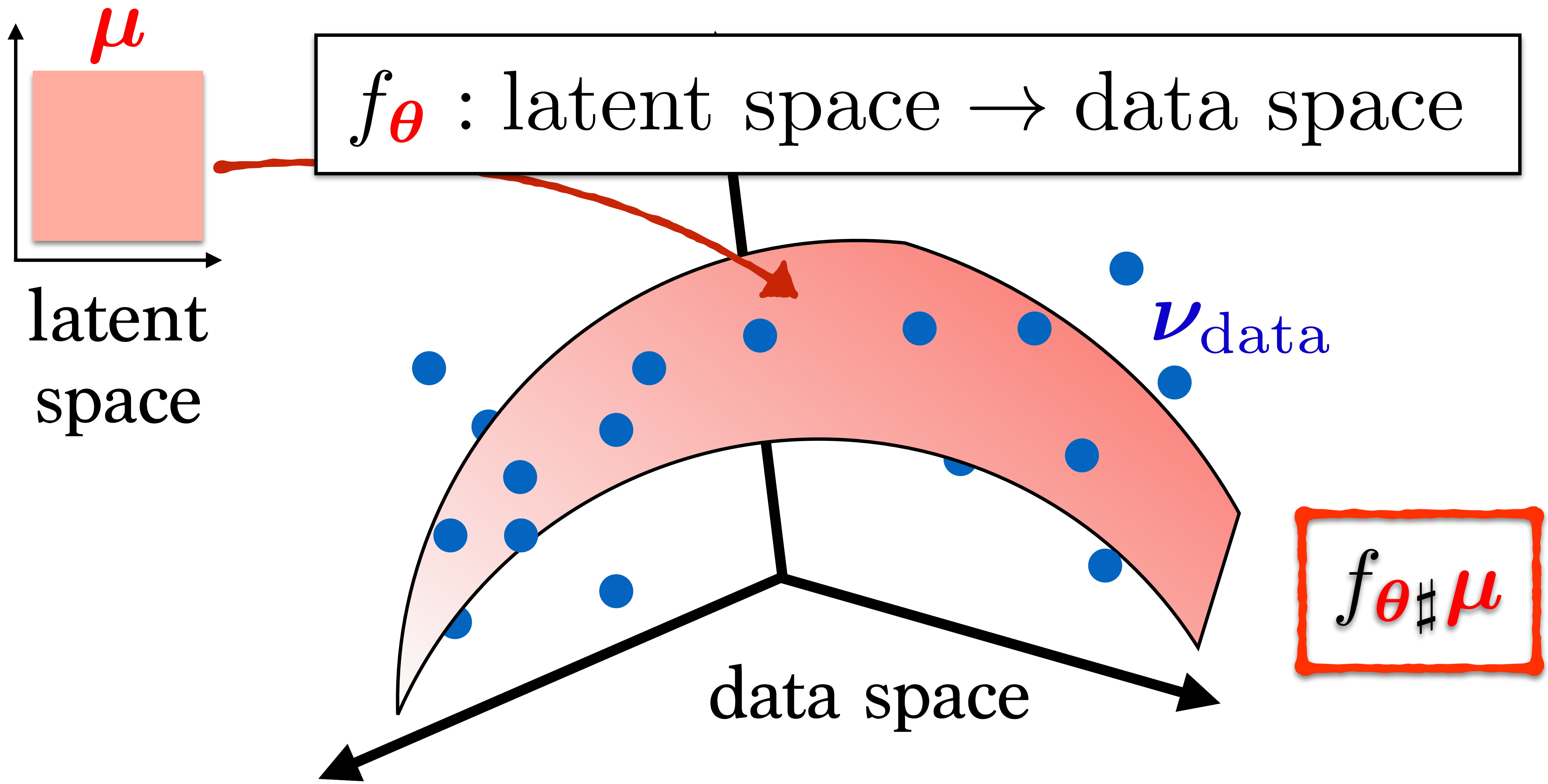
Generative Models



Generative Models

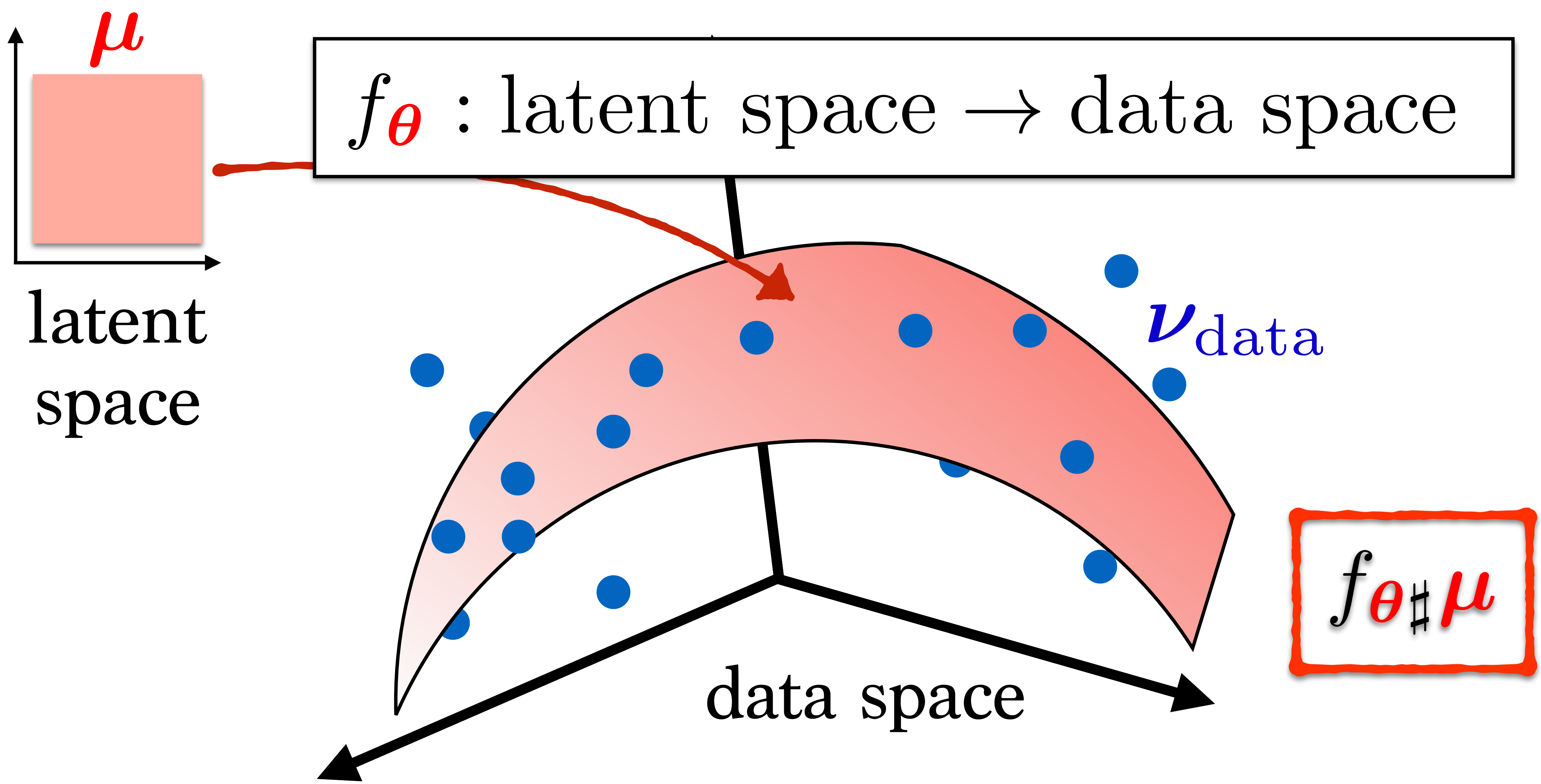


Generative Models



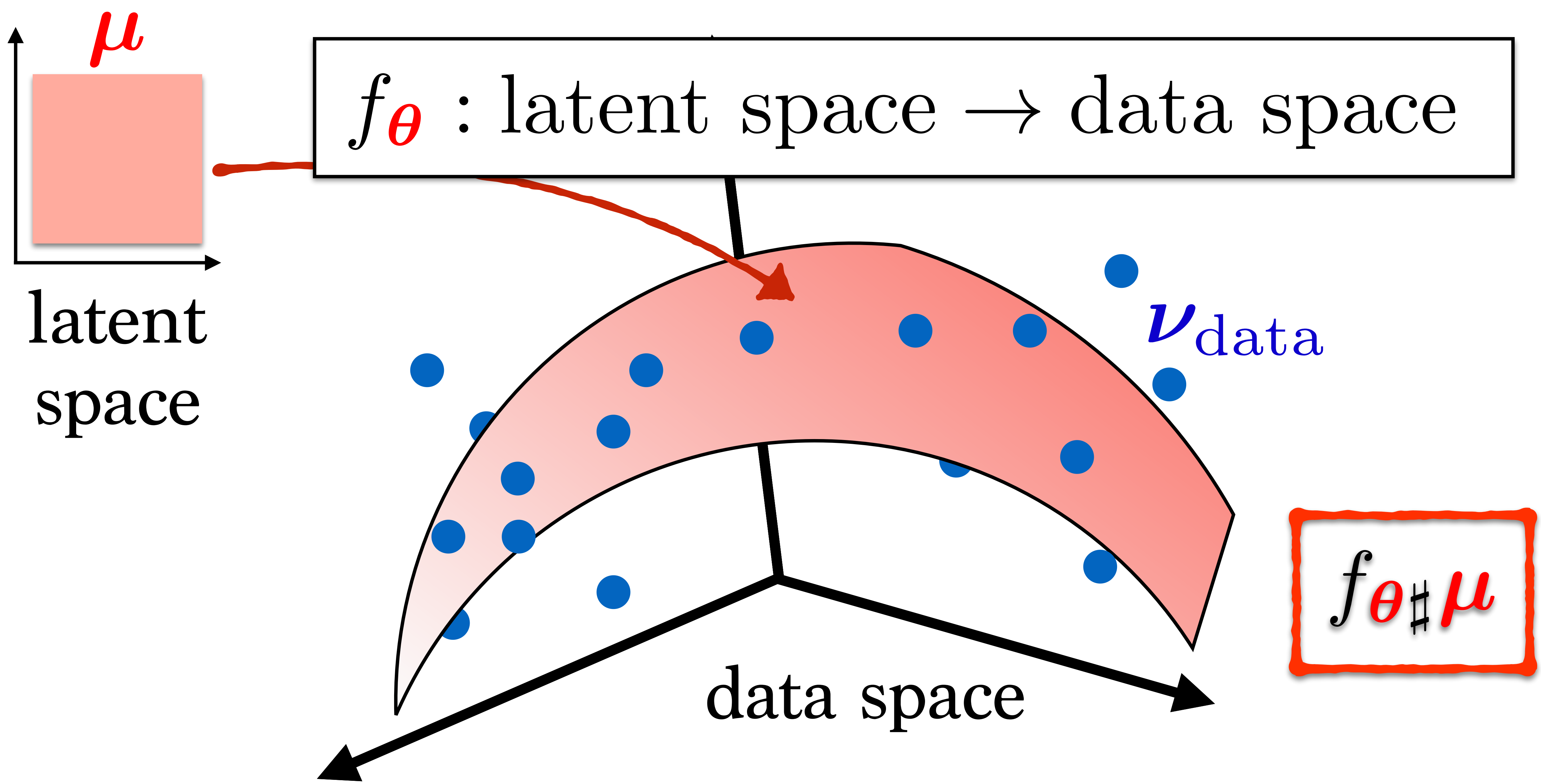
$$\text{Push-forward: } \forall B \subset \Omega,\ f_{\sharp}\mu(B) := \mu(f^{-1}(B))$$

Generative Models



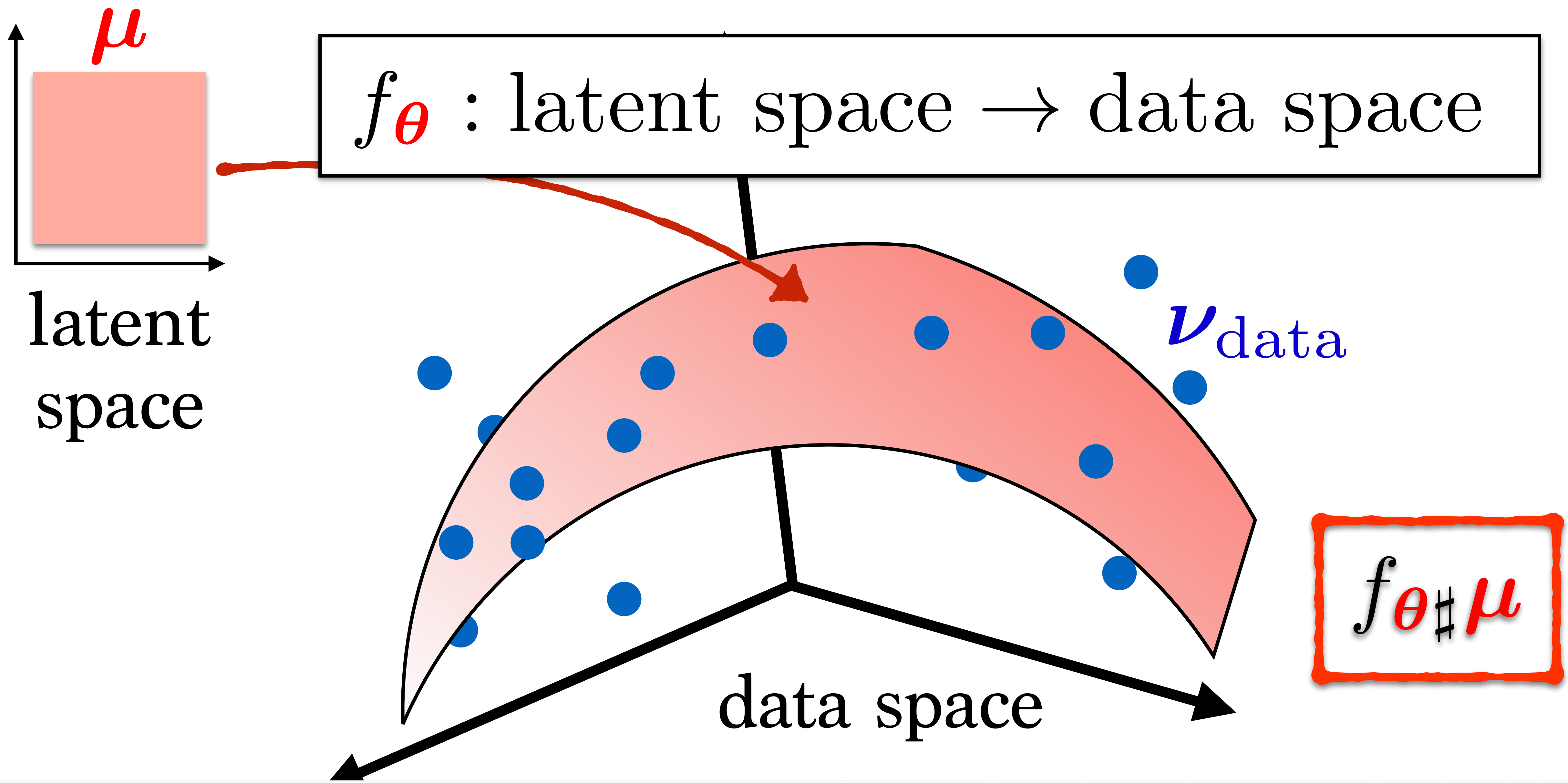
Goal: find θ such that $f_{\theta\sharp}\mu$ fits ν_{data}

Generative Models



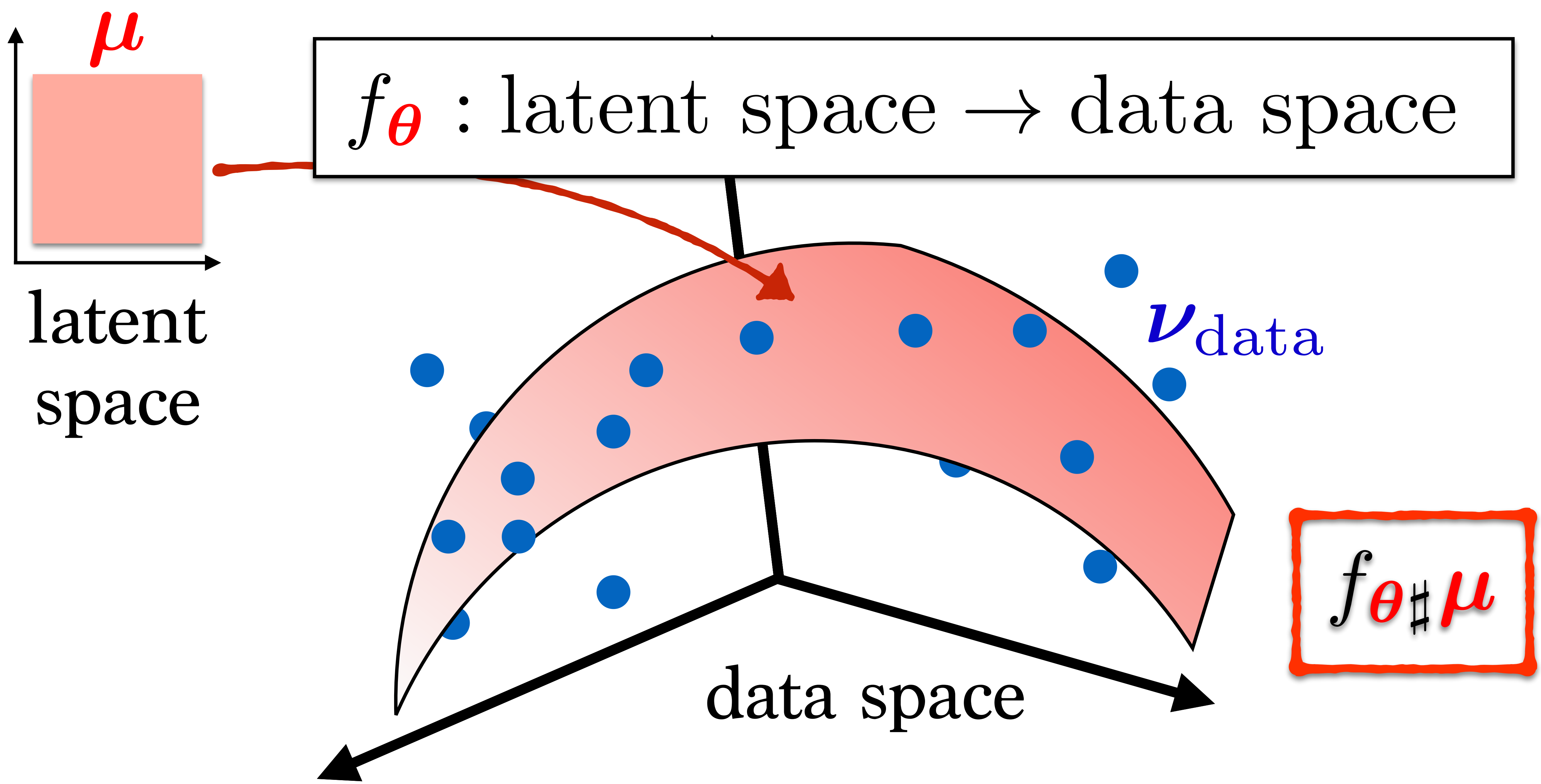
$\boxed{\text{Goal: find } \theta \text{ such that } f_{\theta\sharp}\mu \text{ fits } \nu_{\text{data}}}$

Generative Models



Difference between fitting a push forward measure $f_{\theta\sharp}\mu$ *vs.* a density p_{θ}?

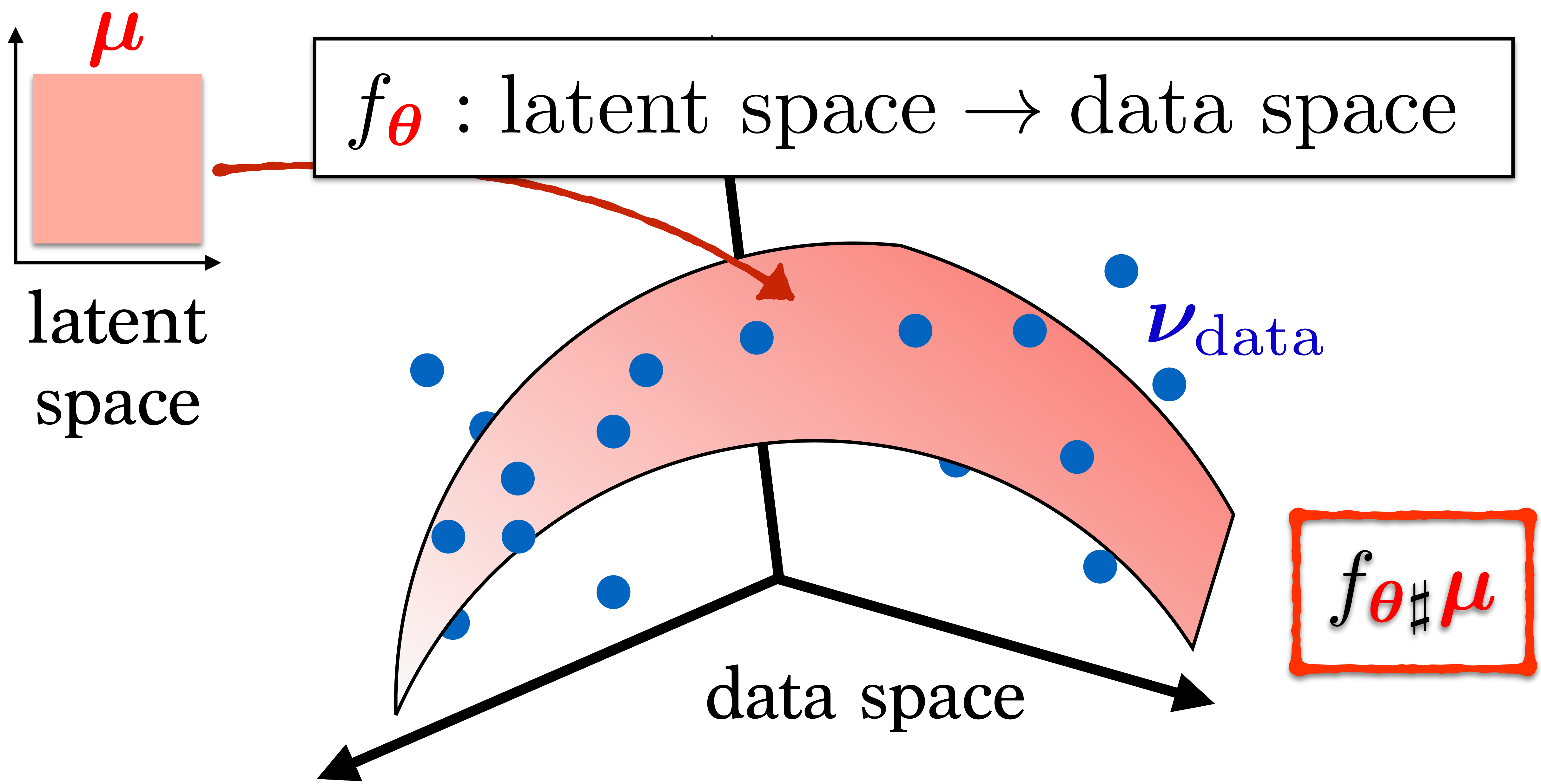
Generative Models



MLE

$$\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(x_i) \quad = \quad \min_{\theta\in\Theta} \mathrm{KL}(\nu_{\text{data}} \| p_\theta)$$

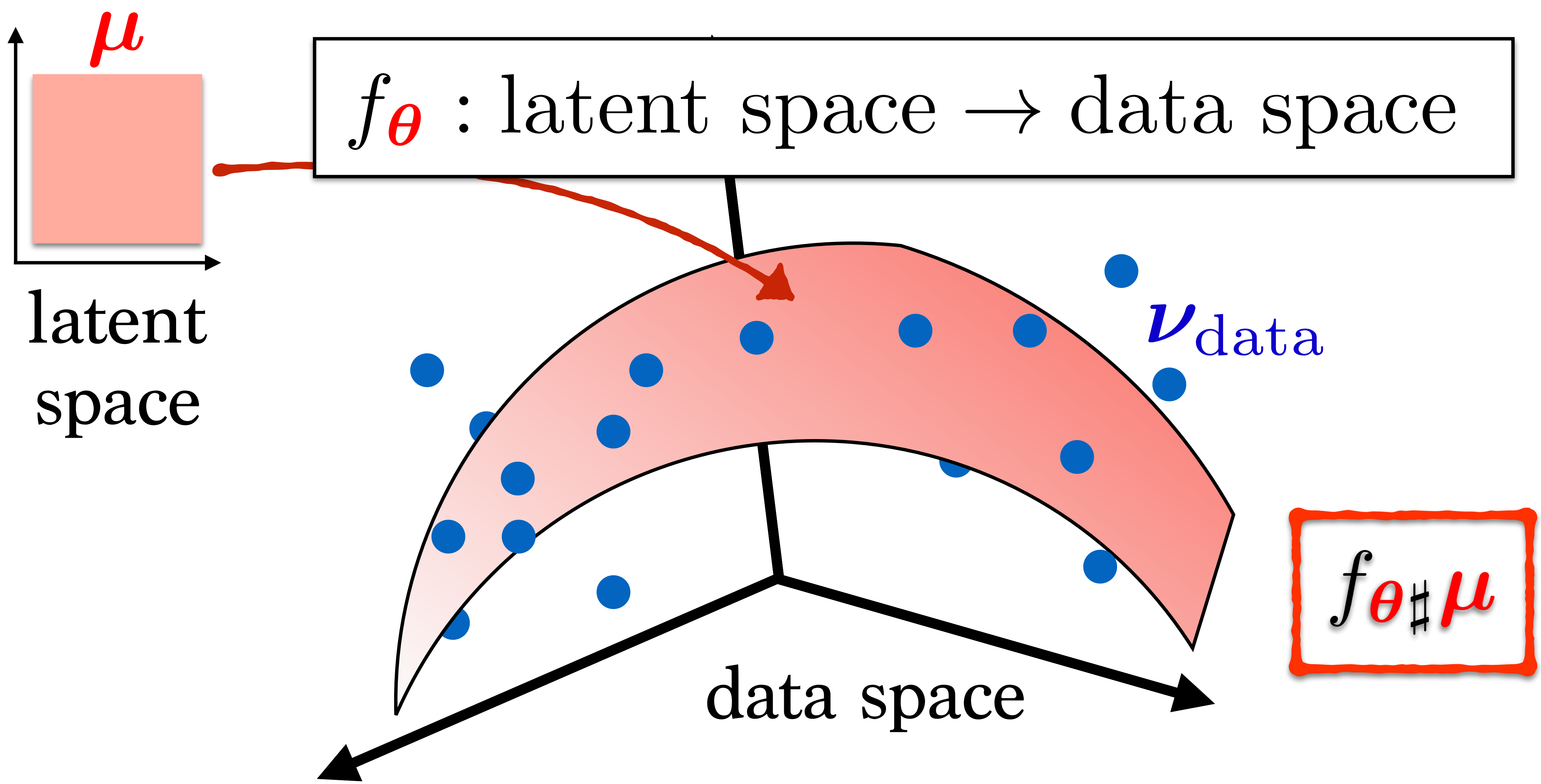
Generative Models



MLE

$$\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log f_{\theta\sharp}\mu(x_i) \qquad \min_{\theta\in\Theta} \mathrm{KL}(\nu_{\text{data}} \| f_{\theta\sharp}\mu)$$

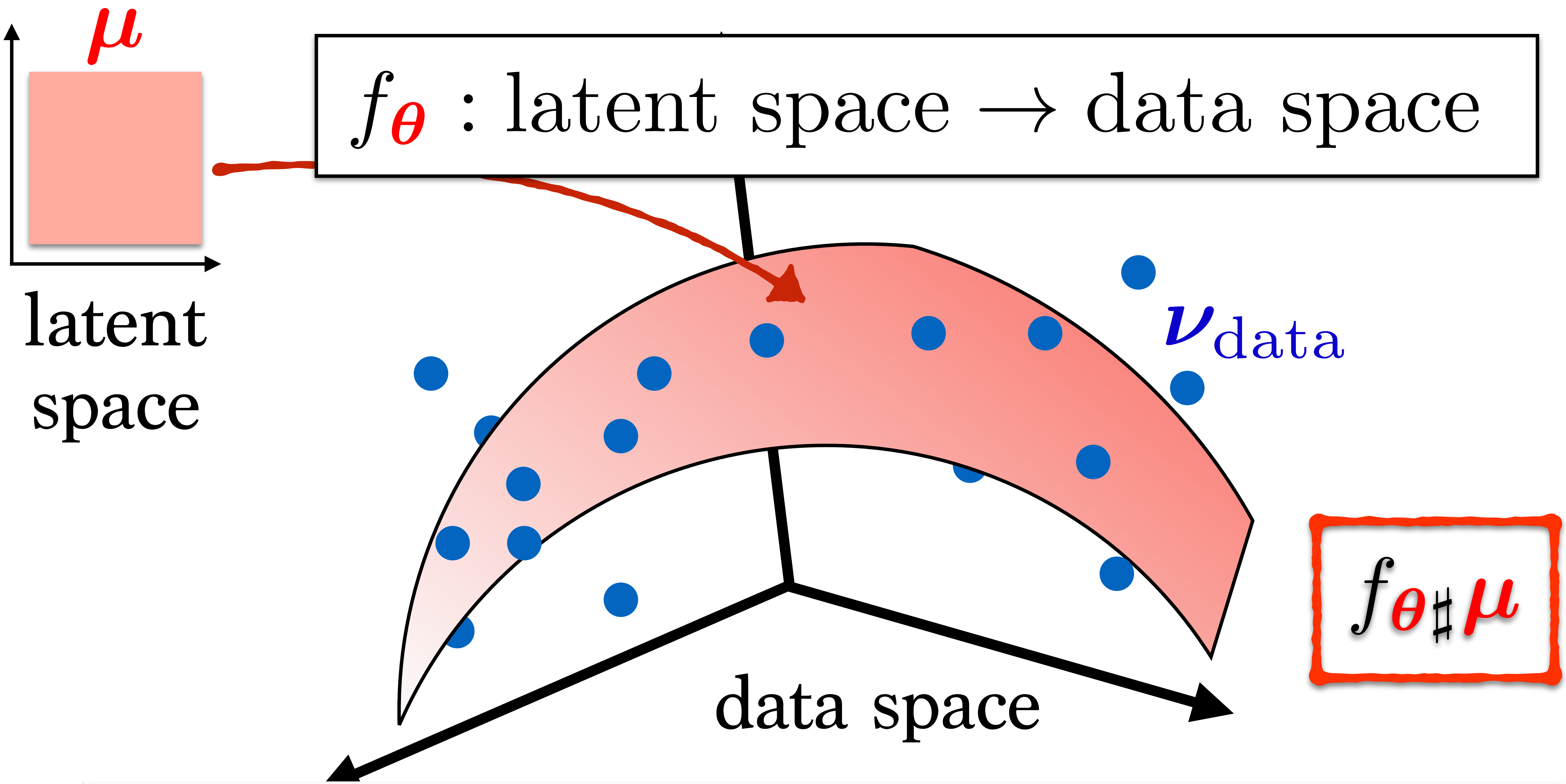
Generative Models



~~MLE~~

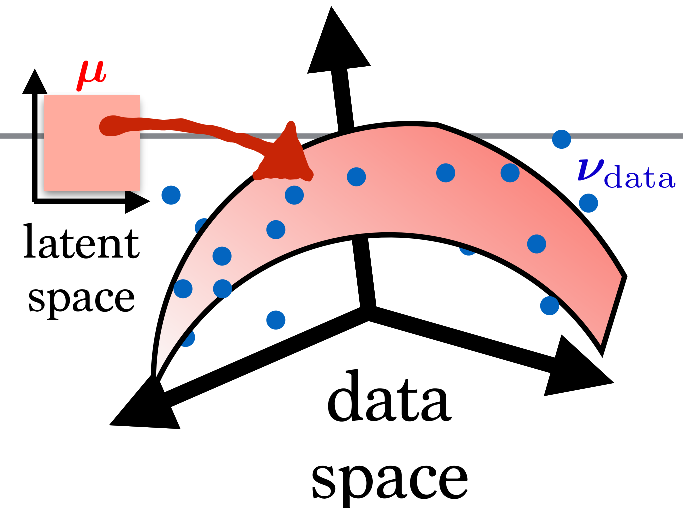
$$\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log \underline{f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu}}(\boldsymbol{x}_i) \qquad \min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\boldsymbol{\nu}_{\text{data}} \| \underline{f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu}})$$

Generative Models



Need a more flexible **discrepancy function** to compare ν_{data} and $f_{\theta\sharp}\mu$

Workarounds?



- Formulation as **adversarial problem** **[GPM…'14]**

$$\min_{\theta\in\Theta} \max_{\text{classifiers } g} \text{Accuracy}_{g}\left(\left(f_{\theta\sharp}\mu, +1\right), \left(\nu_{\text{data}}, -1\right)\right)$$

- Use a **richer metric** Δ for probability measures, able to handle measures with non-overlapping supports:

$$\min_{\theta\in\Theta} \Delta(\nu_{\text{data}}, p_\theta), \quad \textbf{not} \min_{\theta\in\Theta} \mathrm{KL}(\nu_{\text{data}} \| p_\theta)$$

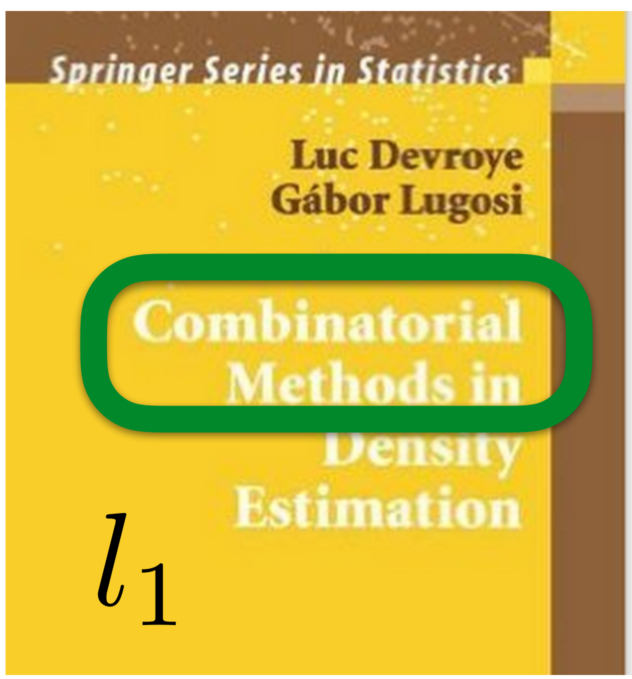
Minimum △ Estimation

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,*]

Minimum Kantorovich Estimation

- Use optimal transport theory, namely *Wasserstein distances* to define discrepancy Δ.

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

- Optimal transport? fertile field in mathematics.

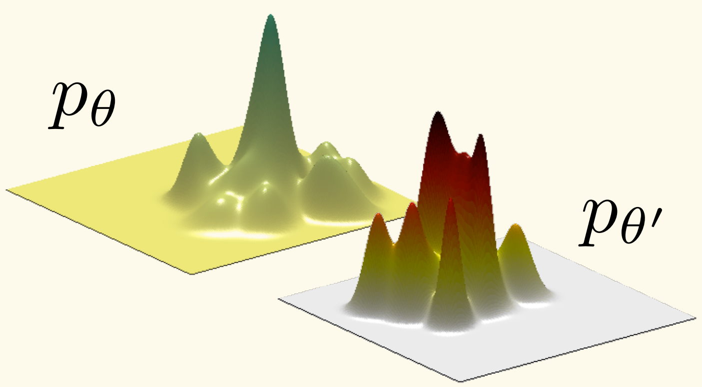
Monge Kantorovich Koopmans Dantzig Brenier Otto McCann Villani

Nobel '75

Fields '10

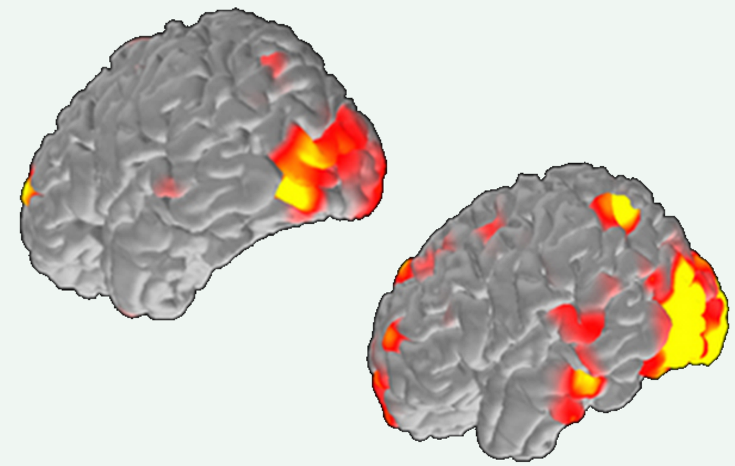
What is Optimal Transport?

A geometric toolbox to
compare **probability measures**
supported on a metric space.



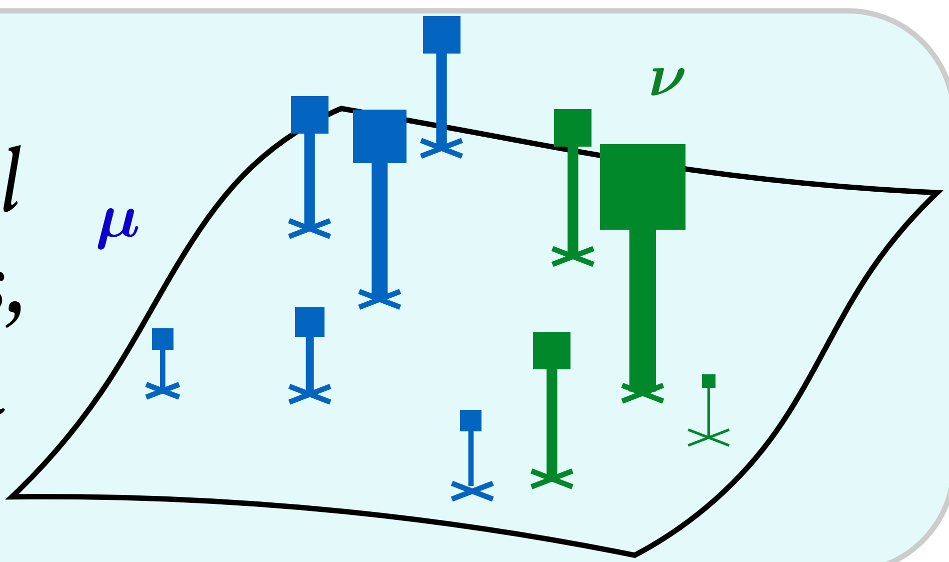
Statistical Models

Bags of features



Brain Activation Maps

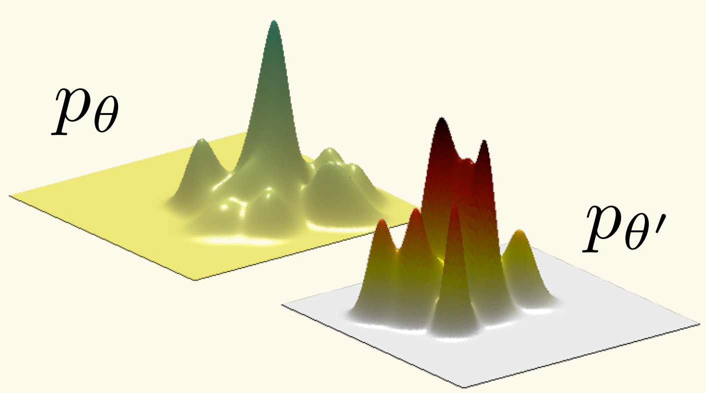
Empirical Measures, i.e. **data**



Color Histograms

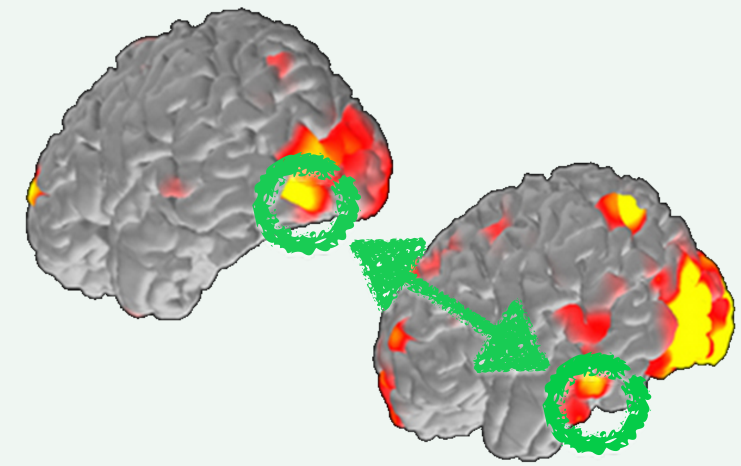
What is Optimal Transport?

A geometric toolbox to
compare probability measures
supported on a **metric space**.



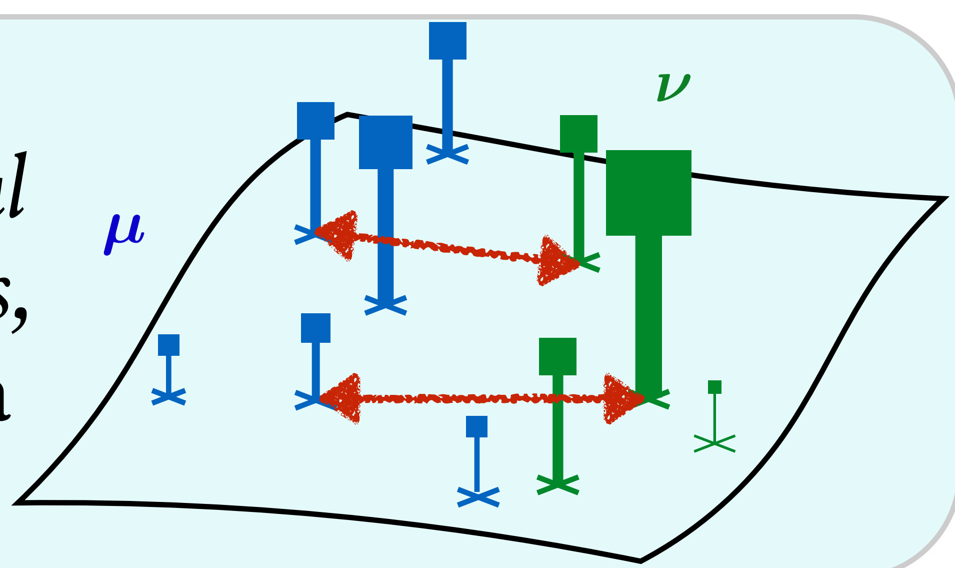
Statistical Models

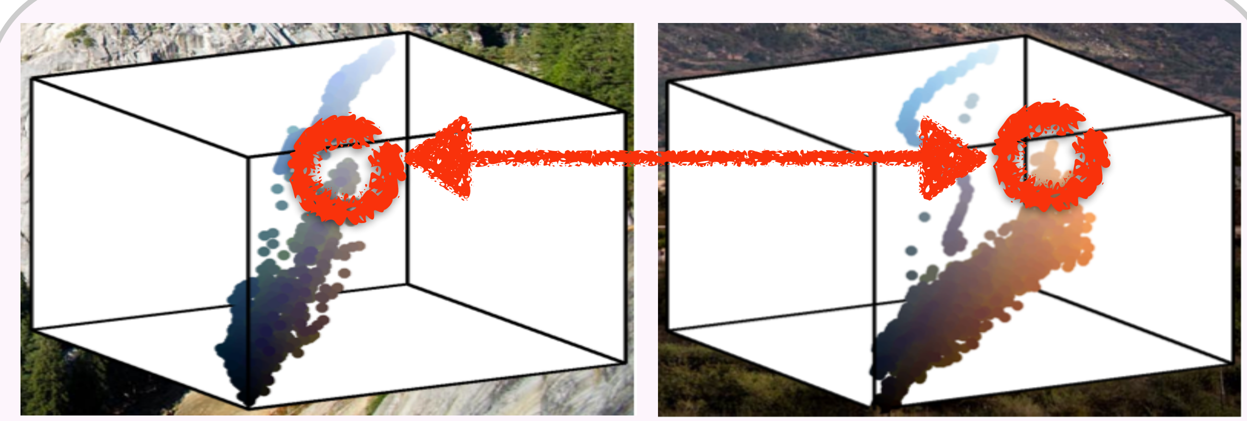
Bags of features



Brain Activation Maps

Empirical Measures, i.e. **data**

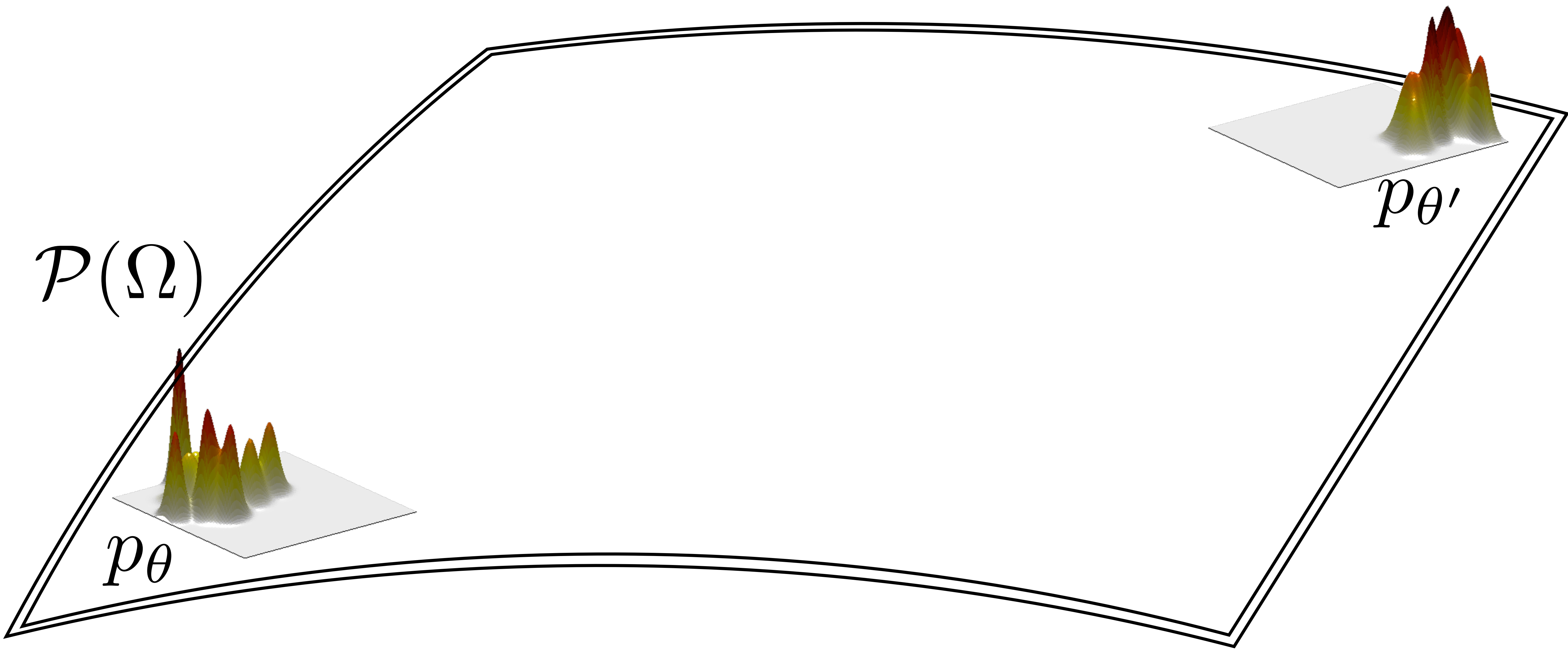




Color Histograms

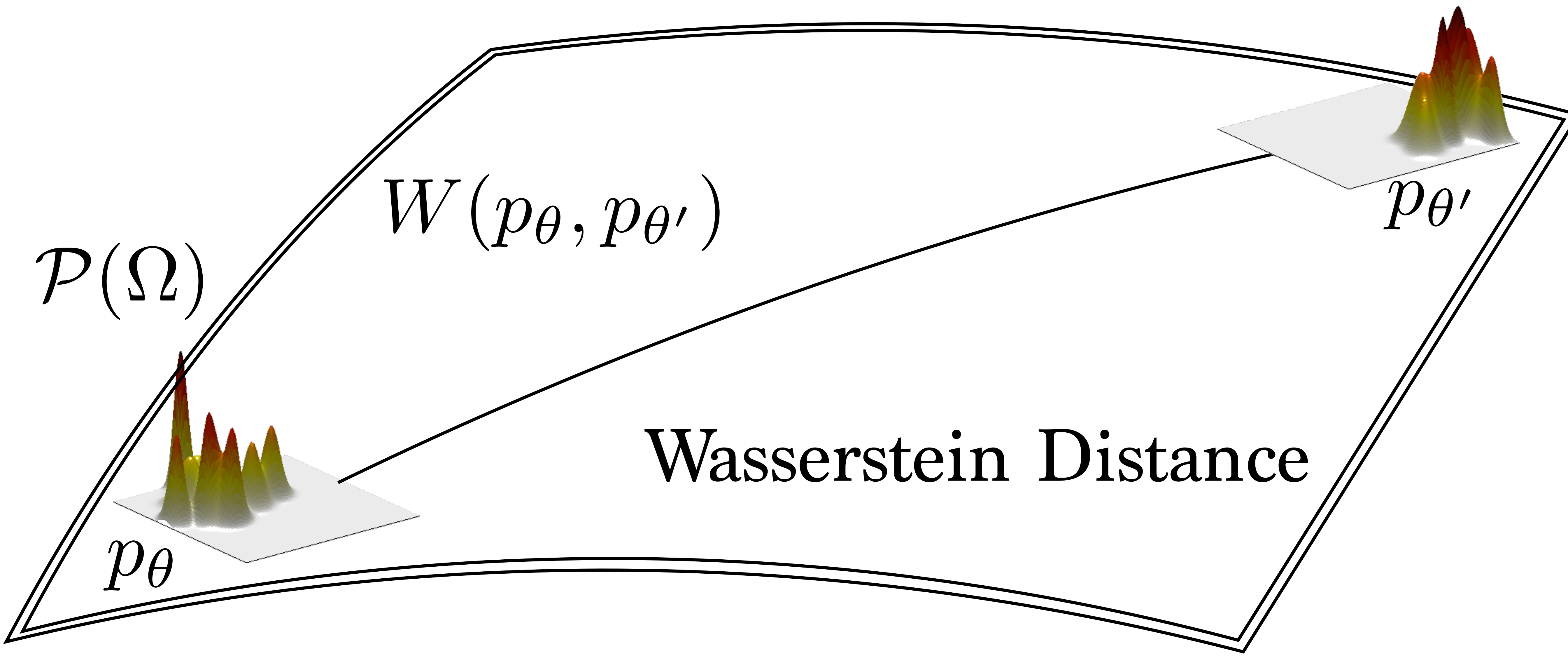
Optimal Transport Geometry

A **geometric toolbox** to
compare probability measures
supported on a metric space.



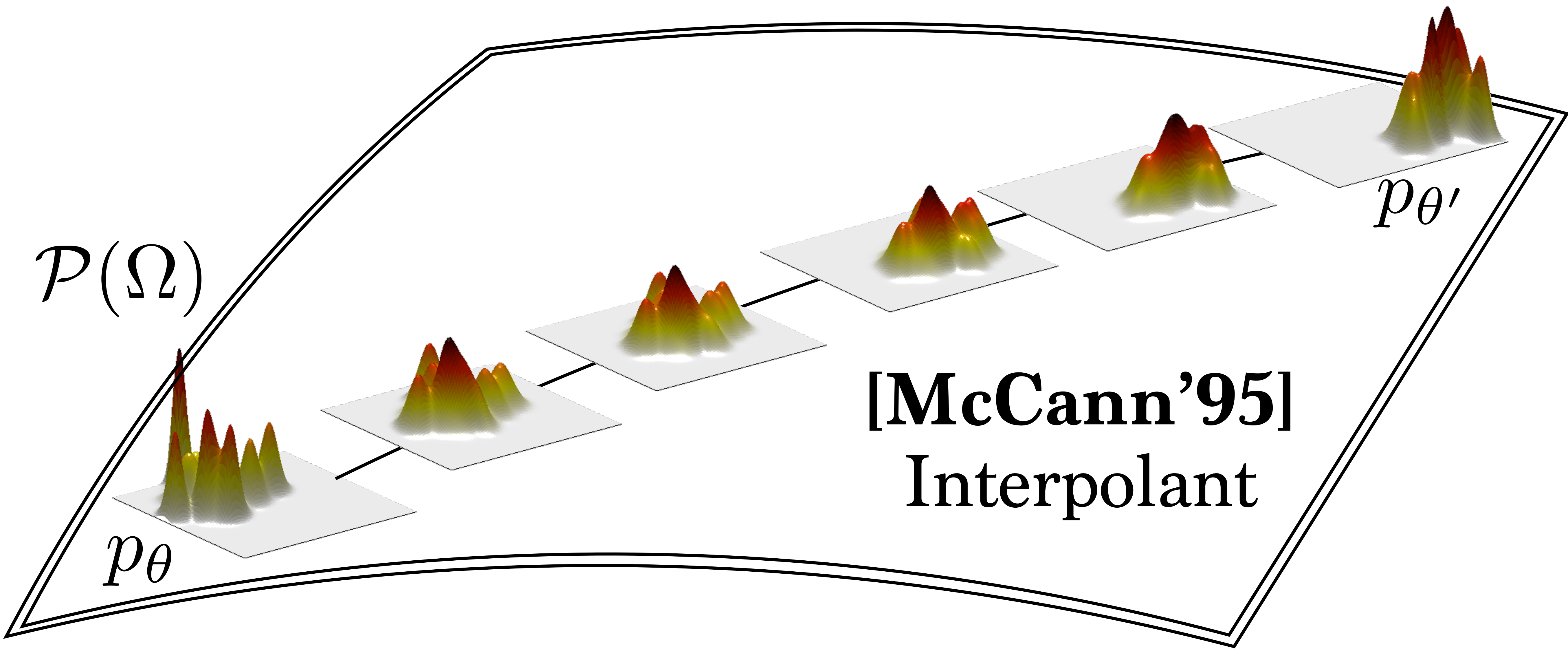
Optimal Transport Geometry

A **geometric toolbox** to compare probability measures supported on a metric space.



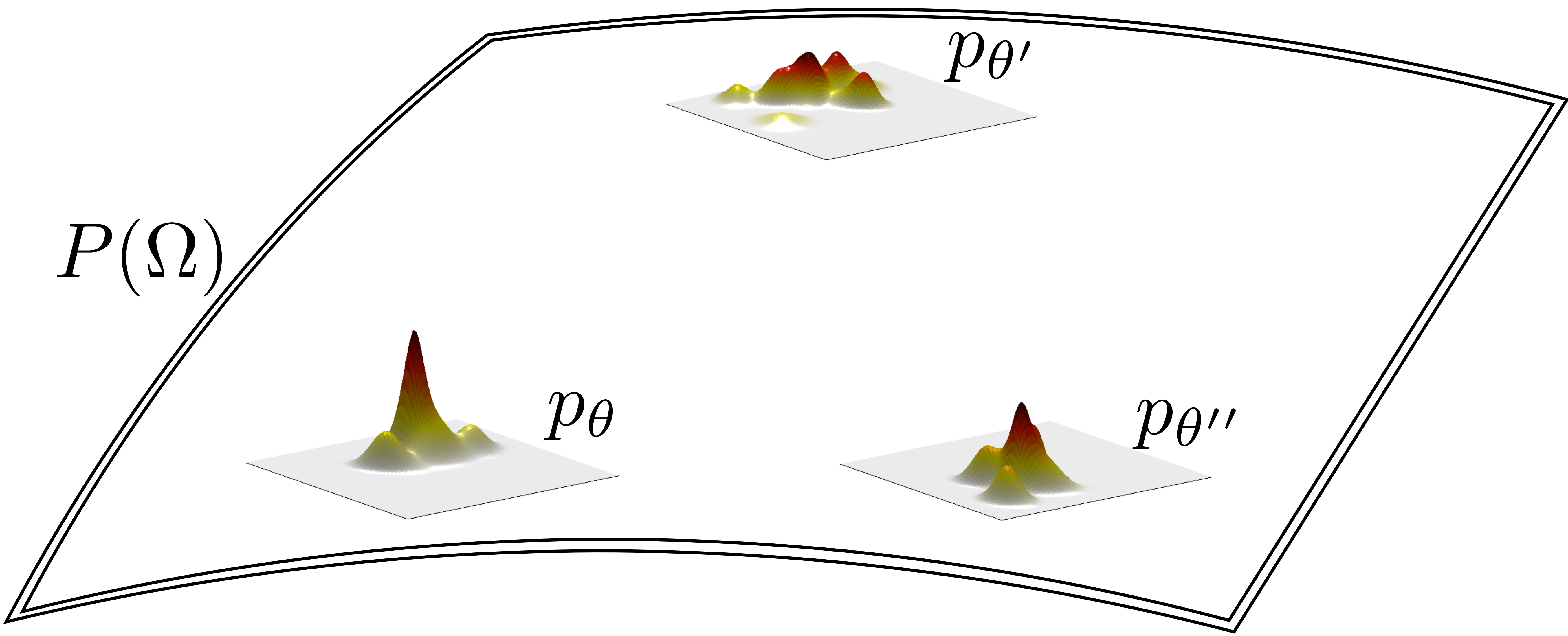
Optimal Transport Geometry

A **geometric toolbox** to
compare probability measures
supported on a metric space.



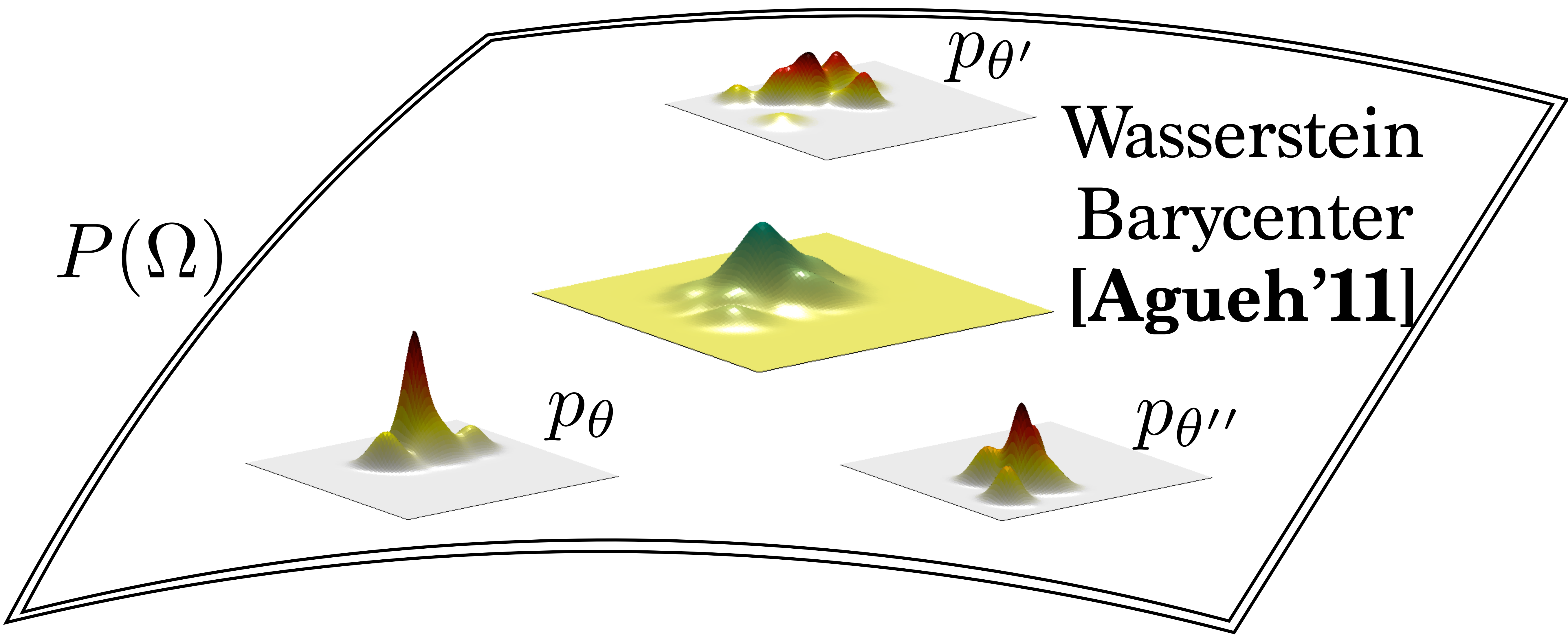
Optimal Transport Geometry

A **geometric toolbox** to compare probability measures supported on a metric space.



Optimal Transport Geometry

A **geometric toolbox** to compare probability measures supported on a metric space.



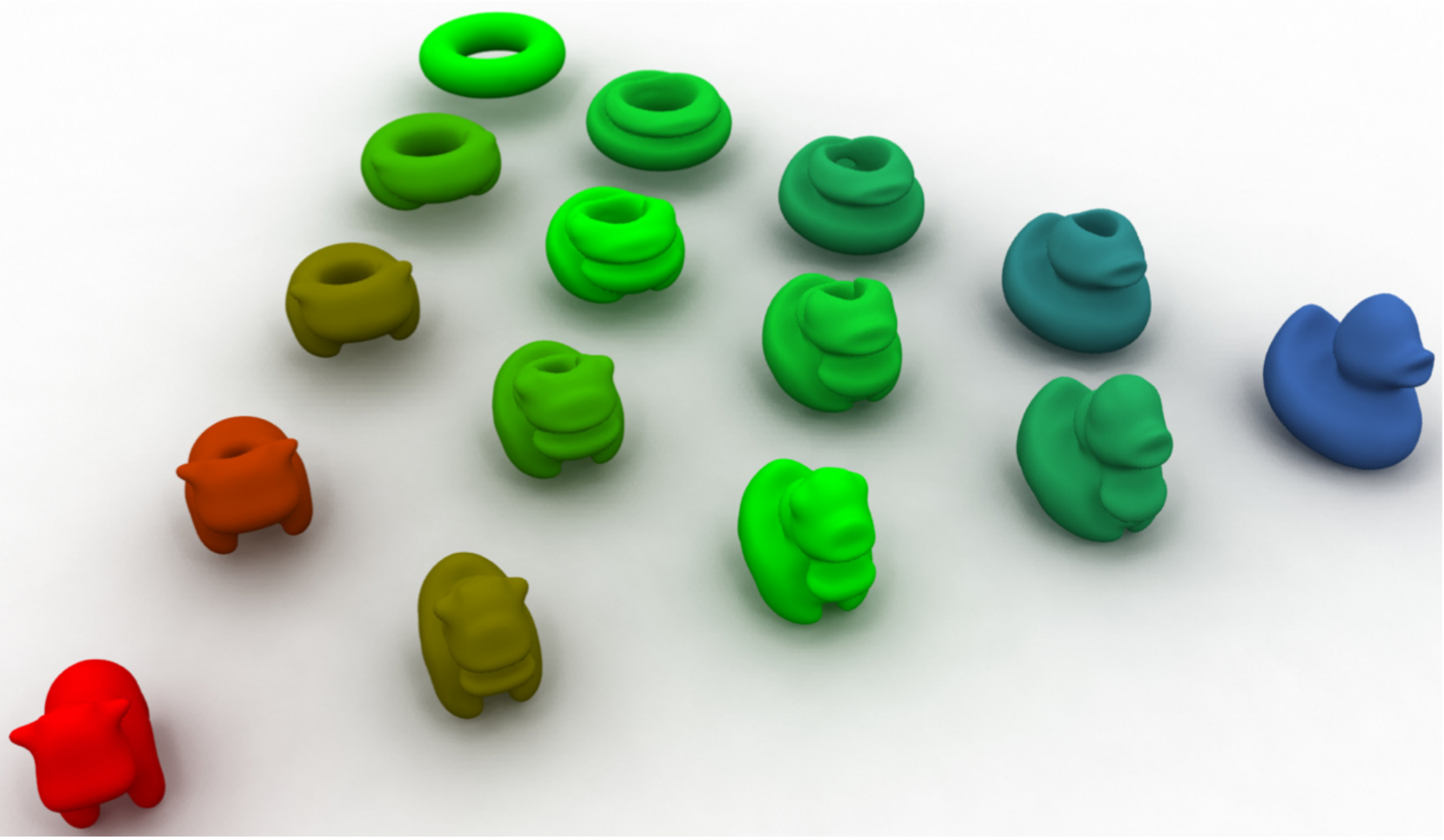
Optimal Transport Geometry

A **geometric toolbox** to compare probability measures supported on a metric space.

[SDPC..'15]

Optimal Transport Geometry

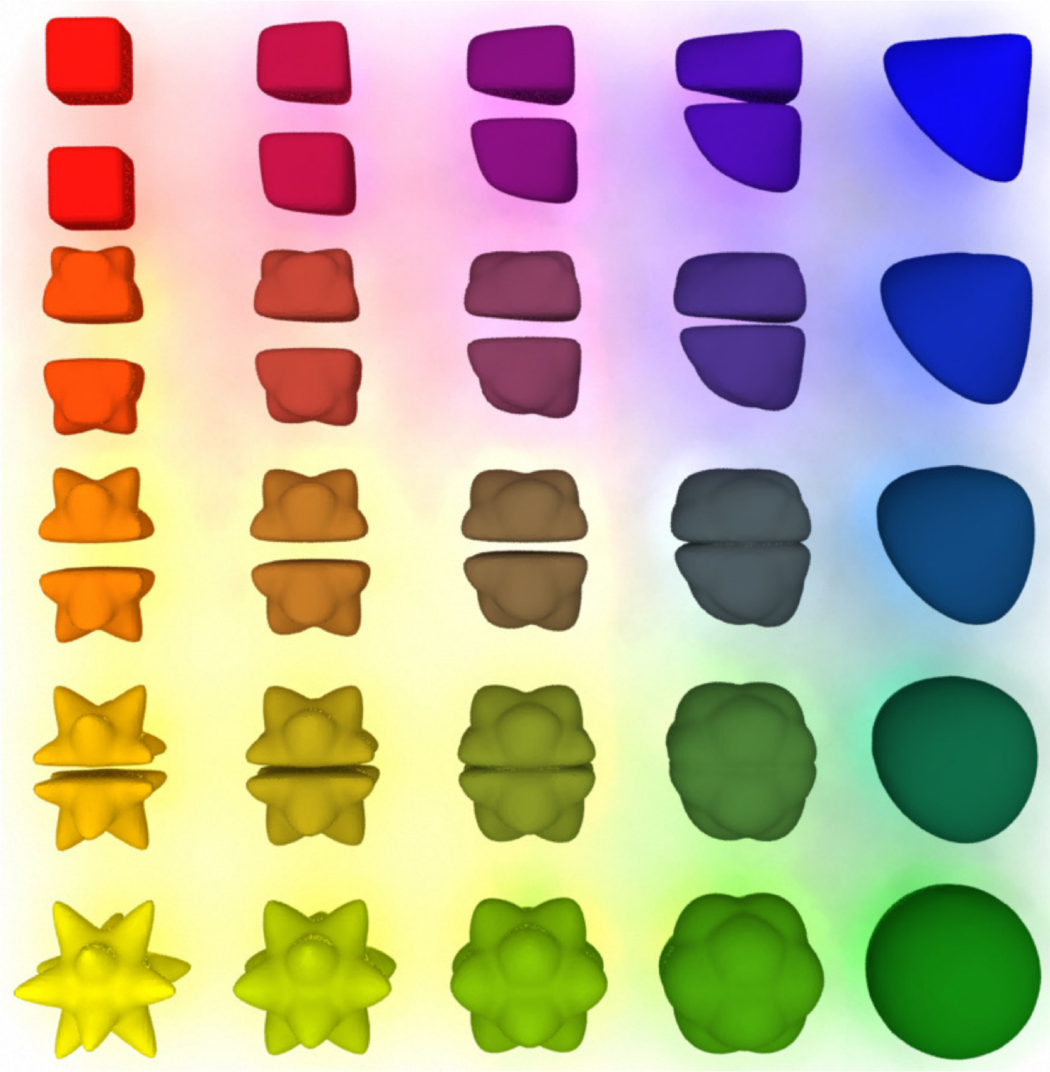
A **geometric toolbox** to
compare probability measures
supported on a metric space.



[SDPC..'15]

Optimal Transport Geometry

A **geometric toolbox** to
compare probability measures
supported on a metric space.



[SDPC..'15]

Origins: Monge's Problem

666. MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

THÉORIE DES DÉBLAIS

ET DES REMBLAIS.

Par M. MONGE.

LORSQU'ON doit tranſporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranſporter, & le nom de *Remblai* à l'eſpace qu'elles doivent occuper après le tranſport.

Origins: Monge's Problem

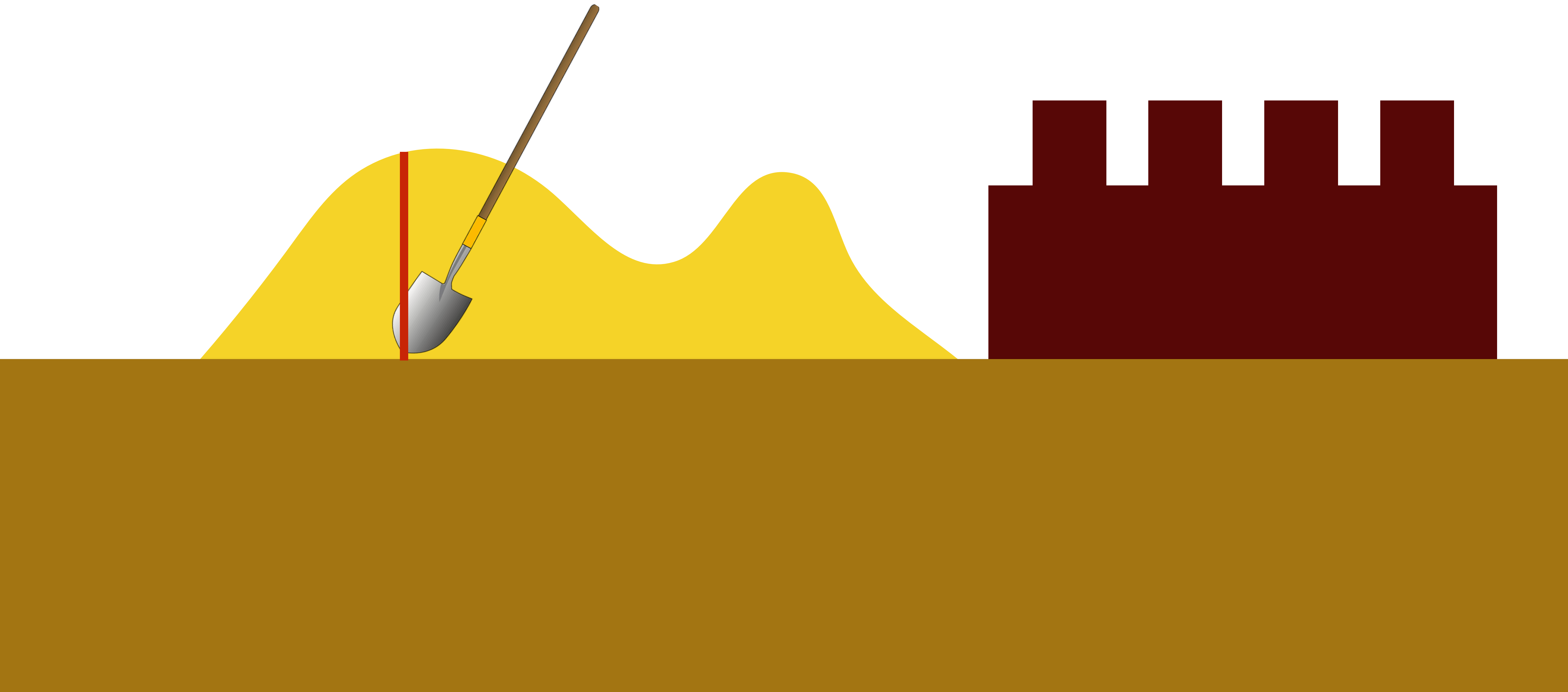
Origins: Monge's Problem

Origins: Monge's Problem

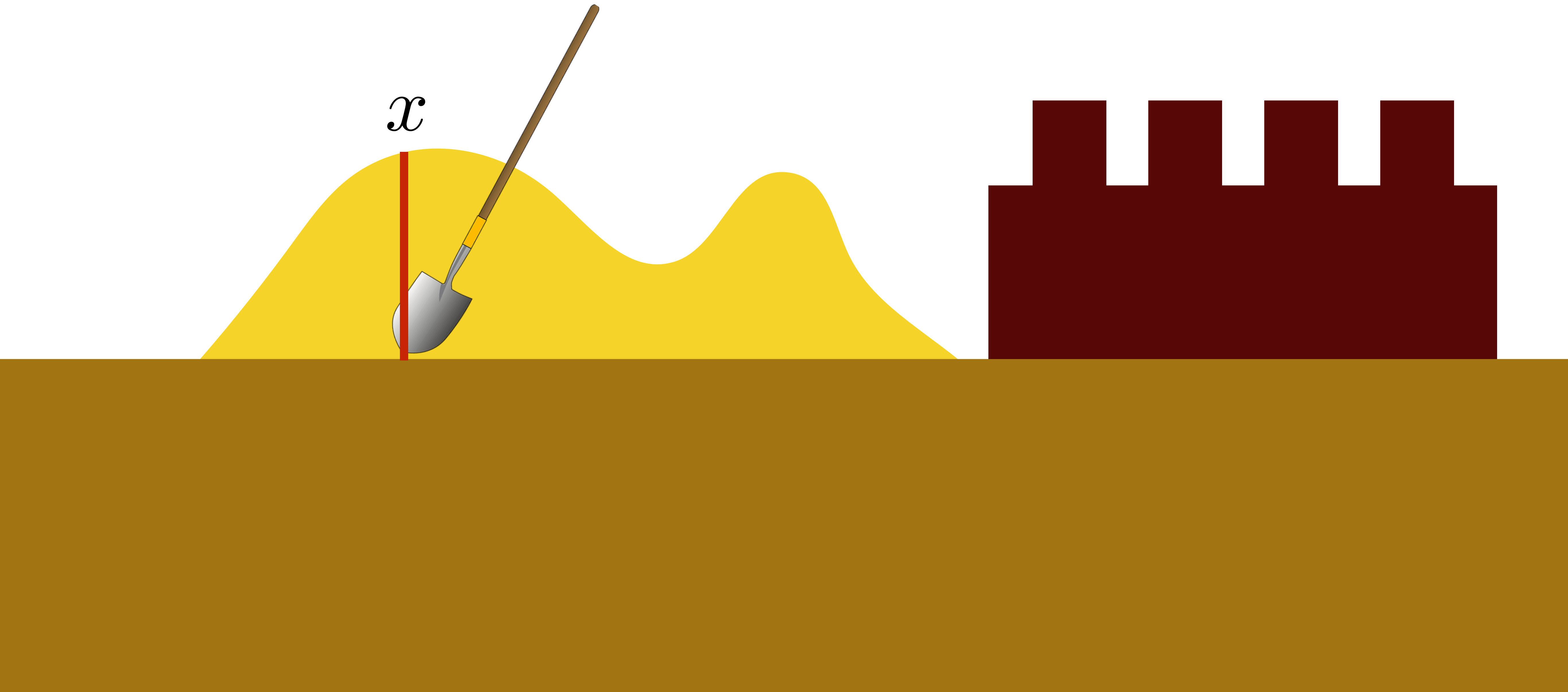
Origins: Monge's Problem

Origins: Monge's Problem

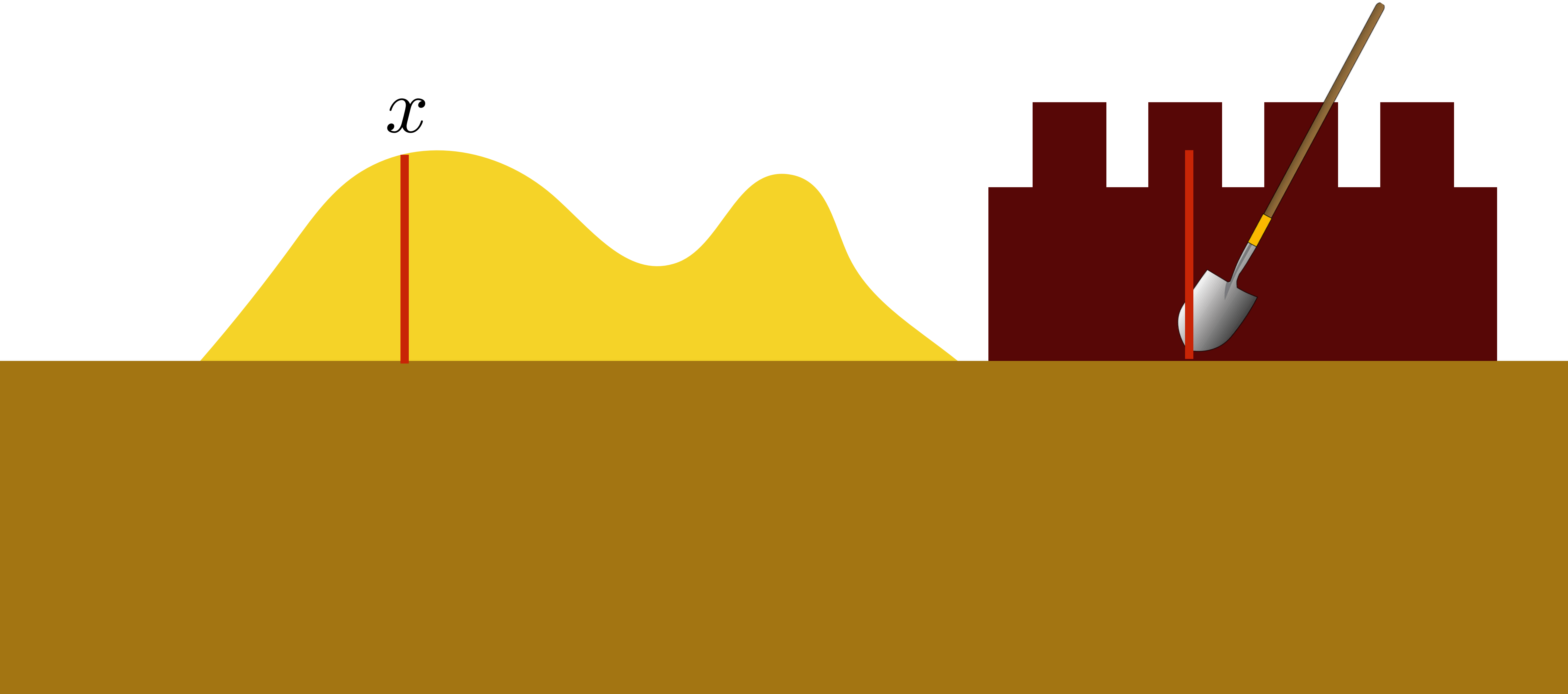
Origins: Monge's Problem



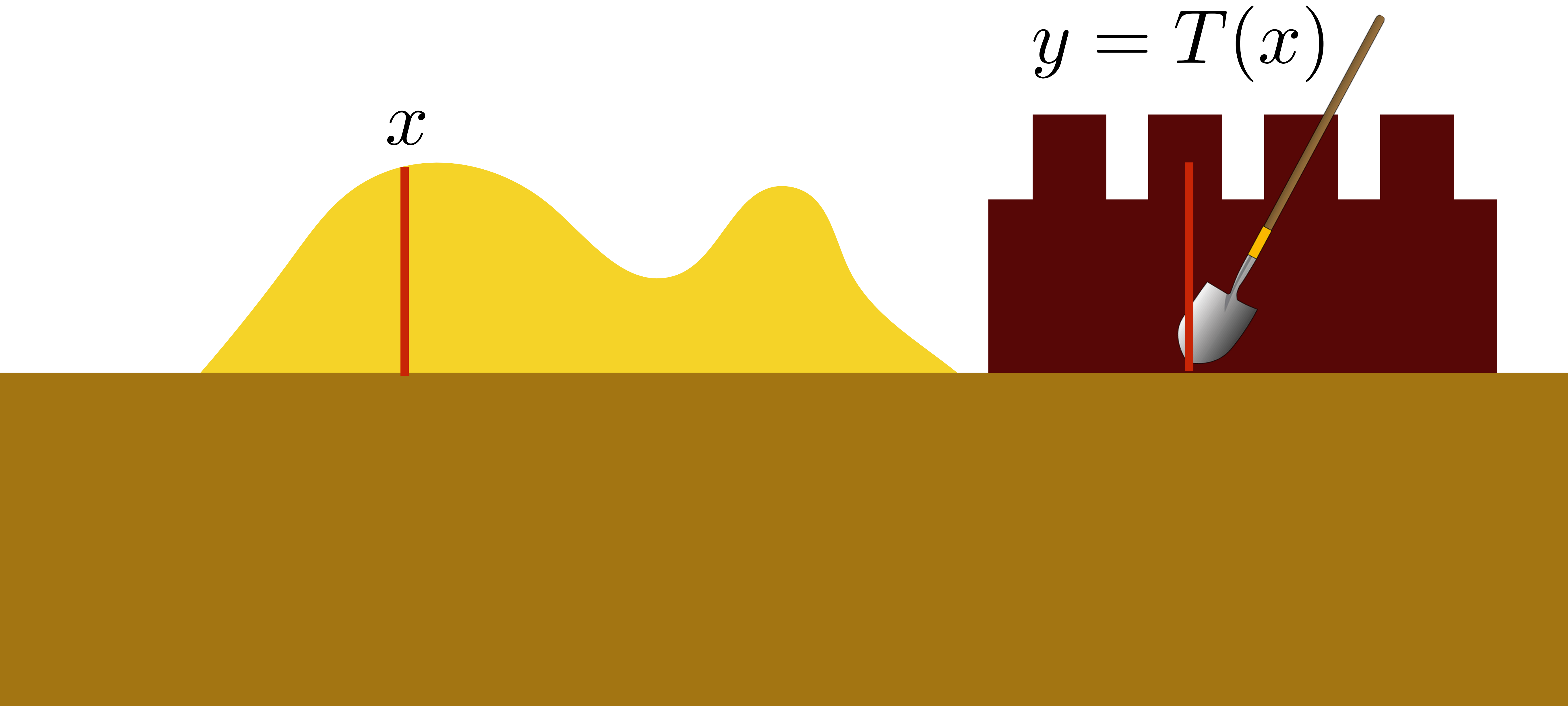
Origins: Monge's Problem



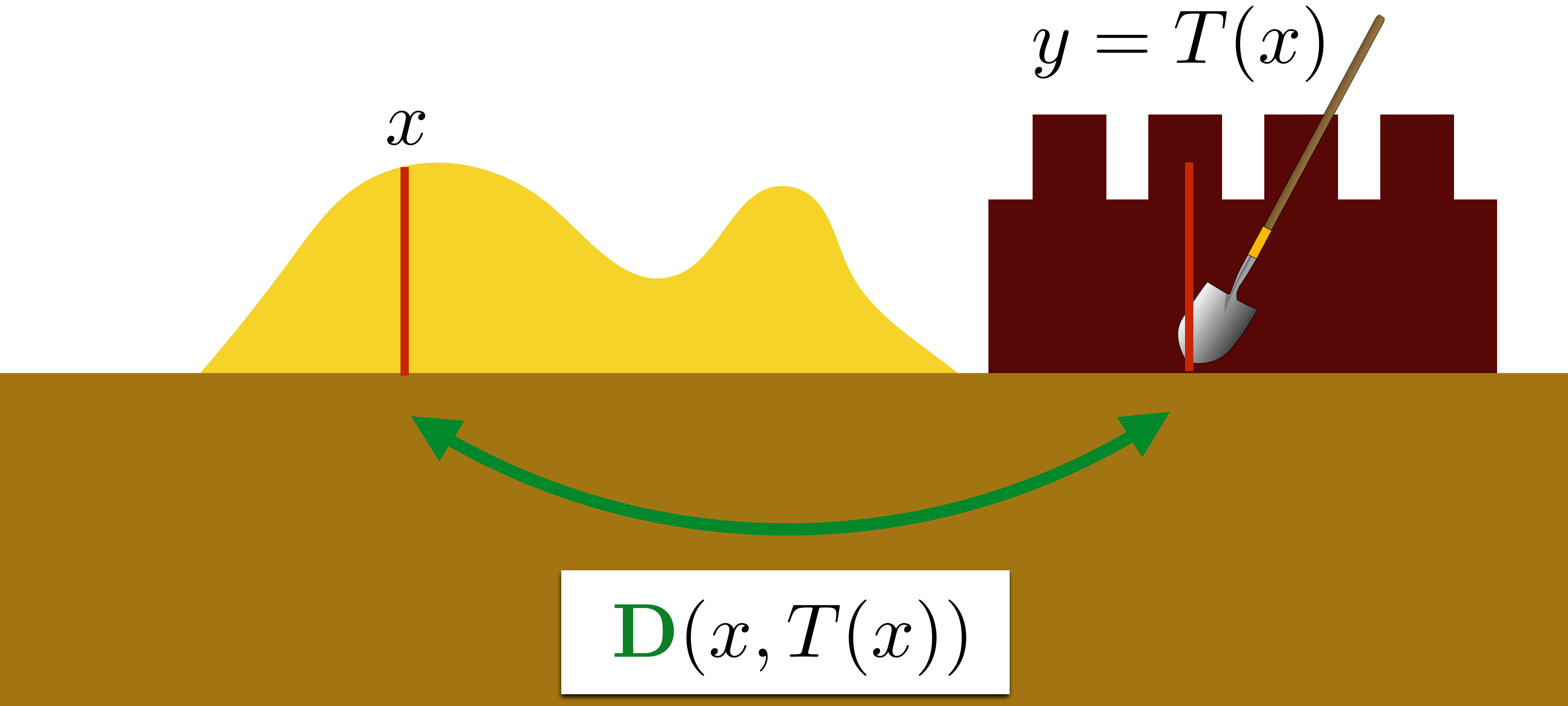
Origins: Monge's Problem



Origins: Monge's Problem



Origins: Monge's Problem

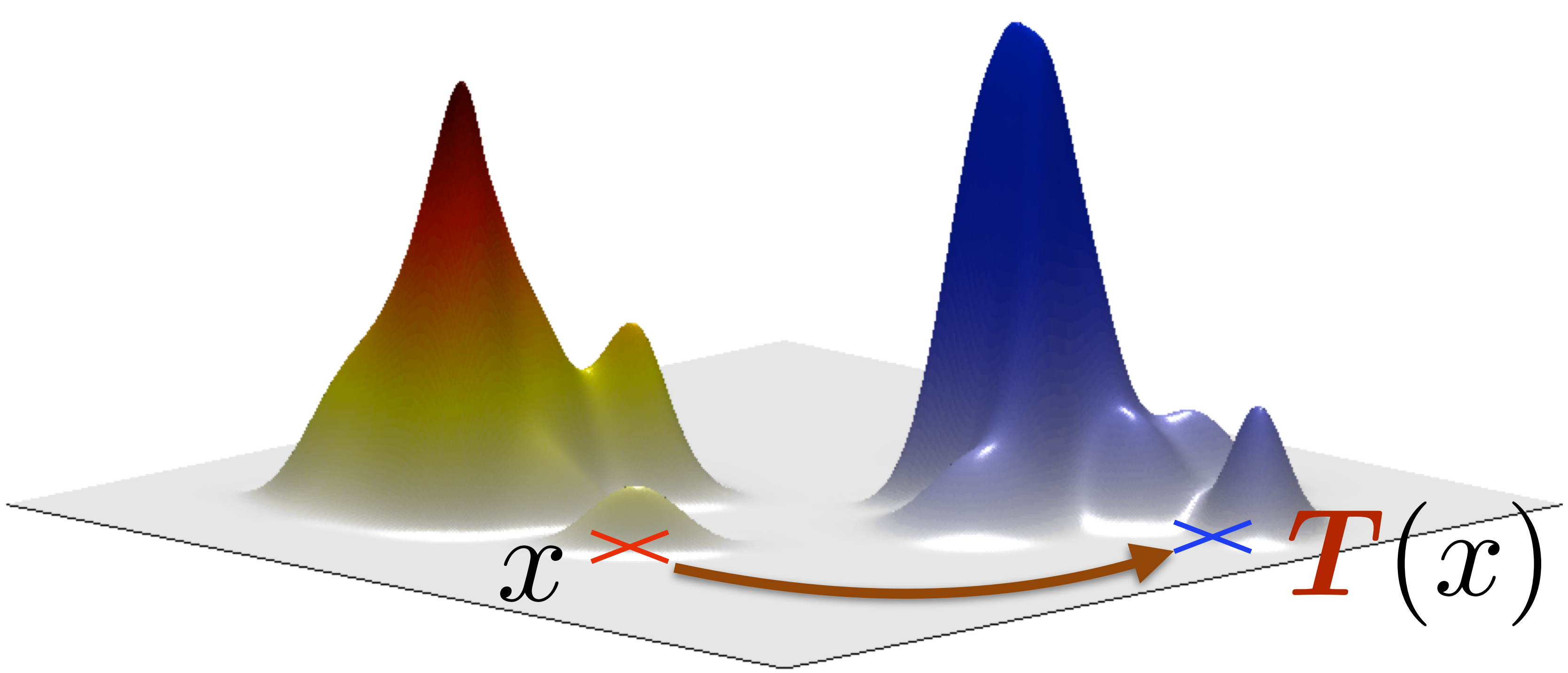


Origins: Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

$$\inf_{\boldsymbol{T}_\sharp \boldsymbol{\mu} = \boldsymbol{\nu}} \int_\Omega \boldsymbol{c}(x, \boldsymbol{T}(x))\boldsymbol{\mu}(dx)$$

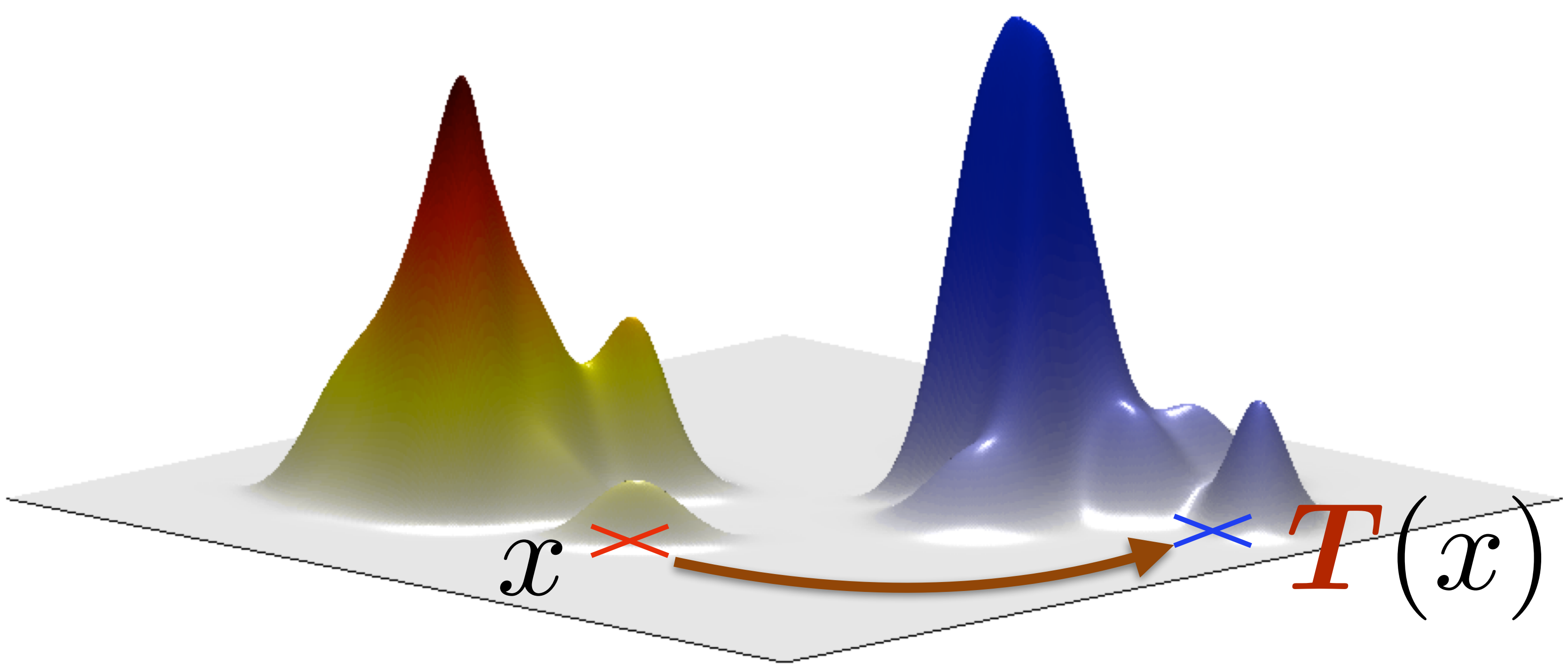


Origins: Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

[Brenier'87] If $\Omega = \mathbb{R}^d$, $\boldsymbol{c} = \|\cdot - \cdot\|^2$, $\boldsymbol{\mu}, \boldsymbol{\nu}$ a.c., then $\boldsymbol{T} = \nabla \boldsymbol{u}$, $\boldsymbol{u}$ **convex**.

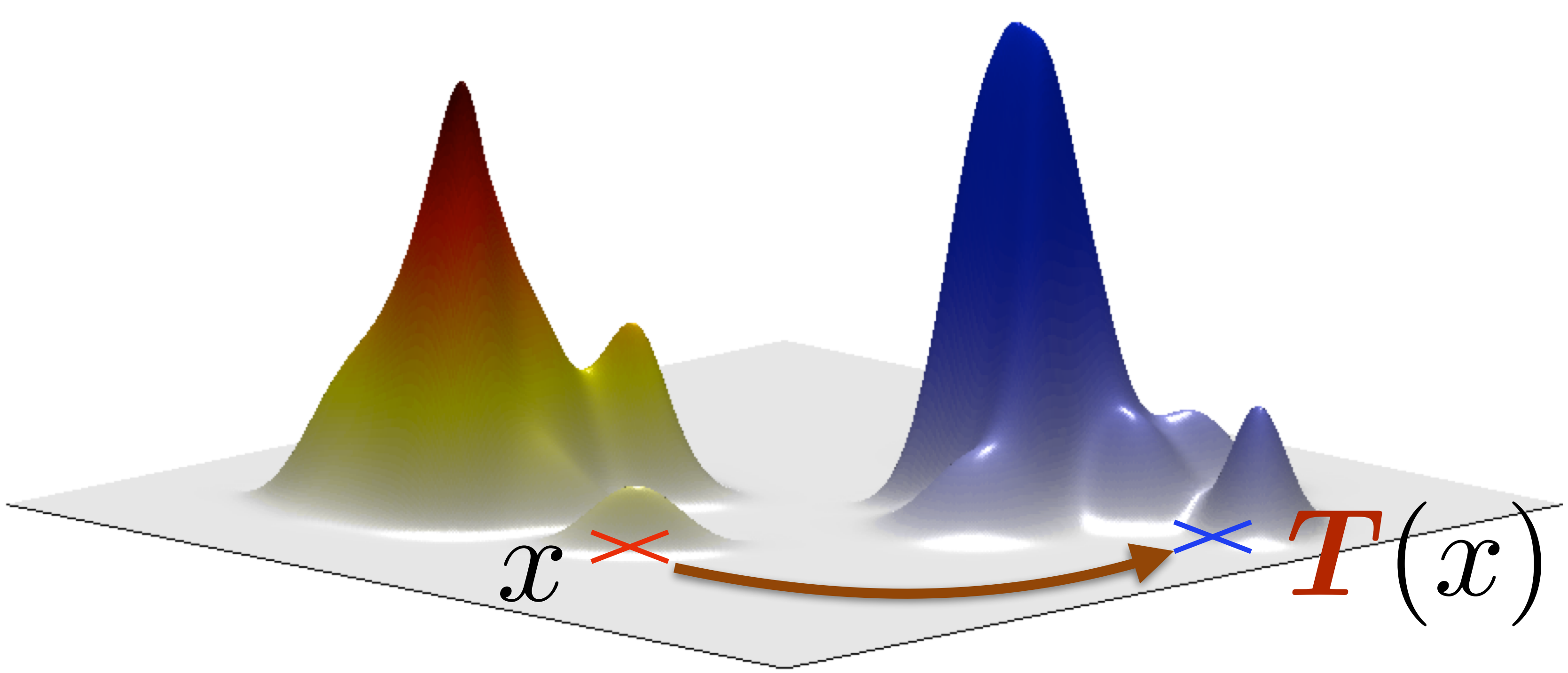


Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

$$\inf_{\boldsymbol{T}_\sharp \boldsymbol{\mu} = \boldsymbol{\nu}} \int_\Omega \boldsymbol{c}(x, \boldsymbol{T}(x))\boldsymbol{\mu}(dx)$$

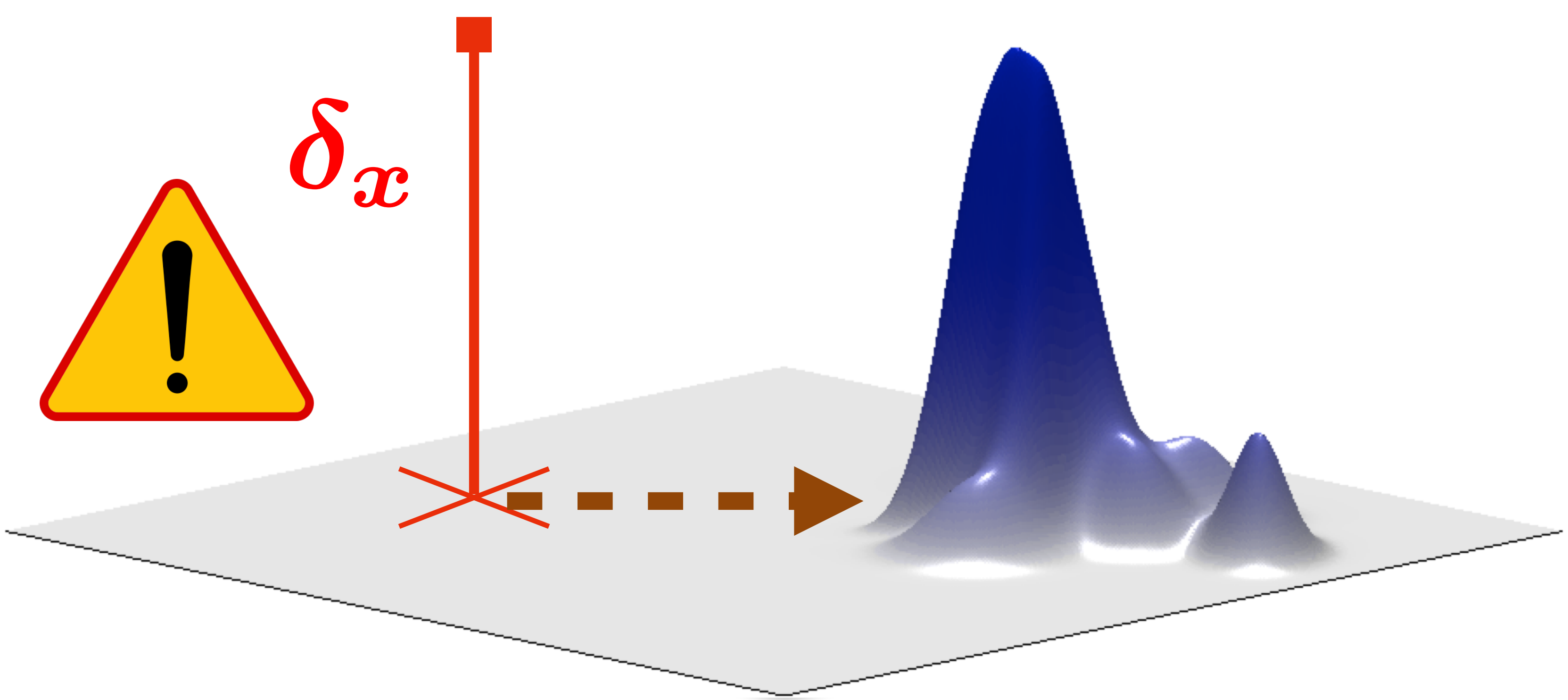


Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

$$\inf_{\boldsymbol{T}_\sharp \boldsymbol{\mu} = \boldsymbol{\nu}} \int_\Omega \boldsymbol{c}(x, \boldsymbol{T}(x)) \boldsymbol{\mu}(dx)$$



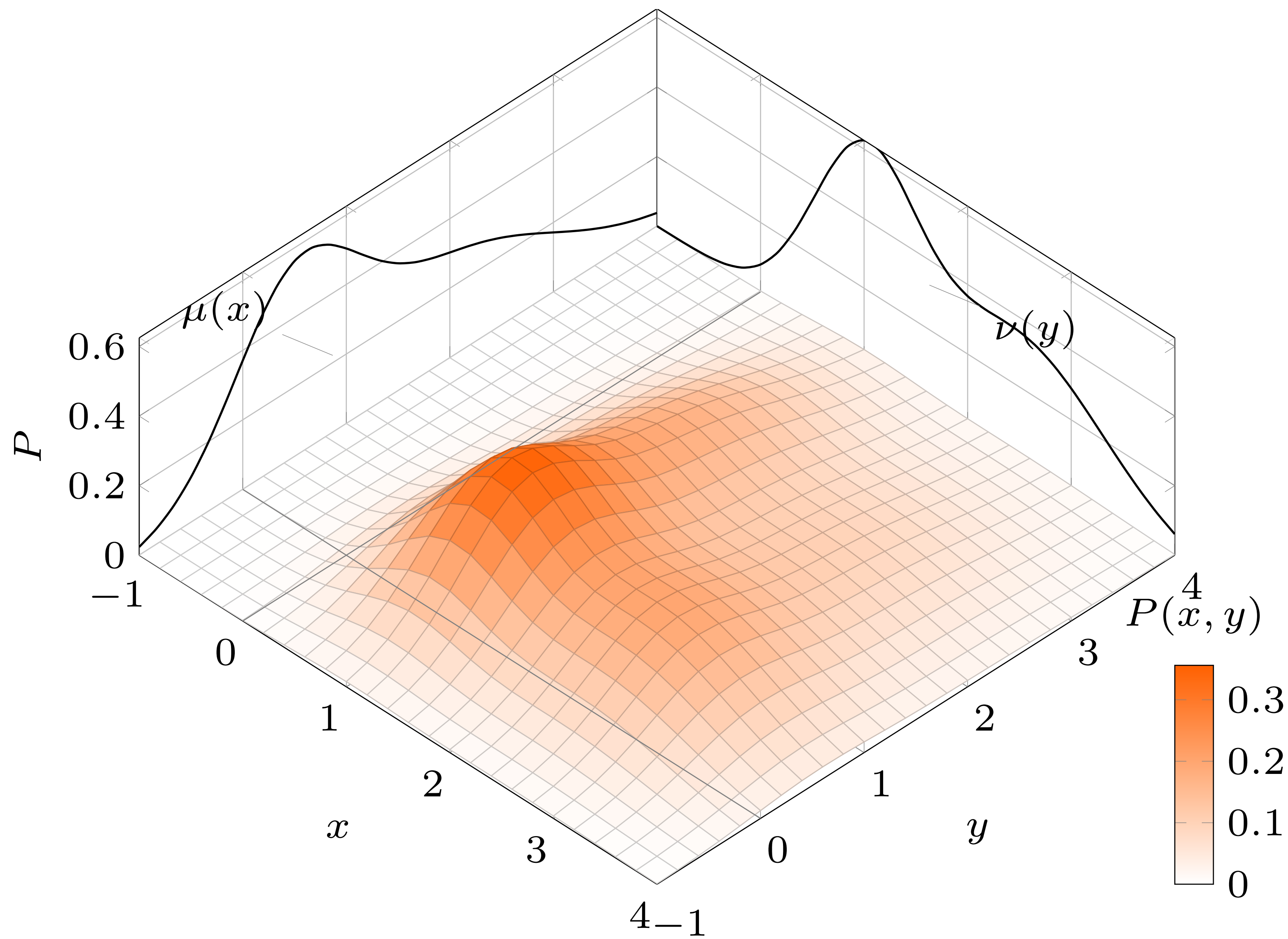
[Kantorovich'42] Relaxation

- Instead of maps $\boldsymbol{T} : \Omega \to \Omega$, consider probabilistic maps, i.e. **couplings** $\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega)$:

$$\boxed{\begin{gathered}\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) \,|\, \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega, \\ \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \\ \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}\end{gathered}}$$

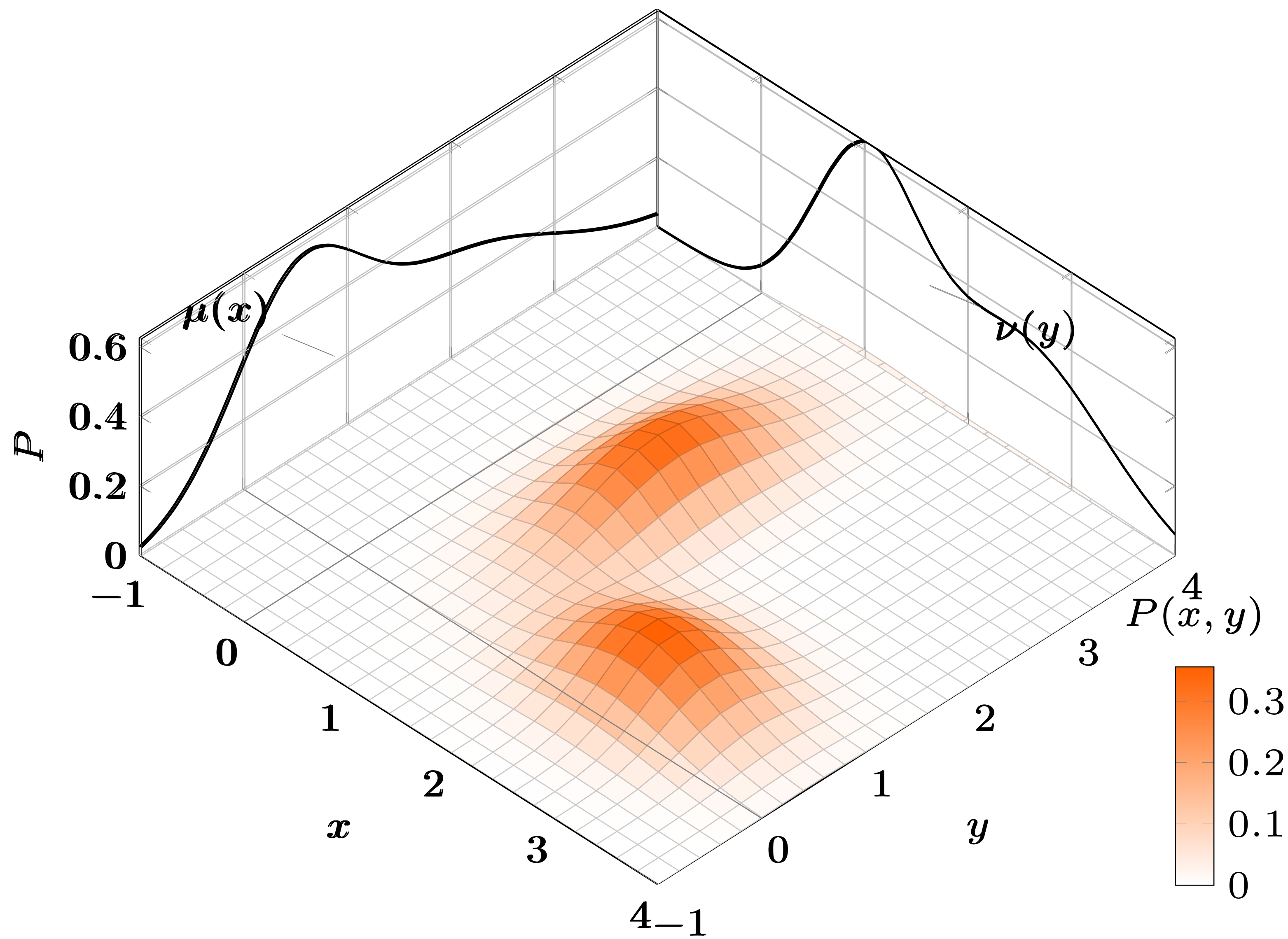
[Kantorovich'42] Relaxation

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) \,|\, \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega, \\ \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$



[Kantorovich'42] Relaxation

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) | \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega,$$
$$\boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$



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,

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy).$$

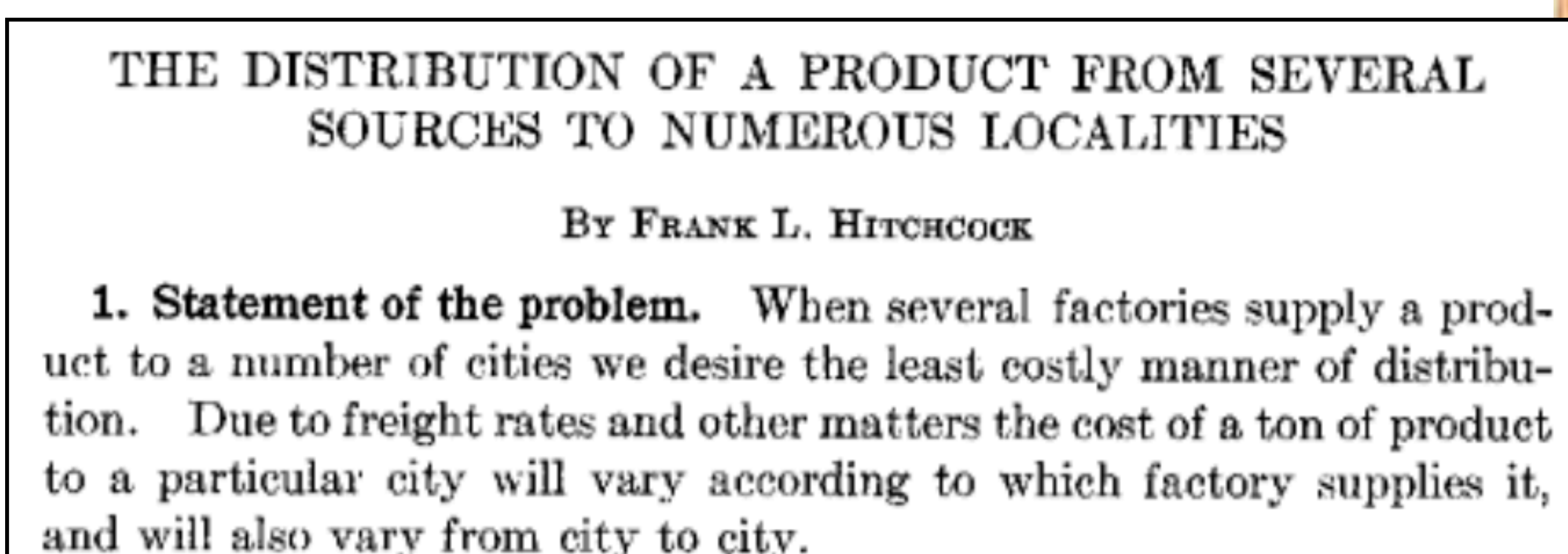
PRIMAL

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,

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy).$$

PRIMAL



THE DISTRIBUTION OF A PRODUCT FROM SEVERAL SOURCES TO NUMEROUS LOCALITIES

BY FRANK L. HITCHCOCK

1. Statement of the problem. When several factories supply a product to a number of cities we desire the least costly manner of distribution. Due to freight rates and other matters the cost of a ton of product to a particular city will vary according to which factory supplies it, and will also vary from city to city.

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,

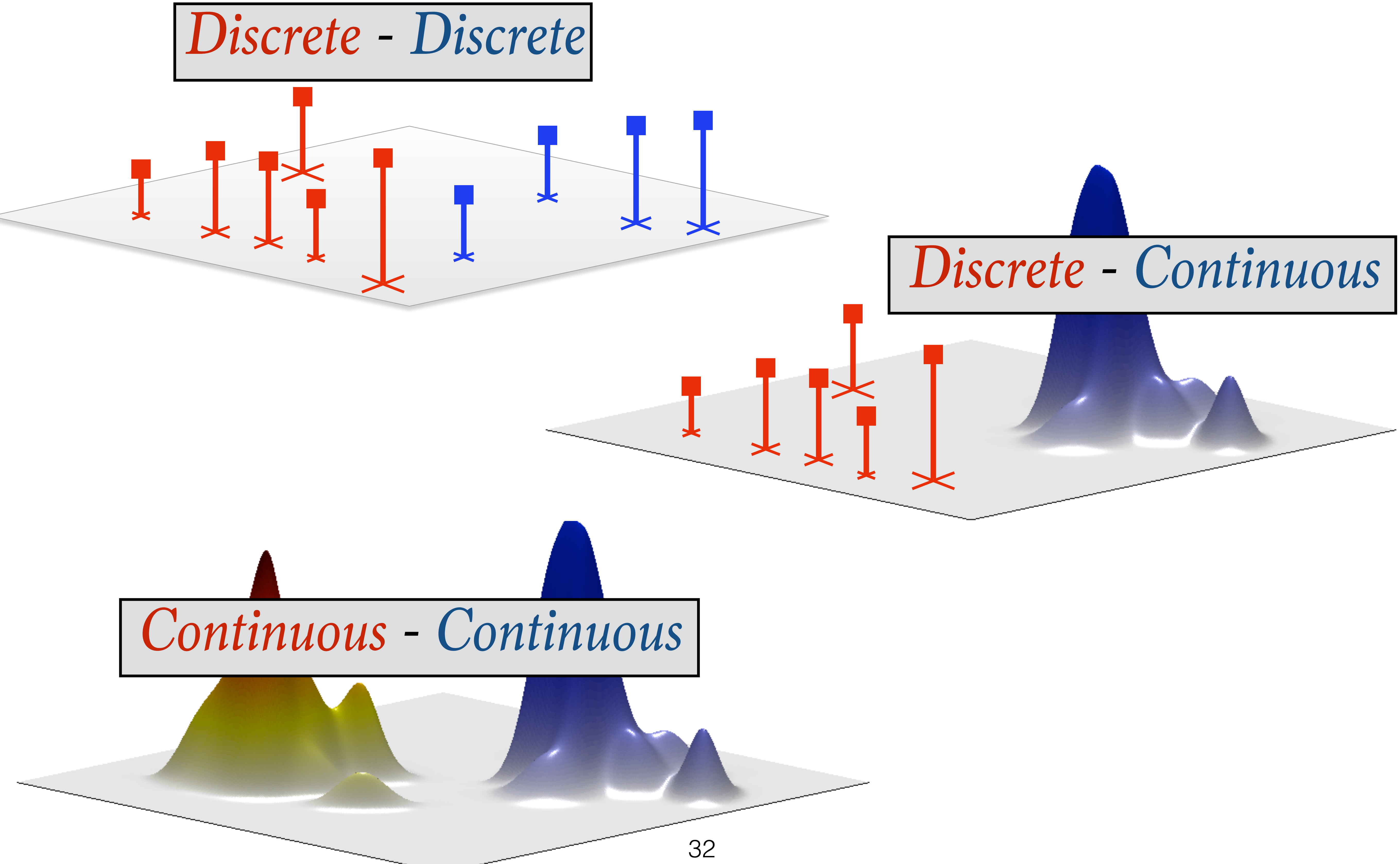
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy).$$

PRIMAL

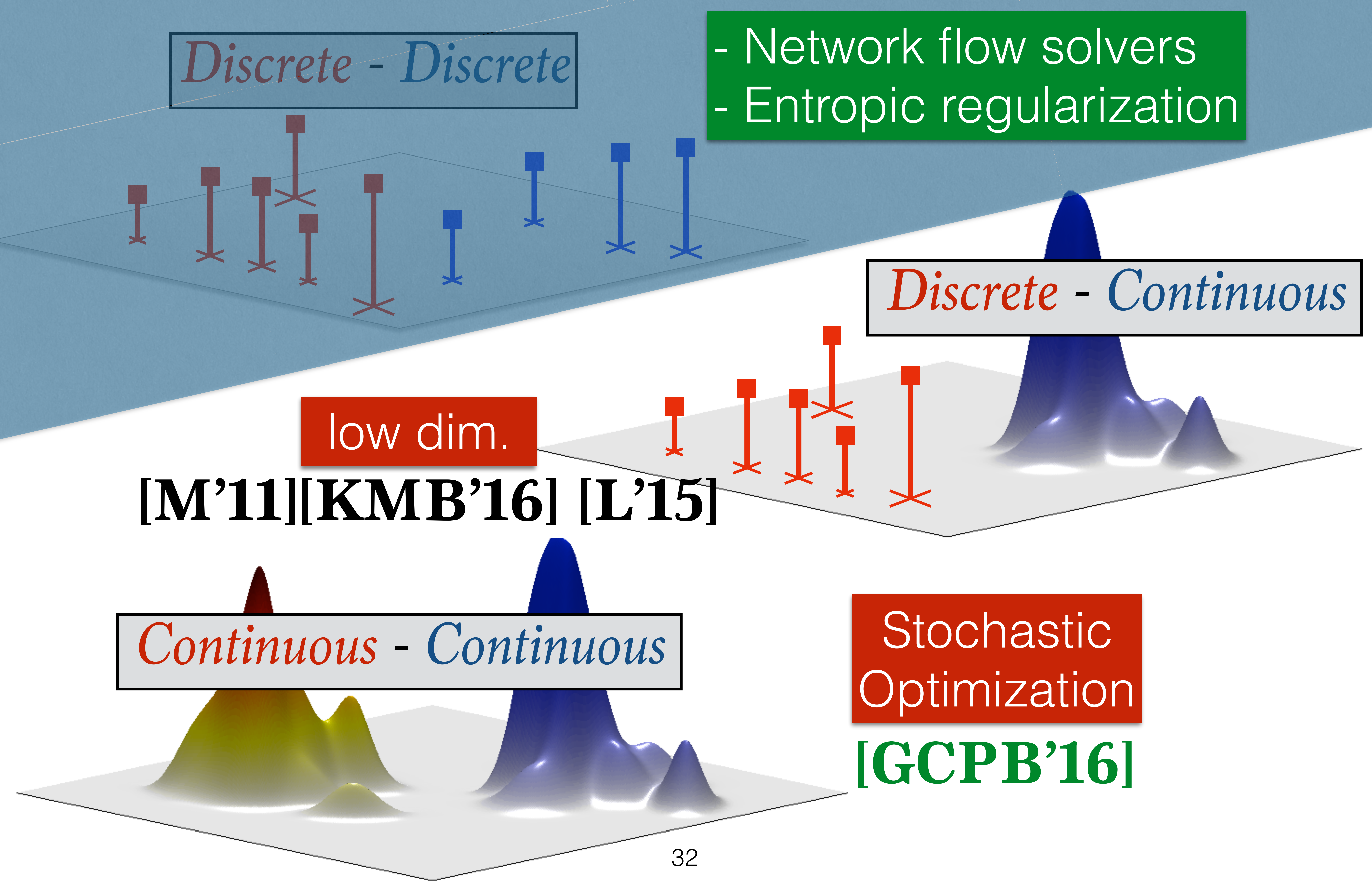
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{D}^p(x, y)}} \int \boldsymbol{\varphi} \, d\boldsymbol{\mu} + \int \boldsymbol{\psi}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

DUAL

W is versatile



W is versatile

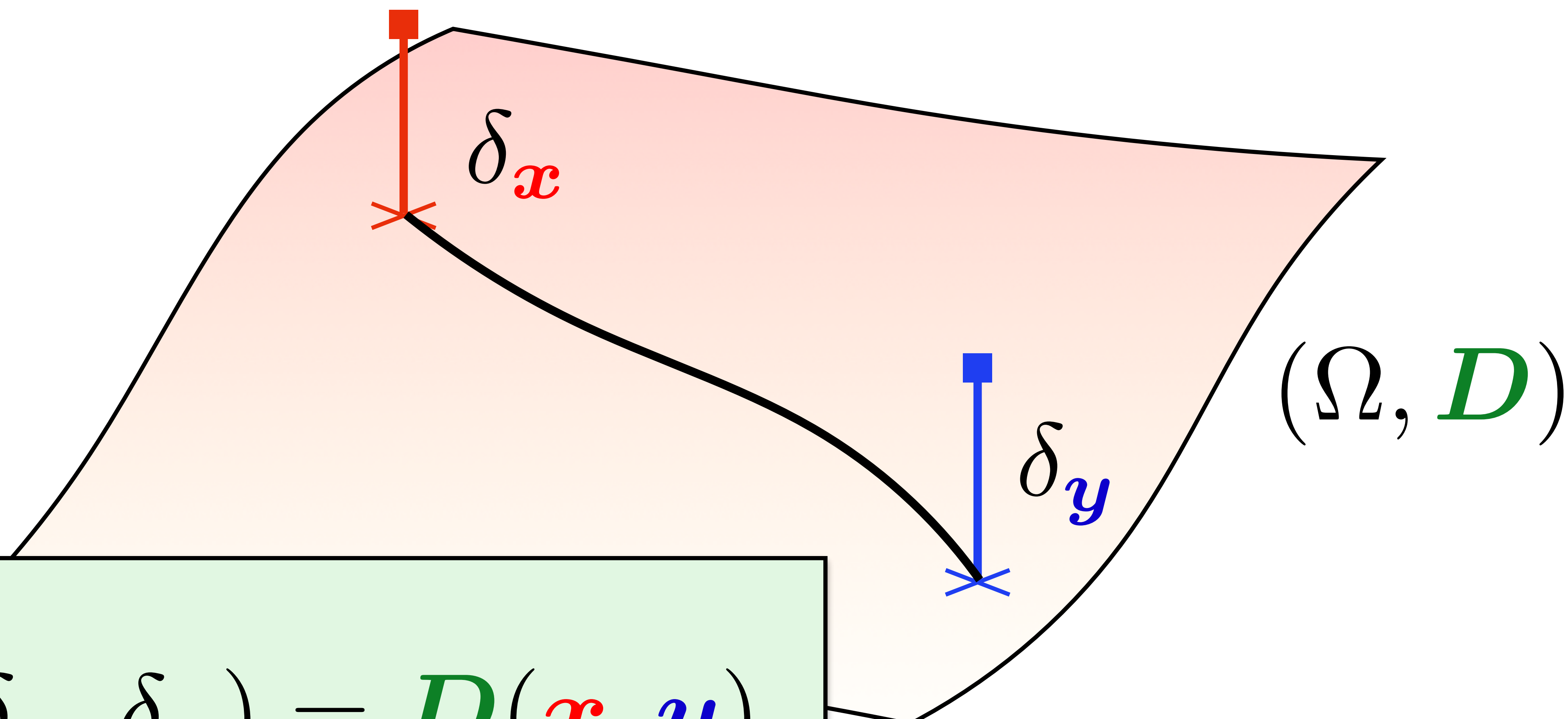


Minimum Kantorovich Estimators

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

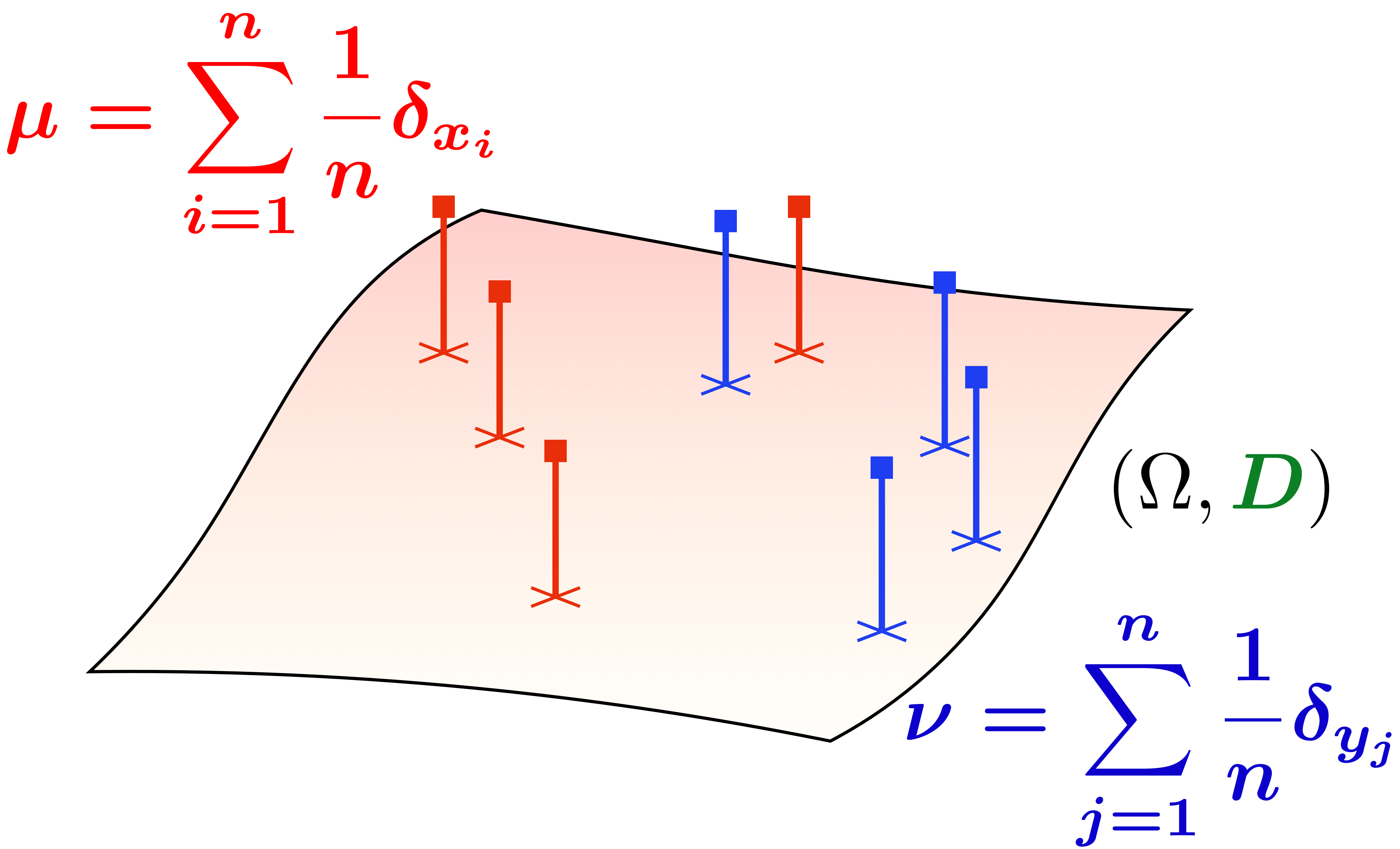
- **[Bassetti'06]** 1st reference discussing this approach.
- **[MMC'16]** use regularization in a finite setting.
- **[ACB'17]** *(WGAN)* **[BJGR'17]** *(Wasserstein ABC).*
- **Hot topics:** *approximate & differentiate W efficiently.*
- Today: ideas from our recent preprint **[GPC'17]**

Wasserstein between 2 Diracs



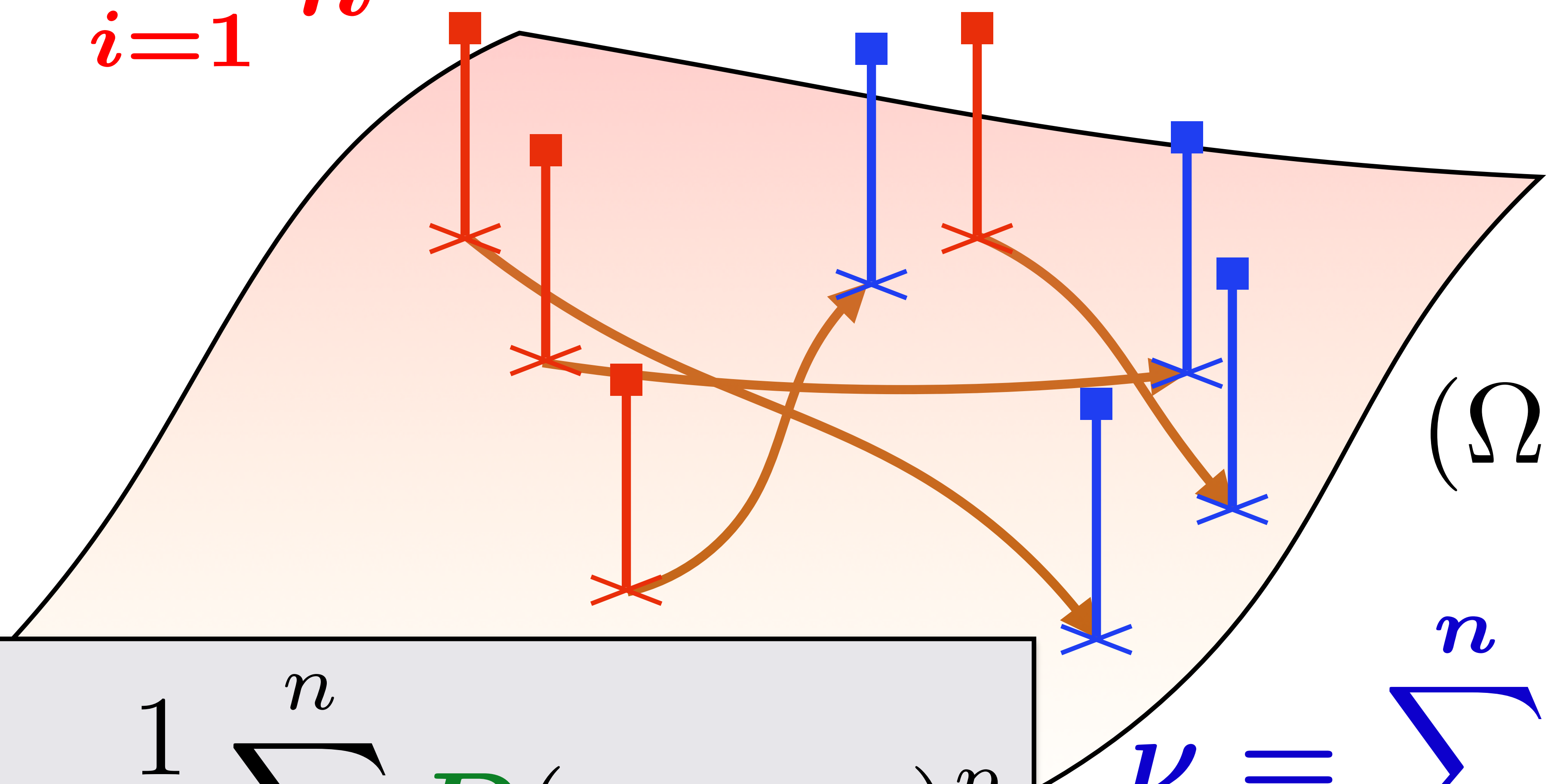
$$W_p^p(\delta_{\boldsymbol{x}}, \delta_{\boldsymbol{y}}) = \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})$$

Wasserstein on Uniform Measures



Wasserstein on Uniform Measures

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



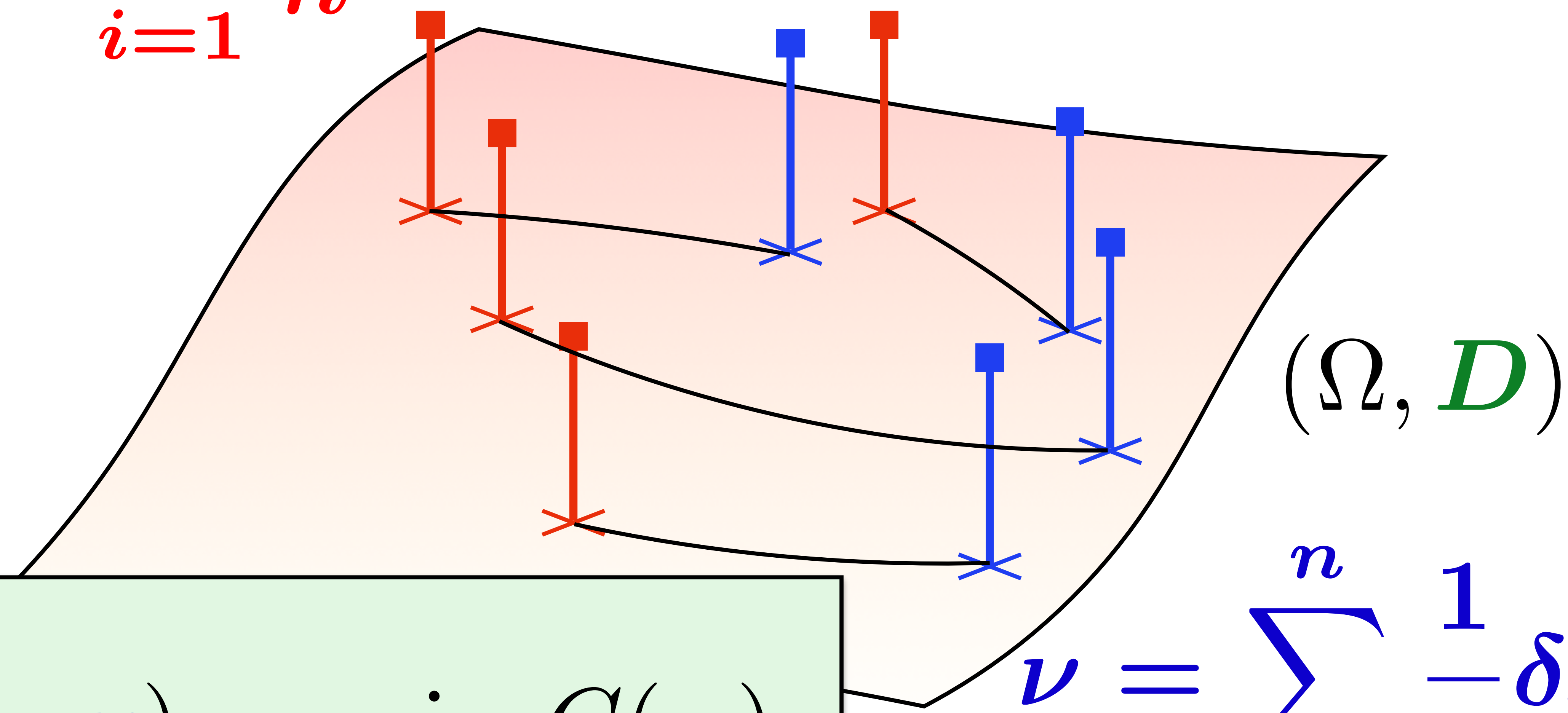
(Ω, D)

$$\nu = \sum_{j=1}^{n} \frac{1}{n} \delta_{y_j}$$

$$C(\sigma) = \frac{1}{n} \sum_{i=1}^{n} D(x_i, y_{\sigma_i})^p$$

Optimal Assignment $\subset$ Wasserstein

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

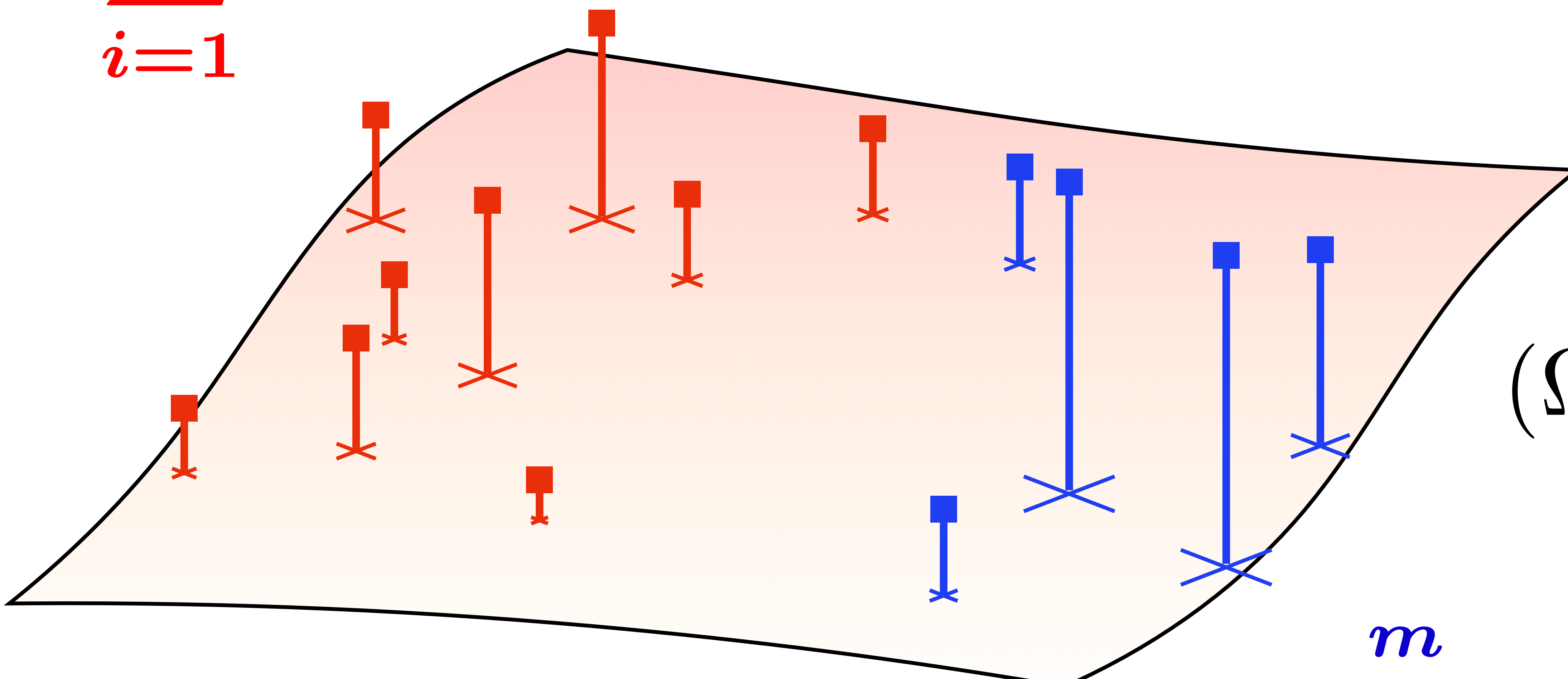


$$\boldsymbol{\nu} = \sum_{j=1}^{n} \frac{1}{n} \delta_{y_j}$$

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{\sigma} \in S_n} C(\boldsymbol{\sigma})$$

OT on Two Empirical Measures

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

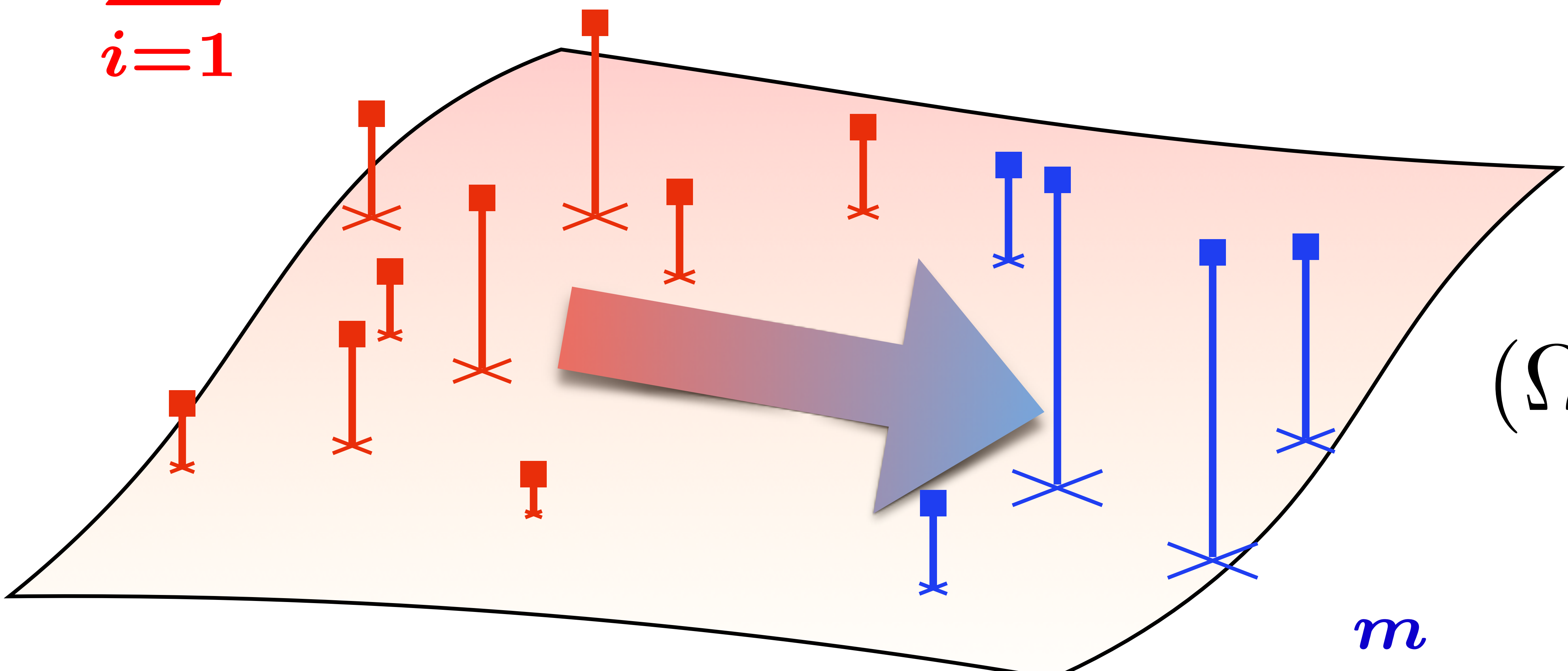


$$(\Omega, D)$$

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

OT on Two Empirical Measures

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



$$(\Omega, D)$$

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

Wasserstein on Empirical Measures

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

$$M_{XY} \stackrel{\text{def}}{=} [D(x_i, y_j)^p]_{ij}$$

$$U(a, b) \stackrel{\text{def}}{=} \{P \in \mathbb{R}_+^{n \times m} \mid P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\}$$

$$\begin{array}{c|c} & \begin{array}{ccc} y_1 & \cdots & y_m \end{array} \\ \hline \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} & \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & D(x_i, y_j)^p & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \end{array} \qquad \begin{array}{c|c} & \begin{array}{ccc} b_1 & \cdots & b_m \end{array} \\ \hline \begin{array}{c} a_1 \\ \vdots \\ a_n \end{array} & \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & P\mathbf{1}_m = a & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \end{array}$$

Wasserstein on Empirical Measures

Consider $\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a_i \delta_{x_i}}$ and $\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b_j \delta_{y_j}}$.

$$M_{\mathbf{XY}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

$$\begin{array}{c c} & \begin{array}{ccc} \boldsymbol{y_1} & \cdots & \boldsymbol{y_m} \end{array} \\ \begin{array}{c} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_n} \end{array} & \left[\begin{array}{ccc} \cdot & \cdot & \cdot \\ \cdot & D(\boldsymbol{x_i}, \boldsymbol{y_j})^p & \cdot \\ \cdot & \cdot & \cdot \end{array}\right] \end{array} \qquad \begin{array}{c c} & \begin{array}{ccc} \boldsymbol{b_1} & \cdots & \boldsymbol{b_m} \end{array} \\ \begin{array}{c} \boldsymbol{a_1} \\ \vdots \\ \boldsymbol{a_n} \end{array} & \left[\begin{array}{ccc} \vdots & \vdots & \vdots \\ \vdots & \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b} & \vdots \\ \vdots & \vdots & \vdots \end{array}\right] \end{array}$$

Wasserstein on Empirical Measures

Consider $\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a_i} \boldsymbol{\delta_{x_i}}$ and $\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b_j} \boldsymbol{\delta_{y_j}}$.

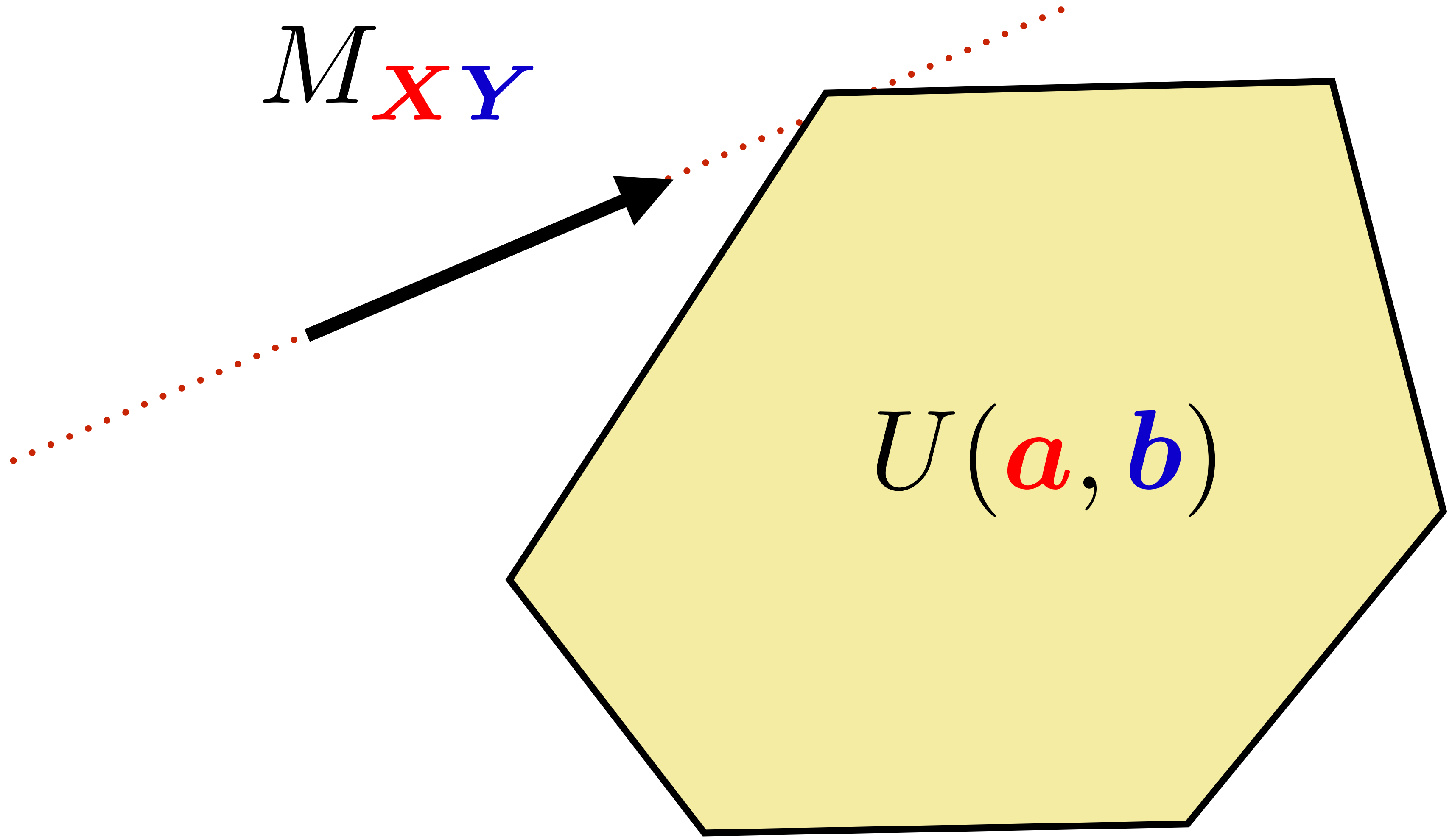
$$M_{\boldsymbol{XY}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

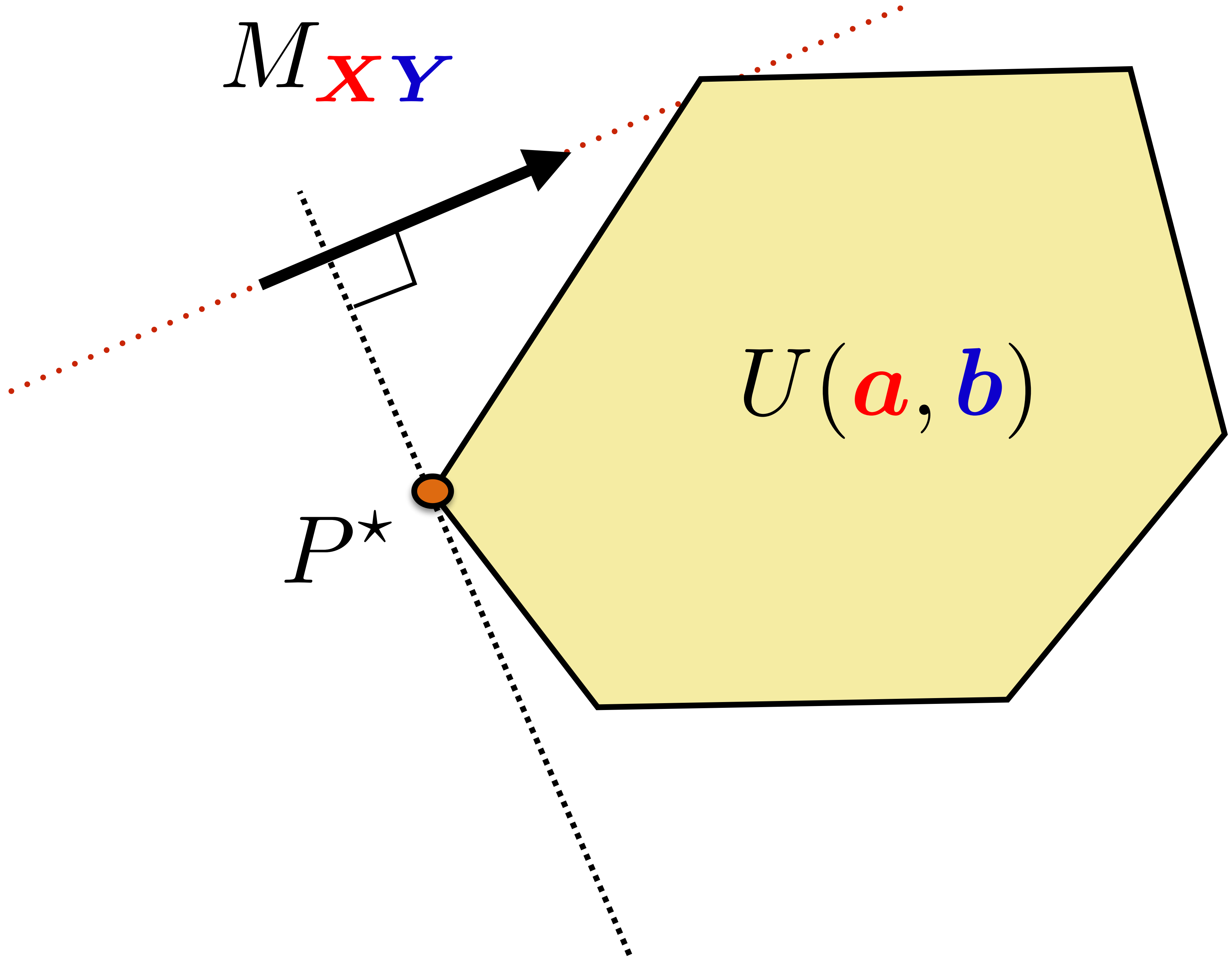
Def. Optimal Transport Problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

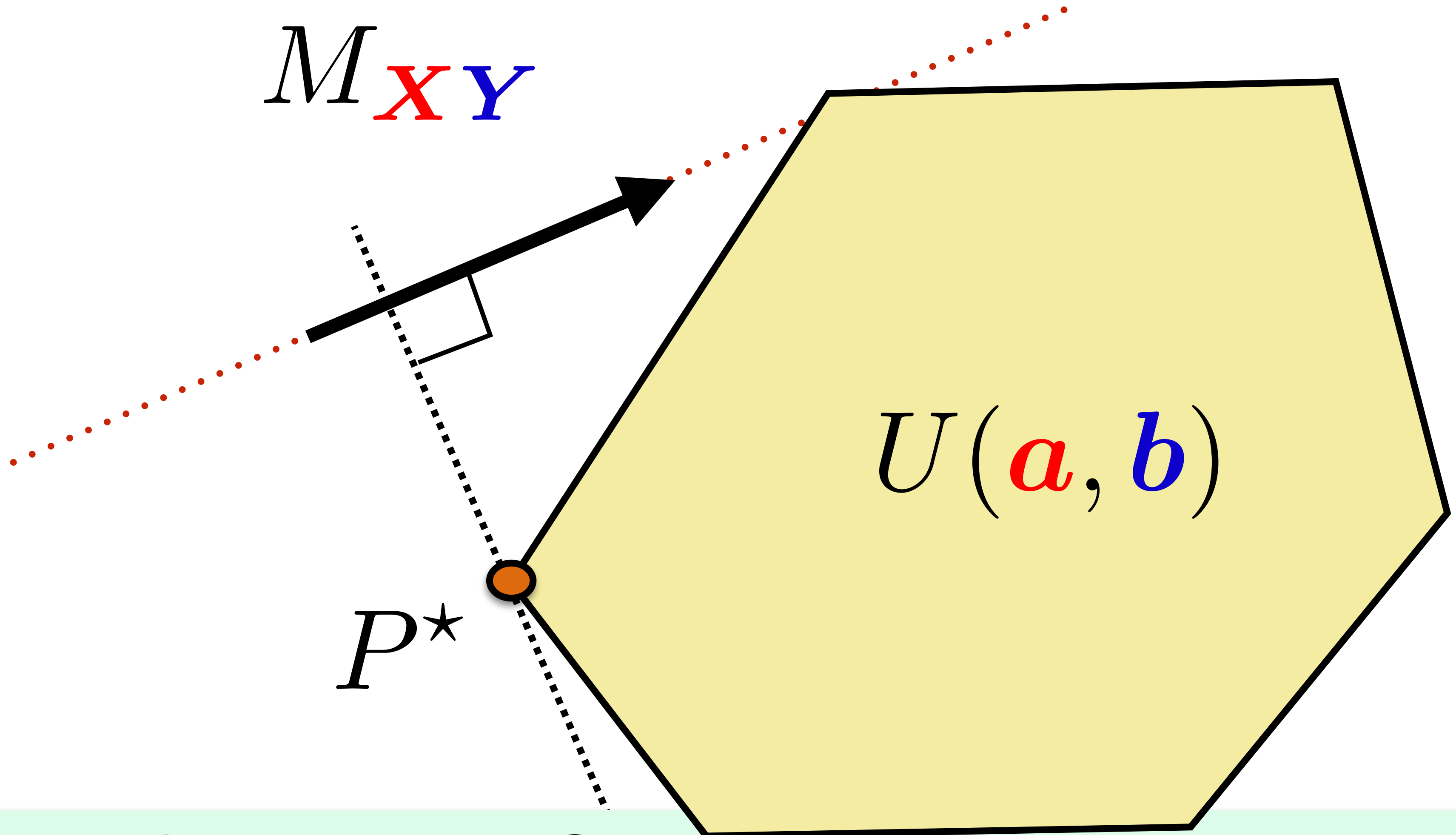
Discrete OT Problem



Discrete OT Problem



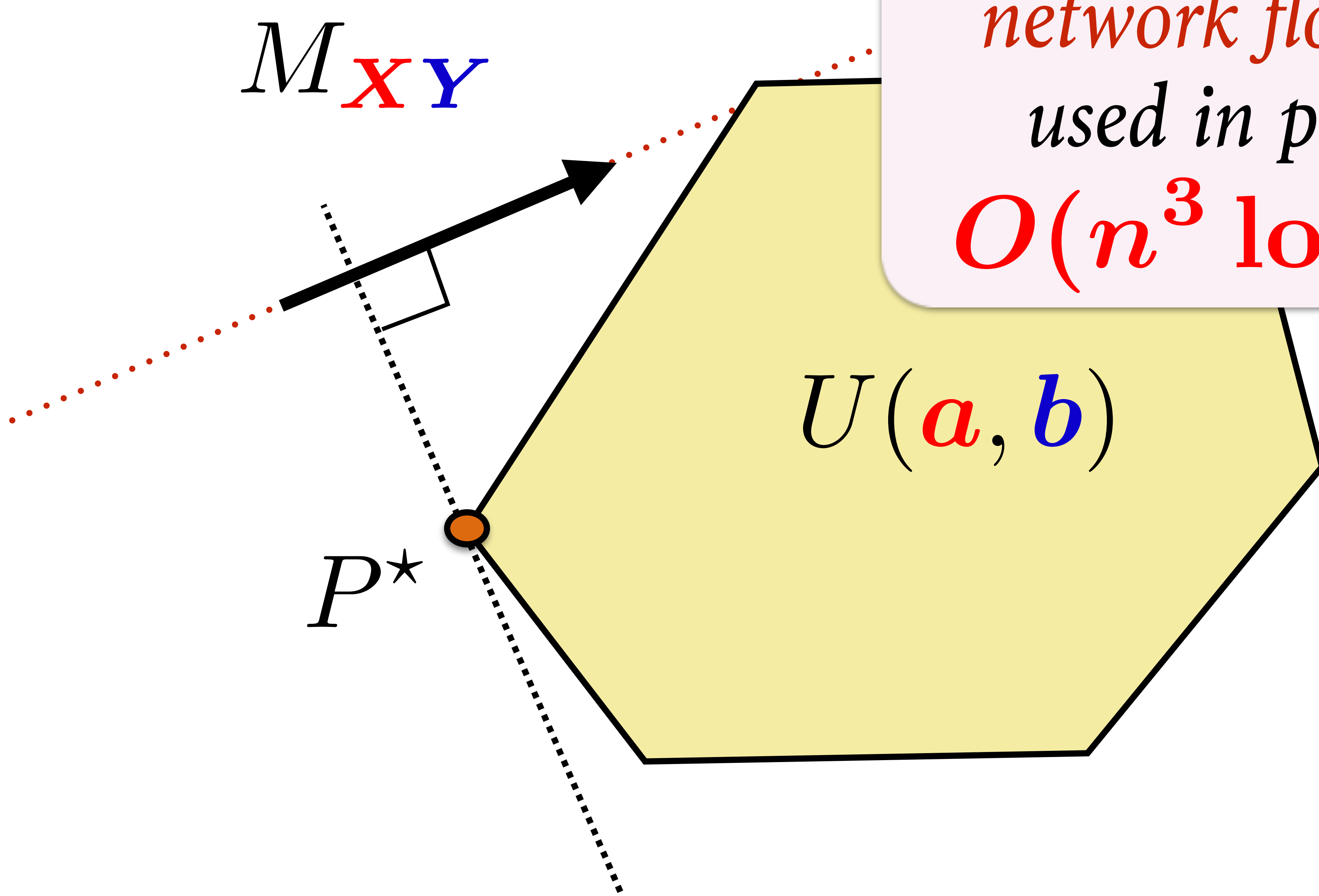
Discrete OT Problem



Def. Dual OT problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^m \\ \boldsymbol{\alpha}_{\boldsymbol{i}} + \boldsymbol{\beta}_{\boldsymbol{j}} \leq D(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{j}})^p}} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b}$$

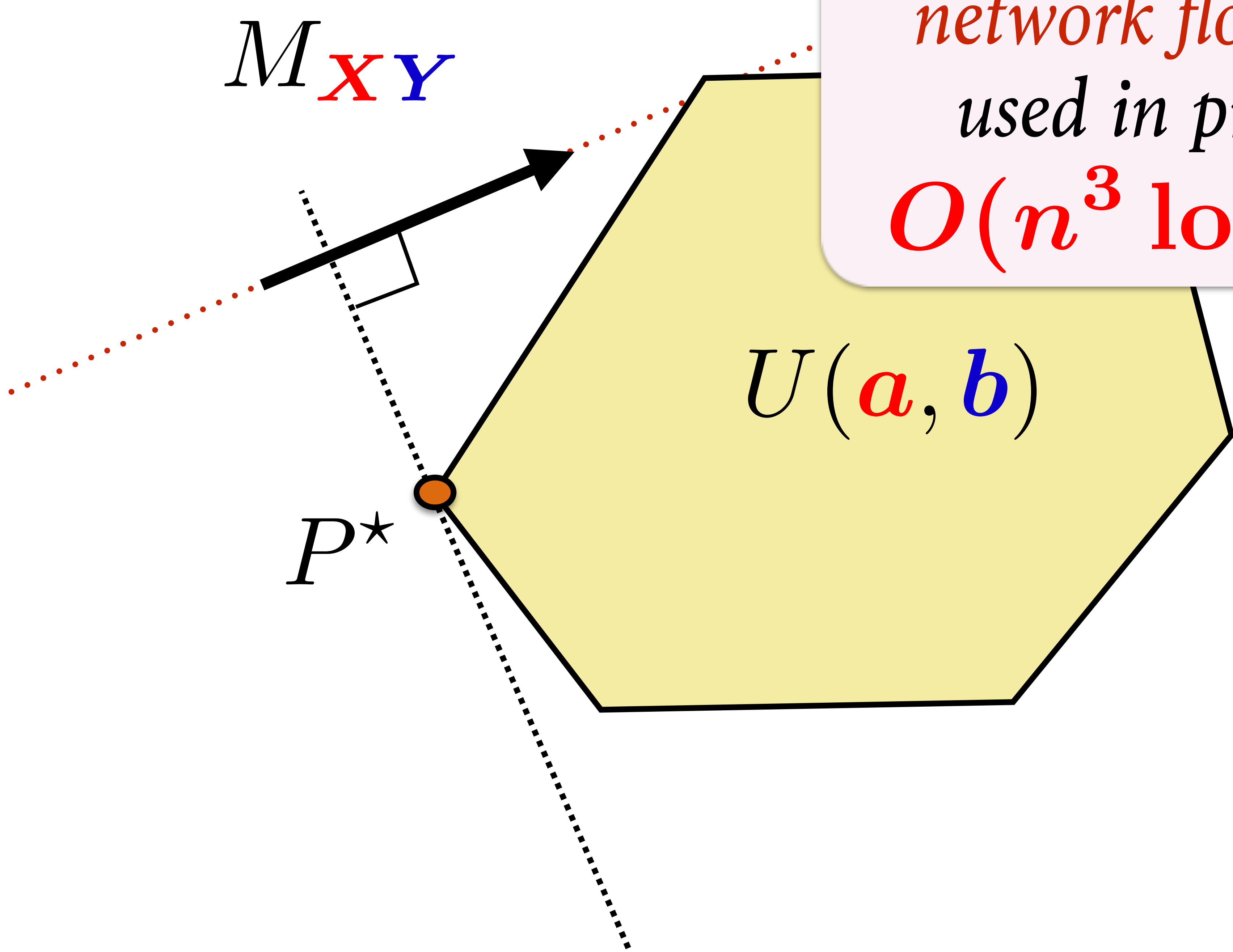
Discrete OT Problem



network flow solver
used in practice.
$$O(n^3 \log(n))$$

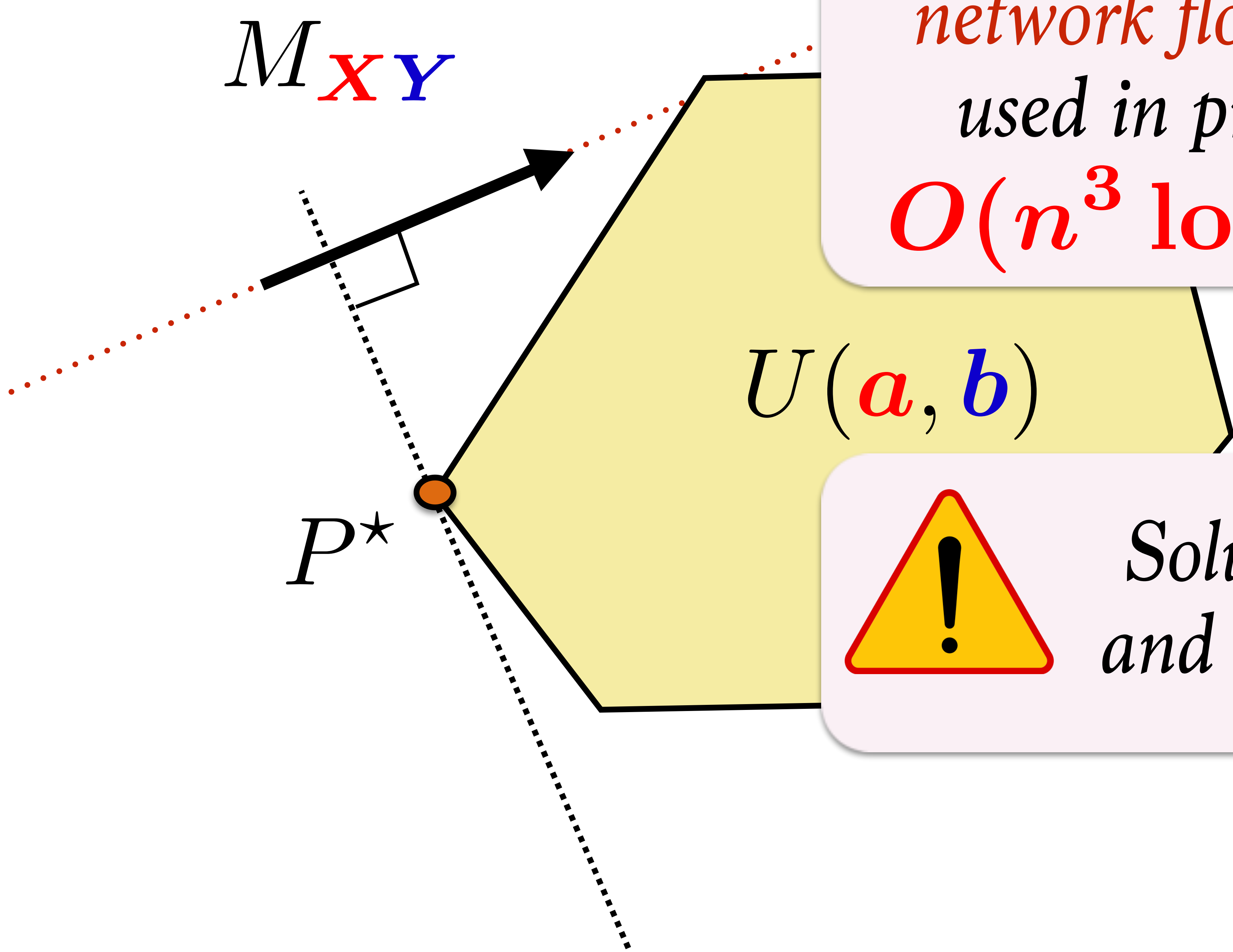
Note: flow/PDE formulations **[Beckman'61]/[Benamou'98]** can be used for *p=1/p=2* for a sparse-graph metric/Euclidean metric.

Discrete OT Problem



network flow solver
used in practice.
$\boldsymbol{O(n^3 \log(n))}$

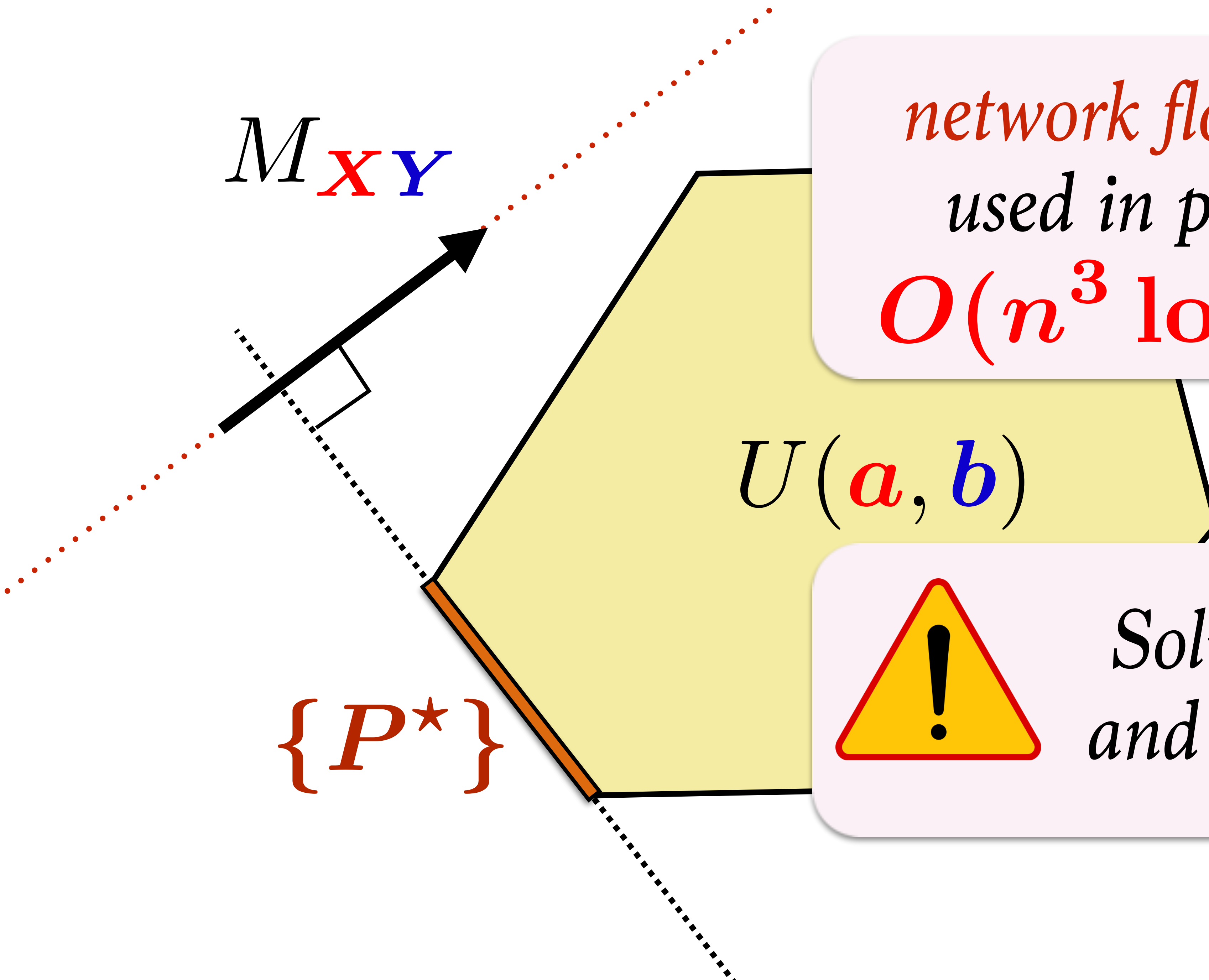
Discrete OT Problem



network flow solver
used in practice.
$O(n^3 \log(n))$

Solution $P^\star$ unstable
and not always unique.

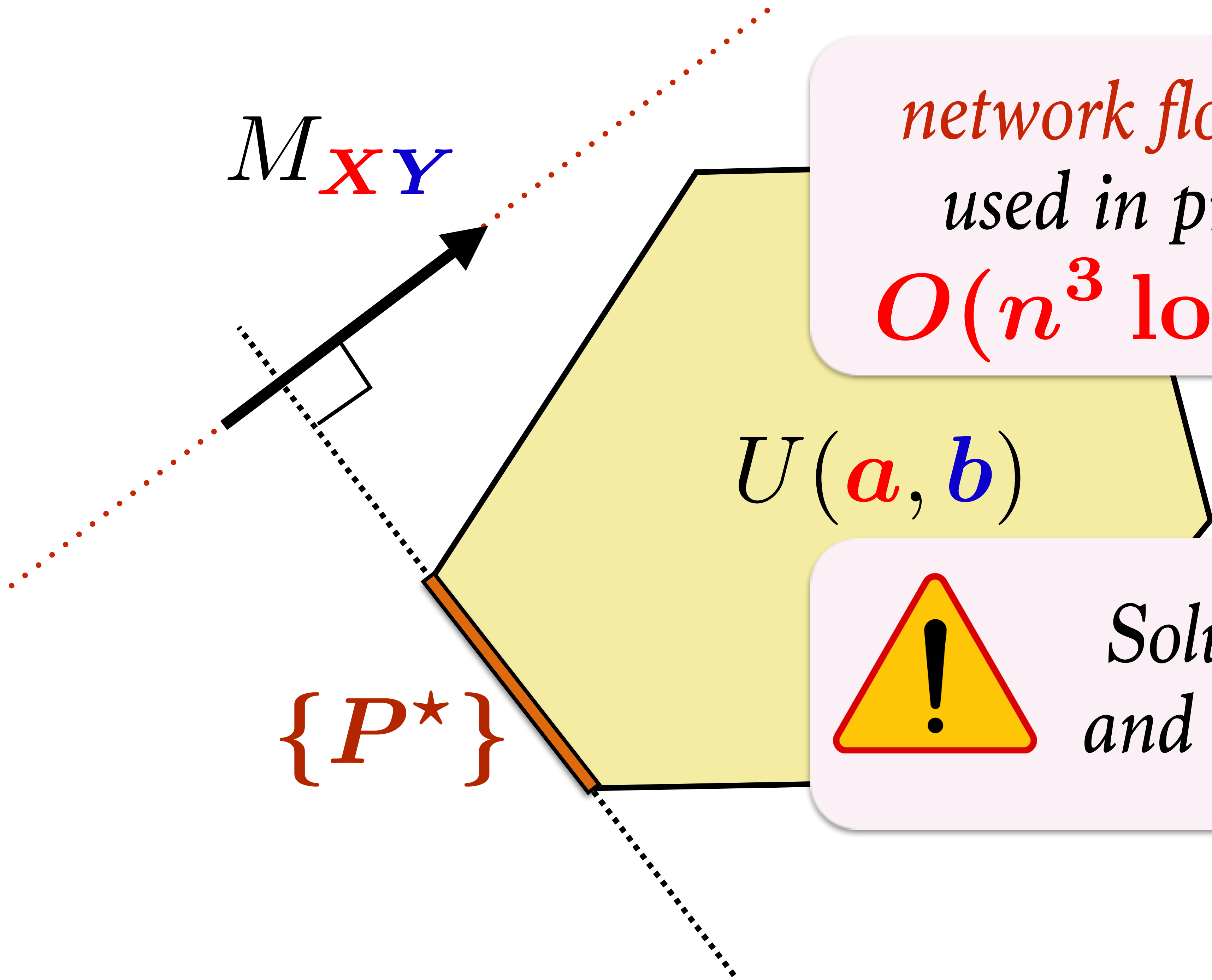
Discrete OT Problem



network flow solver
used in practice.
$O(n^3 \log(n))$

Solution $P^\star$ unstable
and not always unique.

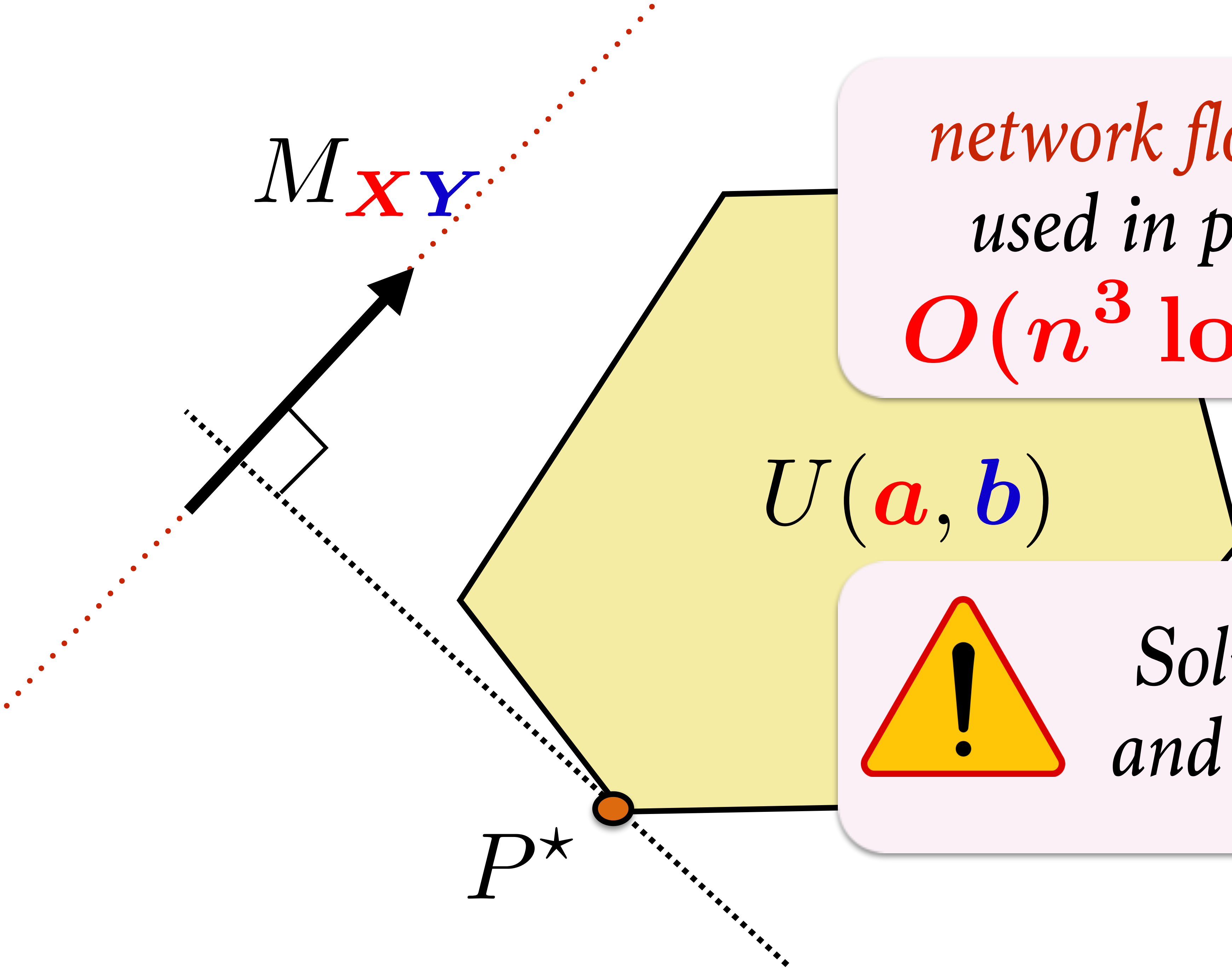
Discrete OT Problem



network flow solver
used in practice.
$O(n^3 \log(n))$

Solution $P^\star$ *unstable*
and not always unique.

Discrete OT Problem

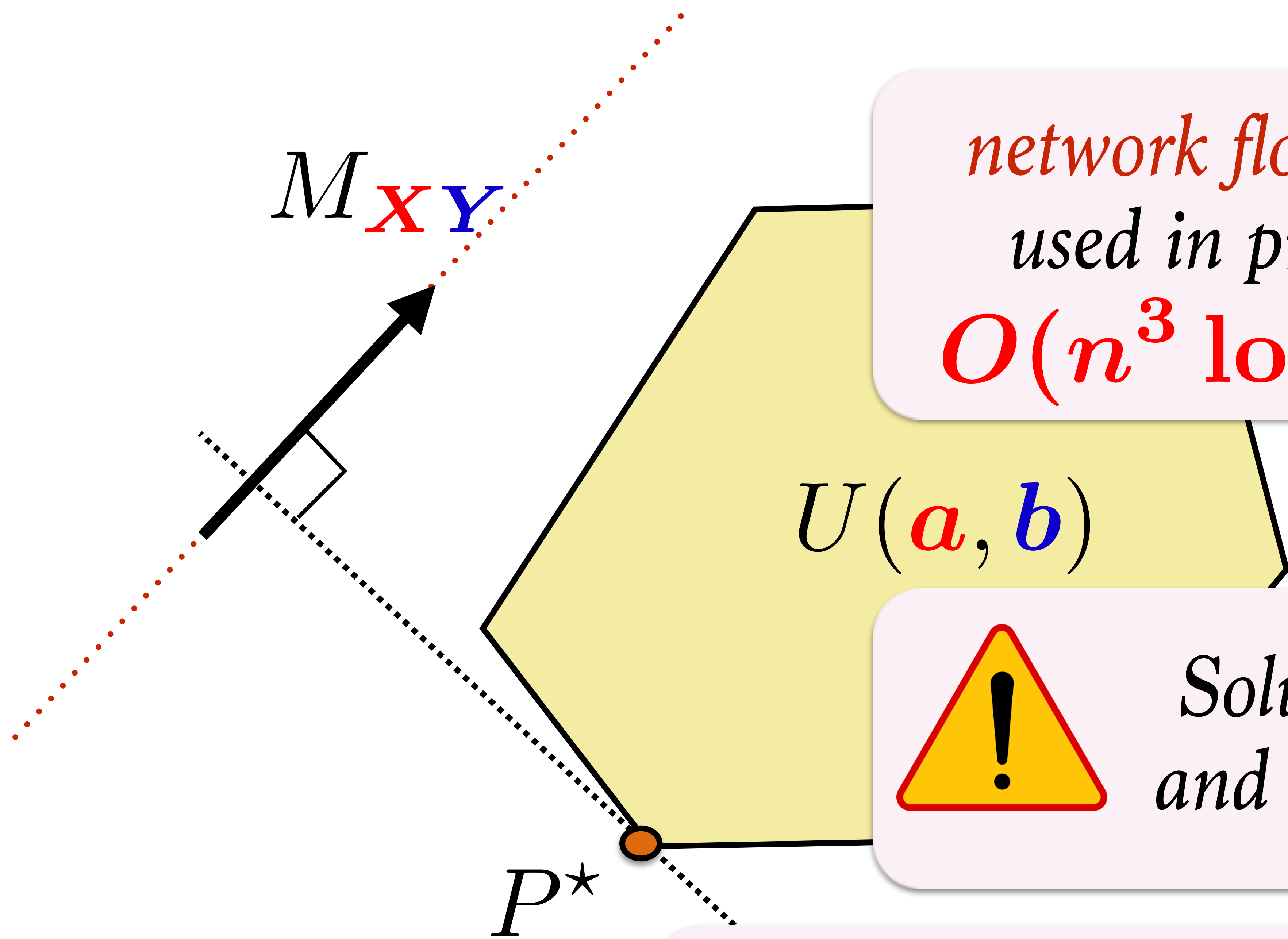


network flow solver *used in practice.*

$O(n^3 \log(n))$

Solution $P^\star$ *unstable and not always unique.*

Discrete OT Problem



network flow solver used in practice. $O(n^3 \log(n))$

Solution $P^\star$ unstable and not always unique.

$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu})$ not differentiable.

Discrete OT Problem

```c
/*
    emd.c

    Last update: 3/14/98

    An implementation of the Earth Movers Distance.
    Based of the solution for the Transportation problem as described in
    "Introduction to Mathematical Programming" by F. S. Hillier and
    G. J. Lieberman, McGraw-Hill, 1990.

    Copyright (C) 1998 Yossi Rubner
    Computer Science Department, Stanford University
    E-Mail: rubner@cs.stanford.edu   URL: http://vision.stanford.edu/~rubner
*/

/*#include <stdio.h>
#include <stdlib.h>*/
#include <math.h>

#include "emd.h"

#define DEBUG_LEVEL 0
/*
 DEBUG_LEVEL:
   0 = NO MESSAGES
   1 = PRINT THE NUMBER OF ITERATIONS AND THE FINAL RESULT
   2 = PRINT THE RESULT AFTER EVERY ITERATION
   3 = PRINT ALSO THE FLOW AFTER EVERY ITERATION
   4 = PRINT A LOT OF INFORMATION (PROBABLY USEFUL ONLY FOR THE AUTHOR)
*/


#define MAX_SIG_SIZE1  (MAX_SIG_SIZE+1)  /* FOR THE POSIBLE DUMMY FEATURE */

/* NEW TYPES DEFINITION */

/* node1_t IS USED FOR SINGLE-LINKED LISTS */
typedef struct node1_t {
  int i;
  double val;
  struct node1_t *Next;
} node1_t;

/* node1_t IS USED FOR DOUBLE-LINKED LISTS */
typedef struct node2_t {
  int i, j;
  double val;
  struct node2_t *NextC;               /* NEXT COLUMN */
  struct node2_t *NextR;               /* NEXT ROW */
} node2_t;



/* GLOBAL VARIABLE DECLARATION */
static int _n1, _n2;                          /* SIGNATURES SIZES */
static float _C[MAX_SIG_SIZE1][MAX_SIG_SIZE1];/* THE COST MATRIX */
static node2_t _X[MAX_SIG_SIZE1*2];            /* THE BASIC VARIABLES VECTOR */
```

Discrete OT Problem

```c
/*
    emd.c

    Last update: 3/14/98

    An implementation of the Earth Movers Distance.
    Based of the solution for the Transportation problem as described in
    "Introduction to Mathematical Programming" by F. S. Hillier and
    G. J. Lieberman, McGraw-Hill, 1990.

    Copyright (C) 1998 Yossi Rubner
    Computer Science Department, Stanford University
    E-Mail: rubner@cs.stanford.edu   URL: http://vision.stanford.edu/~rubner
*/

/*#include <stdio.h>
#include <stdlib.h>*/
#include <math.h>

#include "emd.h"

#define DEBUG_LEVEL 0
/*
 DEBUG_LEVEL:
   0 = NO MESSAGES
   1 = PRINT THE NUMBER OF ITERATIONS AND THE FINAL RESULT
   2 = PRINT THE RESULT AFTER EVERY ITERATION
   3 = PRINT ALSO THE FLOW AFTER EVERY ITERATION
   4 = PRINT A LOT OF INFORMATION (PROBABLY USEFUL ONLY FOR THE AUTHOR)
*/


#define MAX_SIG_SIZE1  (MAX_SIG_SIZE+1)  /* FOR THE POSIBLE DUMMY FEATURE */

/* NEW TYPES DEFINITION */

/* node1_t IS USED FOR SINGLE-LINKED LISTS */
typedef struct node1_t {
  int i;
  double val;
  struct node1_t *Next;
} node1_t;

/* node1_t IS USED FOR DOUBLE-LINKED LISTS */
typedef struct node2_t {
  int i, j;
  double val;
  struct node2_t *NextC;              /* NEXT COLUMN */
  struct node2_t *NextR;              /* NEXT ROW */
} node2_t;



/* GLOBAL VARIABLE DECLARATION */
static int _n1, _n2;                          /* SIGNATURES SIZES */
static float _C[MAX_SIG_SIZE1][MAX_SIG_SIZE1];/* THE COST MATRIX */
static node2_t _X[MAX_SIG_SIZE1*2];            /* THE BASIC VARIABLES VECTOR */
```

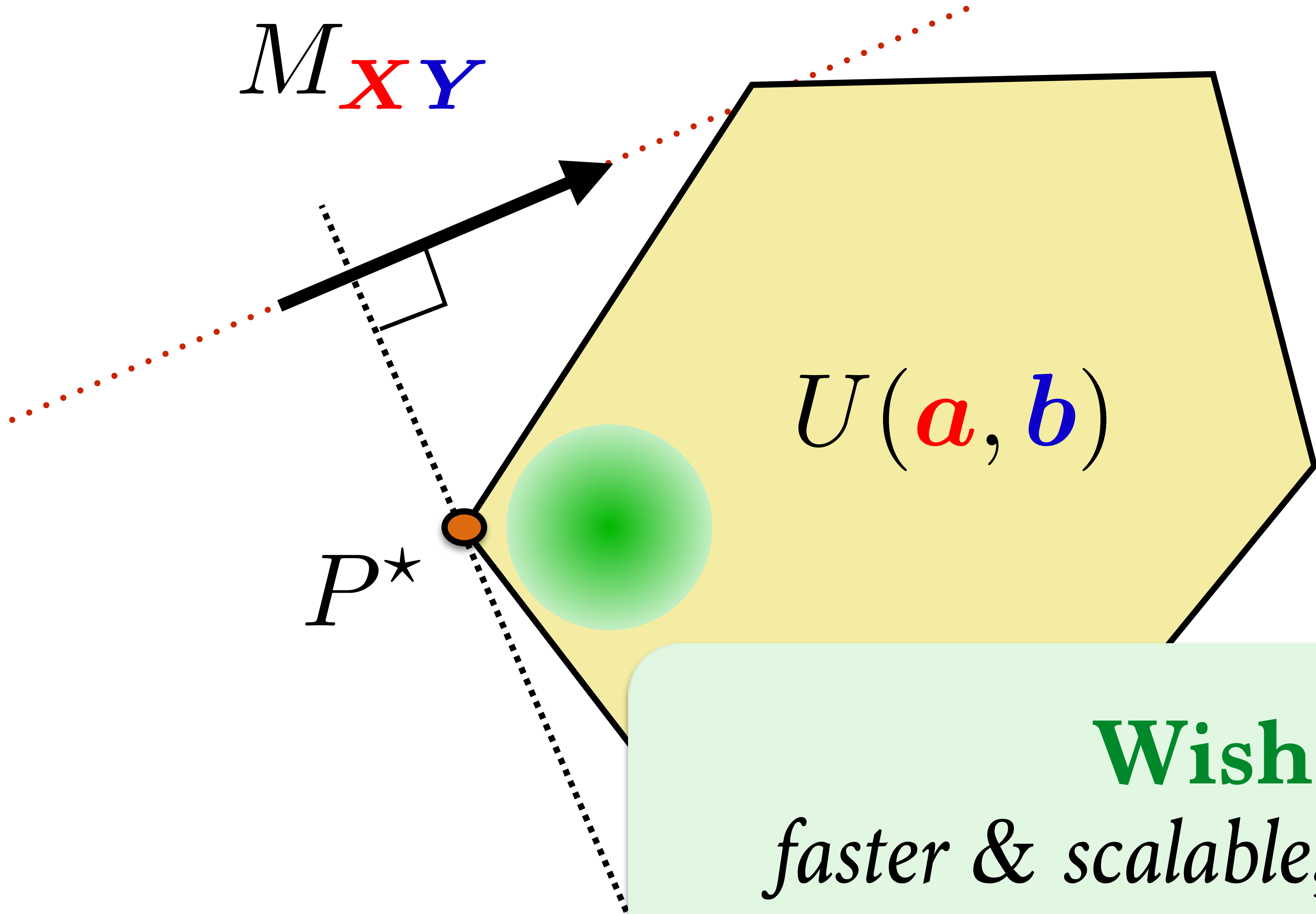
Discrete OT Problem

emd.c

emd.c:6:1 <No selected symbol>

```c
/*
    emd.c

    Last update: 3/14/98

    An implementation of the Earth Movers Distance.
    Based of the solution for the Transportation problem as described in
    "Introduction to Mathematical Programming" by F. S. Hillier and
    G. J. Lieberman, McGraw-Hill, 1990.

    Copyright (C) 1998 Yossi Rubner
    Computer Science Department, Stanford University
    E-Mail: rubner@cs.stanford.edu   URL: http://vision.stanford.edu/~rubner
*/

/*#include <stdio.h>
#include <stdlib.h>*/
#include <math.h>

#include "emd.h"

#define DEBUG_LEVEL 0
/*
 DEBUG_LEVEL:
   0 = NO MESSAGES
   1 = PRINT THE NUMBER OF ITERATIONS AND THE FINAL RESULT
   2 = PRINT THE RESULT AFTER EVERY ITERATION
   3 = PRINT ALSO THE FLOW AFTER EVERY ITERATION
   4 = PRINT A LOT OF INFORMATION (PROBABLY USEFUL ONLY FOR THE AUTHOR)
*/


#define MAX_SIG_SIZE1  (MAX_SIG_SIZE+1)  /* FOR THE POSIBLE DUMMY FEATURE */

/* NEW TYPES DEFINITION */

/* node1_t IS USED FOR SINGLE-LINKED LISTS */
typedef struct node1_t {
  int i;
  double val;
  struct node1_t *Next;
} node1_t;

/* node1_t IS USED FOR DOUBLE-LINKED LISTS */
typedef struct node2_t {
  int i, j;
  double val;
  struct node2_t *NextC;              /* NEXT COLUMN */
  struct node2_t *NextR;              /* NEXT ROW */
} node2_t;



/* GLOBAL VARIABLE DECLARATION */
static int _n1, _n2;                          /* SIGNATURES SIZES */
static float _C[MAX_SIG_SIZE1][MAX_SIG_SIZE1];/* THE COST MATRIX */
static node2_t _X[MAX_SIG_SIZE1*2];            /* THE BASIC VARIABLES VECTOR */
```

Solution: Modify OT Problem



Wishlist:
faster & *scalable*, more *stable*,
differentiable

Entropic Regularization [**Wilson'62**]

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

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle - \gamma E(\boldsymbol{P})$$

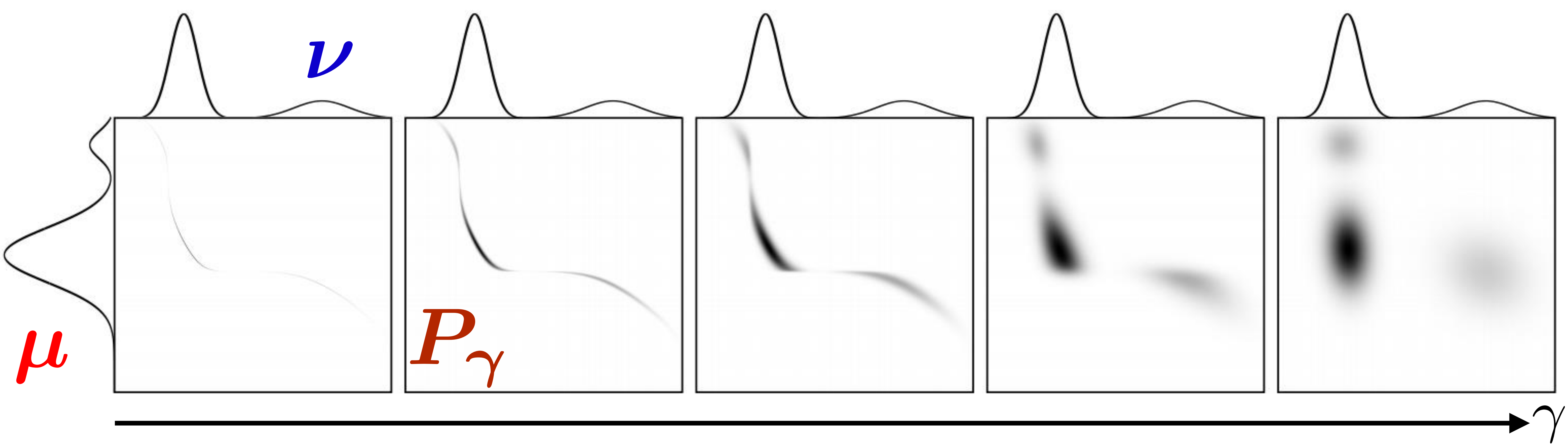
$$E(P) \stackrel{\text{def}}{=} - \sum_{i,j=1}^{nm} P_{ij} (\log P_{ij})$$

Note: Unique optimal solution because of strong concavity of Entropy

Entropic Regularization [Wilson'62]

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

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle - \gamma E(\boldsymbol{P})$$



Note: Unique optimal solution because of strong concavity of Entropy

Fast & Scalable Algorithm

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

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

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

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

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

$$\partial L / \partial P_{ij} = M_{ij} + \gamma(\log P_{ij} + 1) + \alpha_i + \beta_j$$

$$(\boldsymbol{\partial L / \partial P_{ij}} = 0) \Rightarrow P_{ij} = e^{\frac{\alpha_i}{\gamma} + \frac{1}{2}} e^{-\frac{M_{ij}}{\gamma}} e^{\frac{\beta_j}{\gamma} + \frac{1}{2}} = \boldsymbol{u_i} K_{ij} \boldsymbol{v_j}$$

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

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

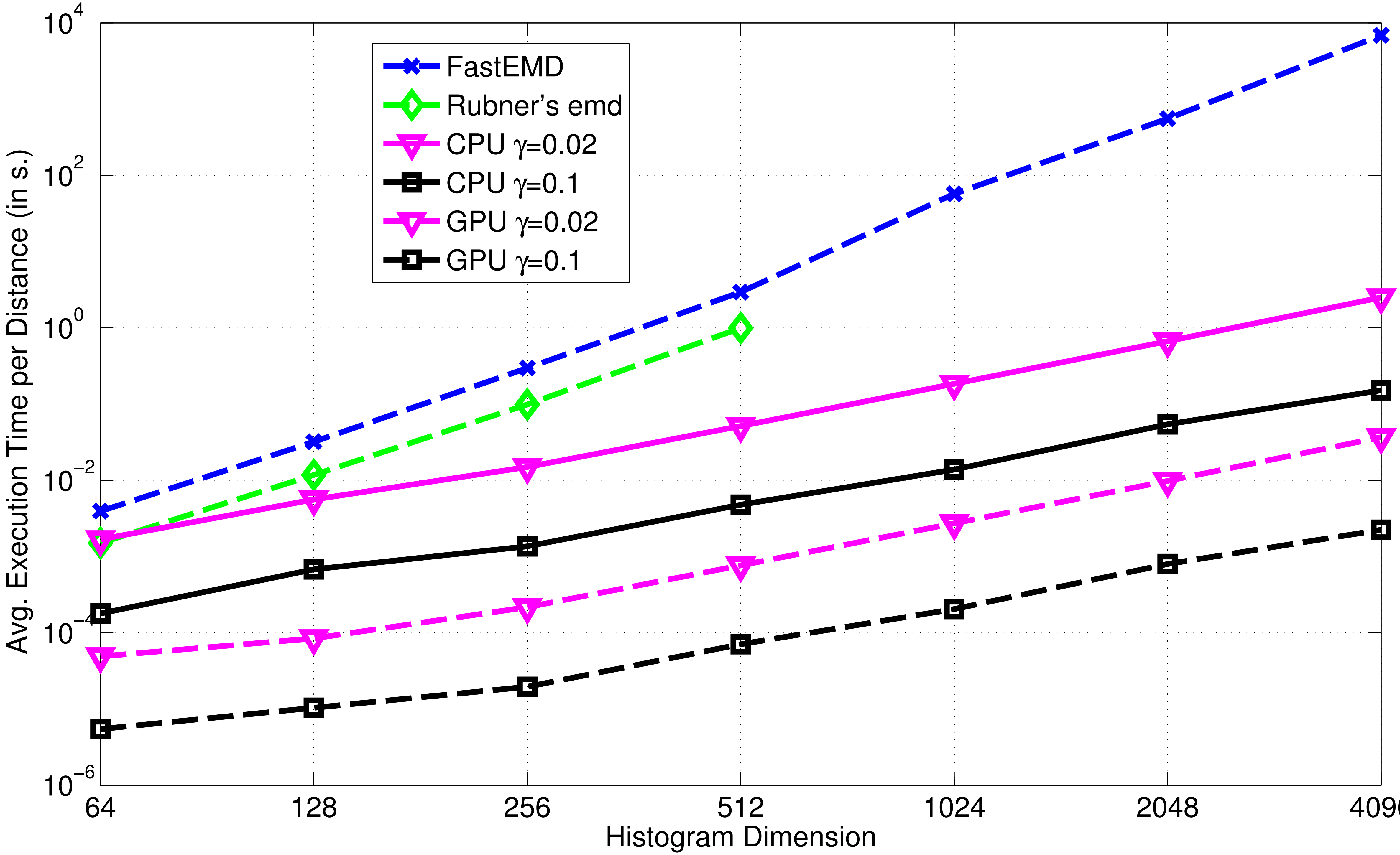
- **[Sinkhorn'64]** fixed-point iterations for $(\boldsymbol{u}, \boldsymbol{v})$

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

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

[S..C..'15]

Very Fast EMD Approx. Solver

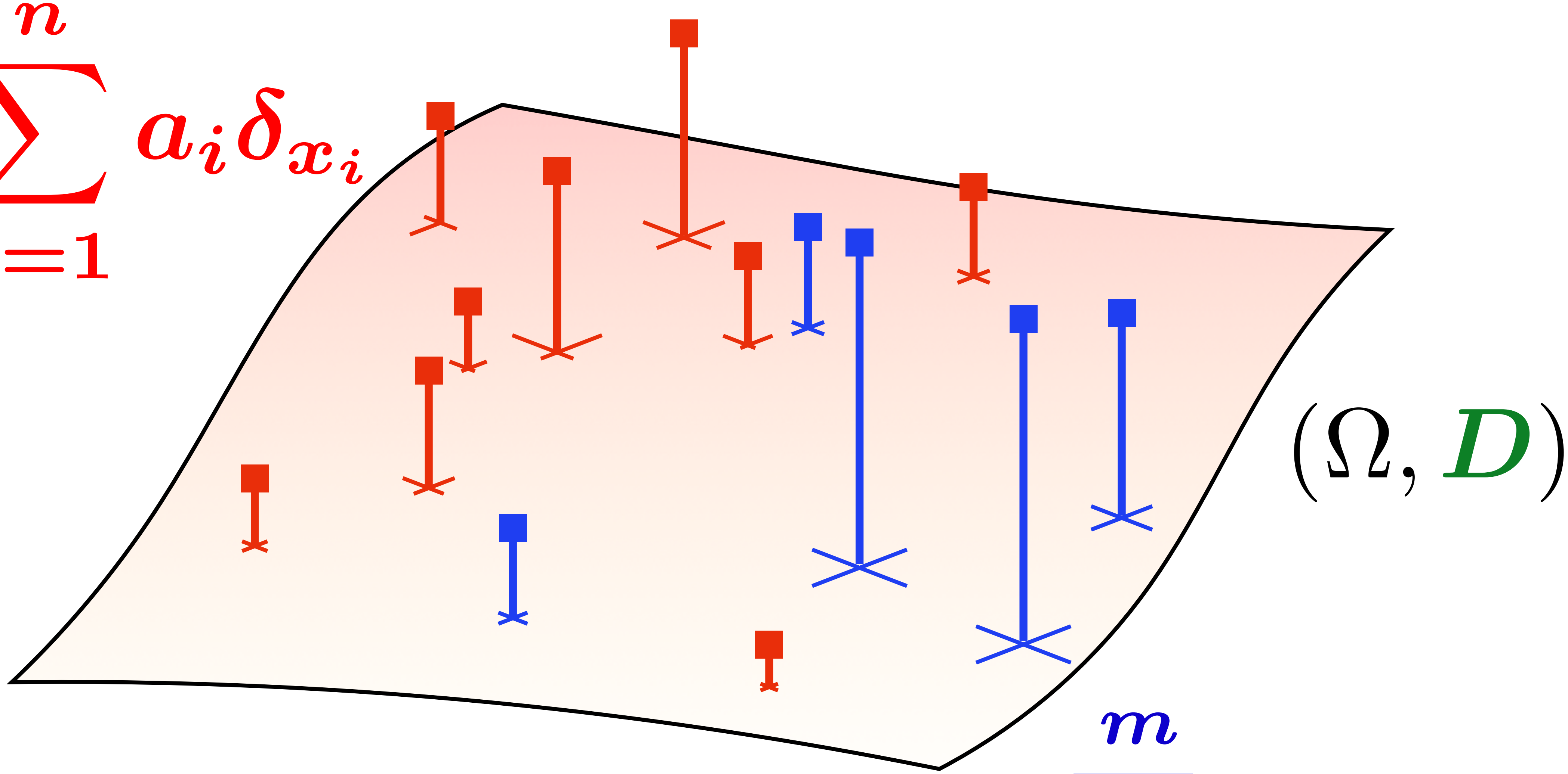


Note. (Ω, D) is a random graph with shortest path metric, histograms sampled uniformly on simplex, Sinkhorn tolerance 10^{-2}.

Regularization ⇝ *Differentiability*

$$W_\gamma((\boldsymbol{a}, \boldsymbol{X}), (\boldsymbol{b}, \boldsymbol{Y})) = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle - \gamma E(\boldsymbol{P})$$

$$\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{\delta}_{\boldsymbol{x}_i}$$

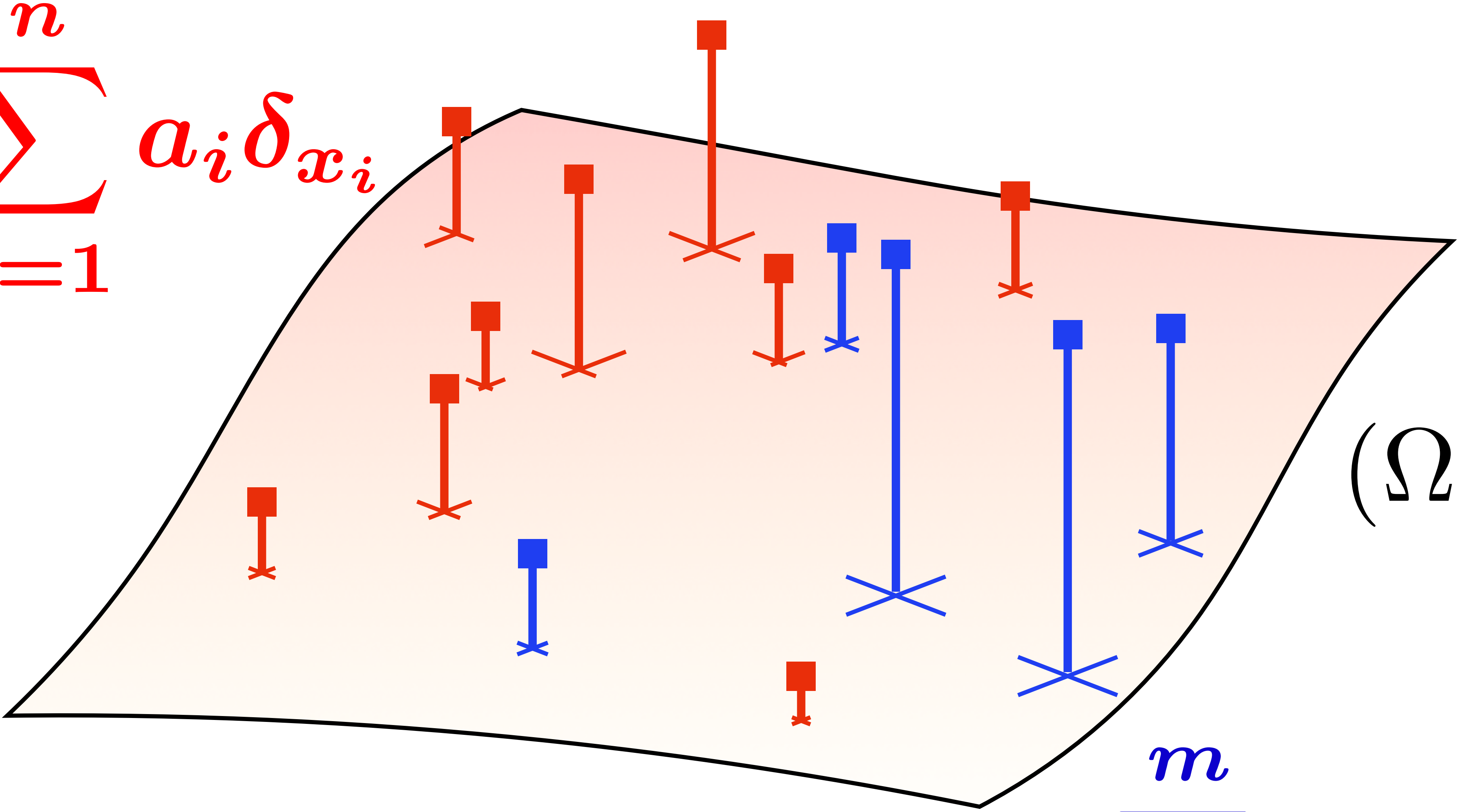


$$\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{\delta}_{\boldsymbol{y}_j}$$

Regularization ⇝ *Differentiability*

$$W_\gamma((\boldsymbol{a} + \boldsymbol{\Delta a}, \boldsymbol{X}), (\boldsymbol{b}, \boldsymbol{Y})) = W_\gamma((\boldsymbol{a}, \boldsymbol{X}), (\boldsymbol{b}, \boldsymbol{Y})) + ??$$

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

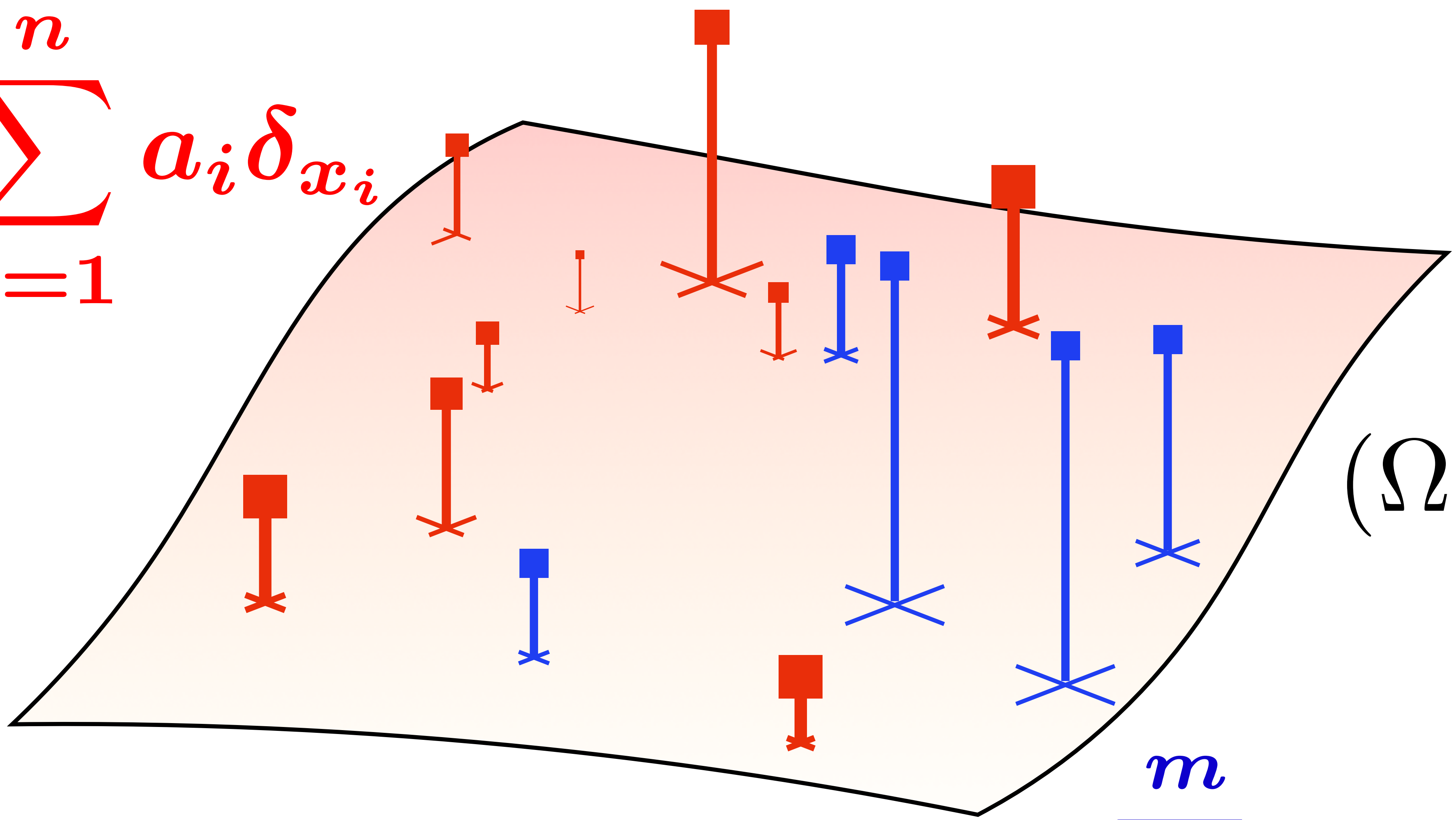


$$\boldsymbol{\nu} = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Regularization ⇝ *Differentiability*

$$W_\gamma((\mathbf{a}+\mathbf{\Delta a},\mathbf{X}),(\mathbf{b},\mathbf{Y})) = W_\gamma((\mathbf{a},\mathbf{X}),(\mathbf{b},\mathbf{Y}))+??$$

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



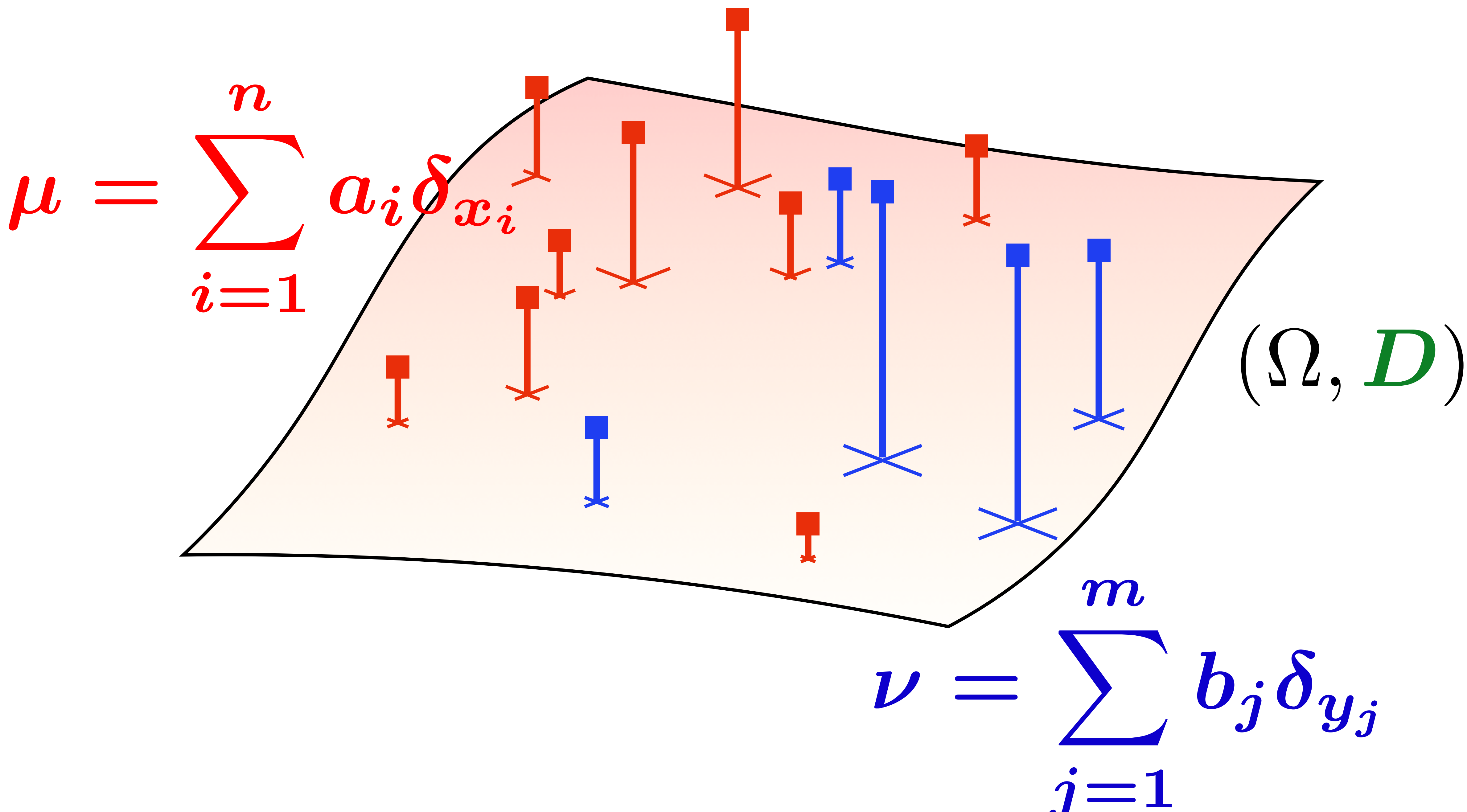
$(\Omega, \boldsymbol{D})$

$$\boldsymbol{a} \leftarrow \boldsymbol{a} + \boldsymbol{\Delta a}$$

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

Regularization ⇢ *Differentiability*

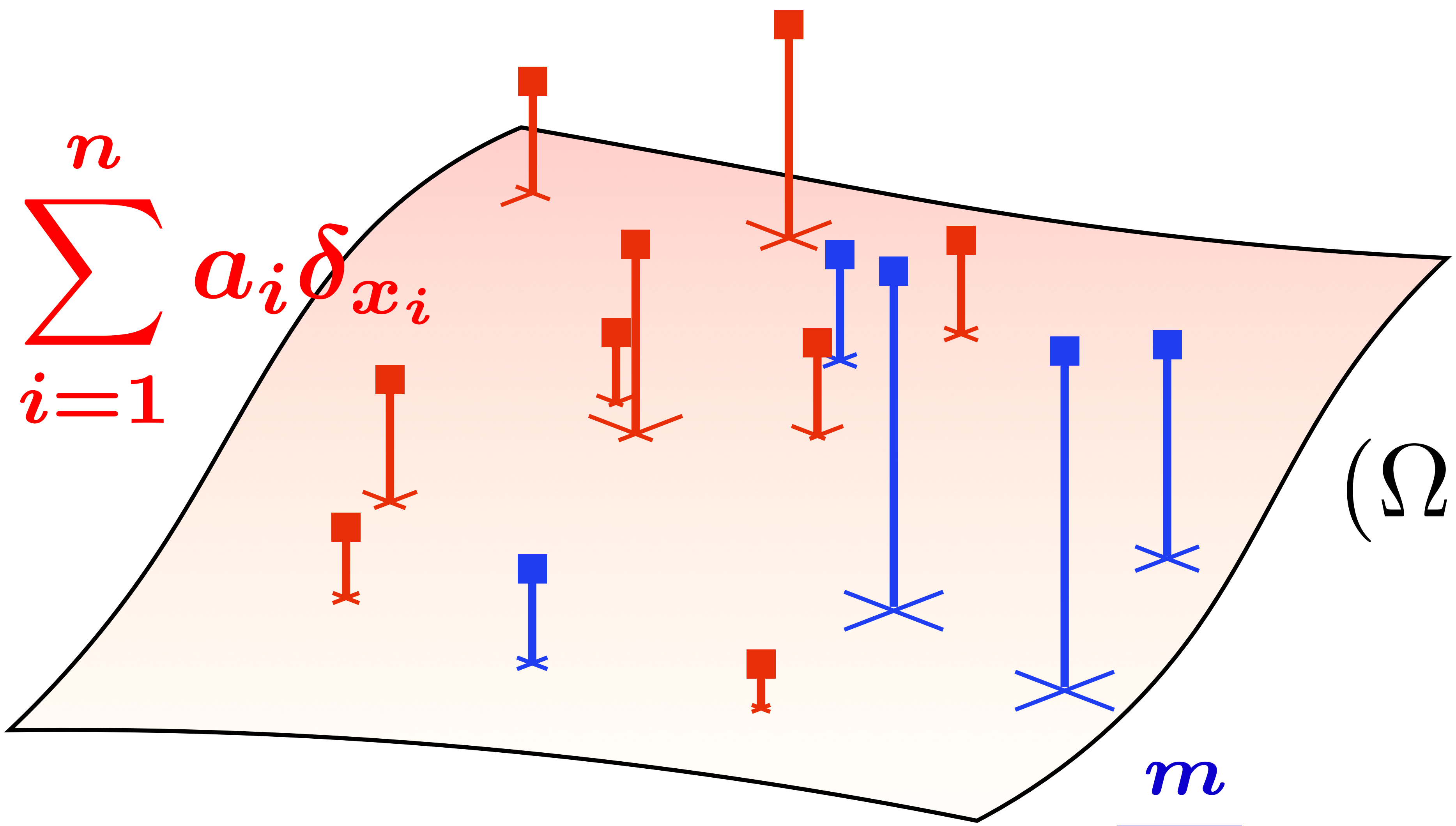
$$W_\gamma((\boldsymbol{a}, \boldsymbol{X} + \boldsymbol{\Delta X}), (\boldsymbol{b}, \boldsymbol{Y})) = W_\gamma((\boldsymbol{a}, \boldsymbol{X}), (\boldsymbol{b}, \boldsymbol{Y})) + ??$$



Regularization ⇝ *Differentiability*

$$W_\gamma((\boldsymbol{a}, \boldsymbol{X} + \boldsymbol{\Delta X}), (\boldsymbol{b}, \boldsymbol{Y})) = W_\gamma((\boldsymbol{a}, \boldsymbol{X}), (\boldsymbol{b}, \boldsymbol{Y})) + ??$$

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



$$(\Omega, \boldsymbol{D})$$

$$\boldsymbol{X} \leftarrow \boldsymbol{X} + \boldsymbol{\Delta X}$$

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

1. Differentiability of Regularized OT

Def. Dual regularized OT Problem

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\alpha,\beta} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b} - \frac{1}{\gamma}(e^{\alpha/\gamma})^T \; K e^{\beta/\gamma}$$

Prop. $W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu})$ is **[CD'14]**

1. convex w.r.t. $\boldsymbol{a}$,

$$\nabla_{\boldsymbol{a}} W_\gamma = \alpha^\star = \gamma \log(\boldsymbol{u}).$$

2. decreased, when $p=2, \Omega=\mathbb{R}^d$, using

$$\boldsymbol{X} \leftarrow \boldsymbol{Y} P_\gamma^T \mathbf{D}(\boldsymbol{a}^{-1}).$$

2. Duality for Discrete Reg. OT's

[CP'16]

Prop. Writing $H_{\boldsymbol{\nu}} : \boldsymbol{a} \mapsto W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$,

1. $H_{\boldsymbol{\nu}}$ has simple Legendre transform:

$$H_{\boldsymbol{\nu}}^* : \boldsymbol{g} \in \mathbb{R}^n \mapsto \gamma \left(E(\boldsymbol{b}) + \boldsymbol{b}^T \log(K e^{\boldsymbol{g}/\gamma}) \right)$$

2. If $A \in \mathbb{R}^{n \times d}$, f convex on $\mathbb{R}^d$,

$$\min_{\boldsymbol{a} \in \Sigma_n} H_{\boldsymbol{\nu}}(\boldsymbol{a}) + f(A\boldsymbol{a}) = \max_{\boldsymbol{g} \in \mathbb{R}^d} -H_{\boldsymbol{\nu}}^*(A^T \boldsymbol{g}) - f^*(-\boldsymbol{g})$$

3. Stochastic Formulation

[GCPB'16

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

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

DUAL

regularizing dual *constraints* $\gamma > 0$

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

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

REGULARIZED DUAL

3. Stochastic Formulation

[GCPB'16

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

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

DUAL

regularizing dual *constraints* $\gamma > 0$

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

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

REGULARIZED DUAL

Smoothed D transforms

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

SEMI-DUAL

$$\gamma > 0$$

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

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

REGULARIZED SEMI-DUAL

Regularized Semidual Wasserstein

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

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

REGULARIZED SEMI-DUAL

substituting

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

REGULARIZED SEMI-DUAL

Stochastic Regularized Semidual

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

REGULARIZED SEMI-DUAL

Stochastic Regularized Semidual

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

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 $\alpha \in \mathbb{R}^n$!

Stochastic Regularized Semidual

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

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

4. Sinkhorn Divergence

Def. For $\gamma > 0$, let $W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \langle \boldsymbol{P}_\gamma, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle$

Prop. $W_\gamma(\boldsymbol{\mu}, \boldsymbol{\mu}) > 0$

Def. Normalized Sinkhorn Divergence

$$\bar{W}_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}\left(W_\gamma(\boldsymbol{\mu}, \boldsymbol{\mu}) + W_\gamma(\boldsymbol{\nu}, \boldsymbol{\nu})\right)$$

Prop. If $p = 1$, $\bar{W}_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \underset{\gamma \to \infty}{\longrightarrow} \mathrm{ED}(\boldsymbol{\mu}, \boldsymbol{\nu})$

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 of Reg. OT

Example: Differentiability w.r.t. a

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

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

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

Algorithmic Formulation of Reg. OT

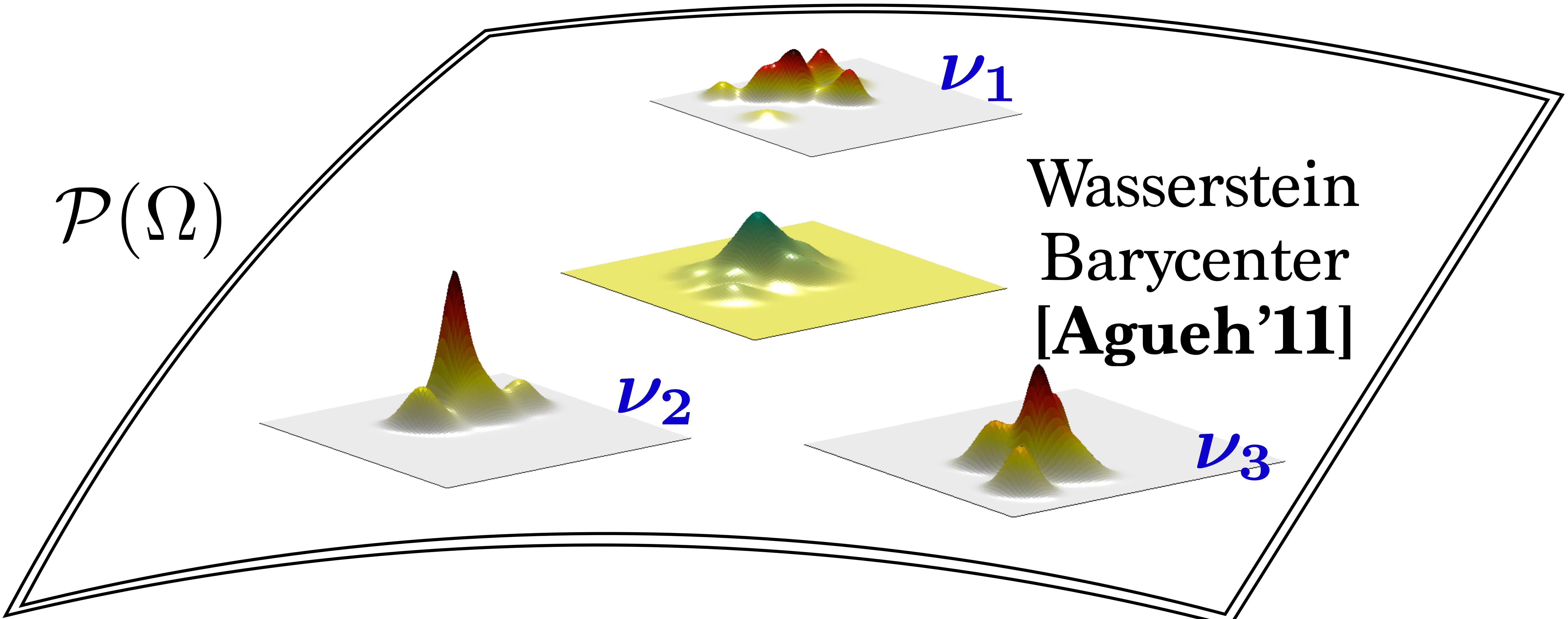
Example: Differentiability w.r.t. a

$$N = K \circ M_{\boldsymbol{X}\boldsymbol{Y}}$$

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

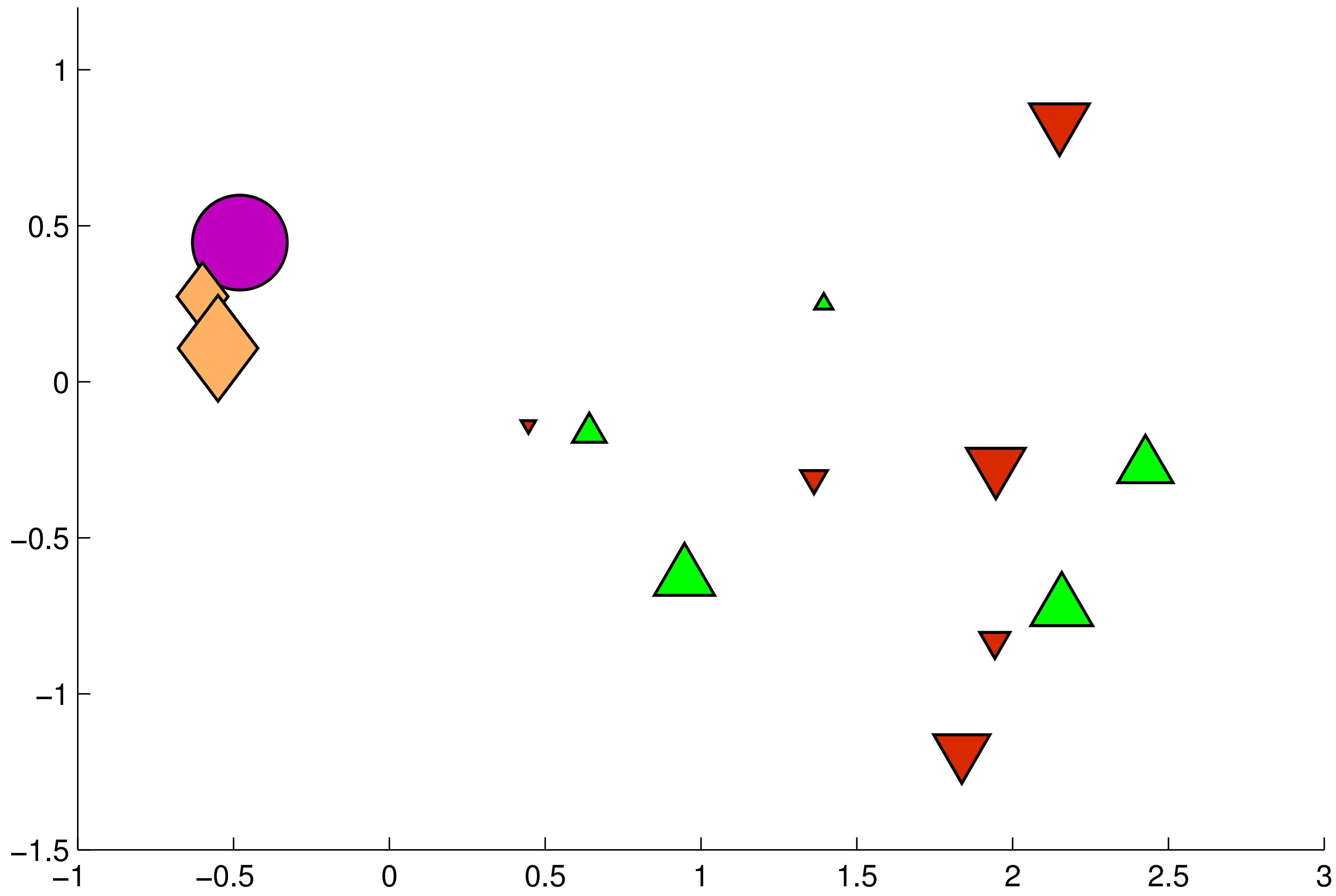
Wasserstein Barycenters

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



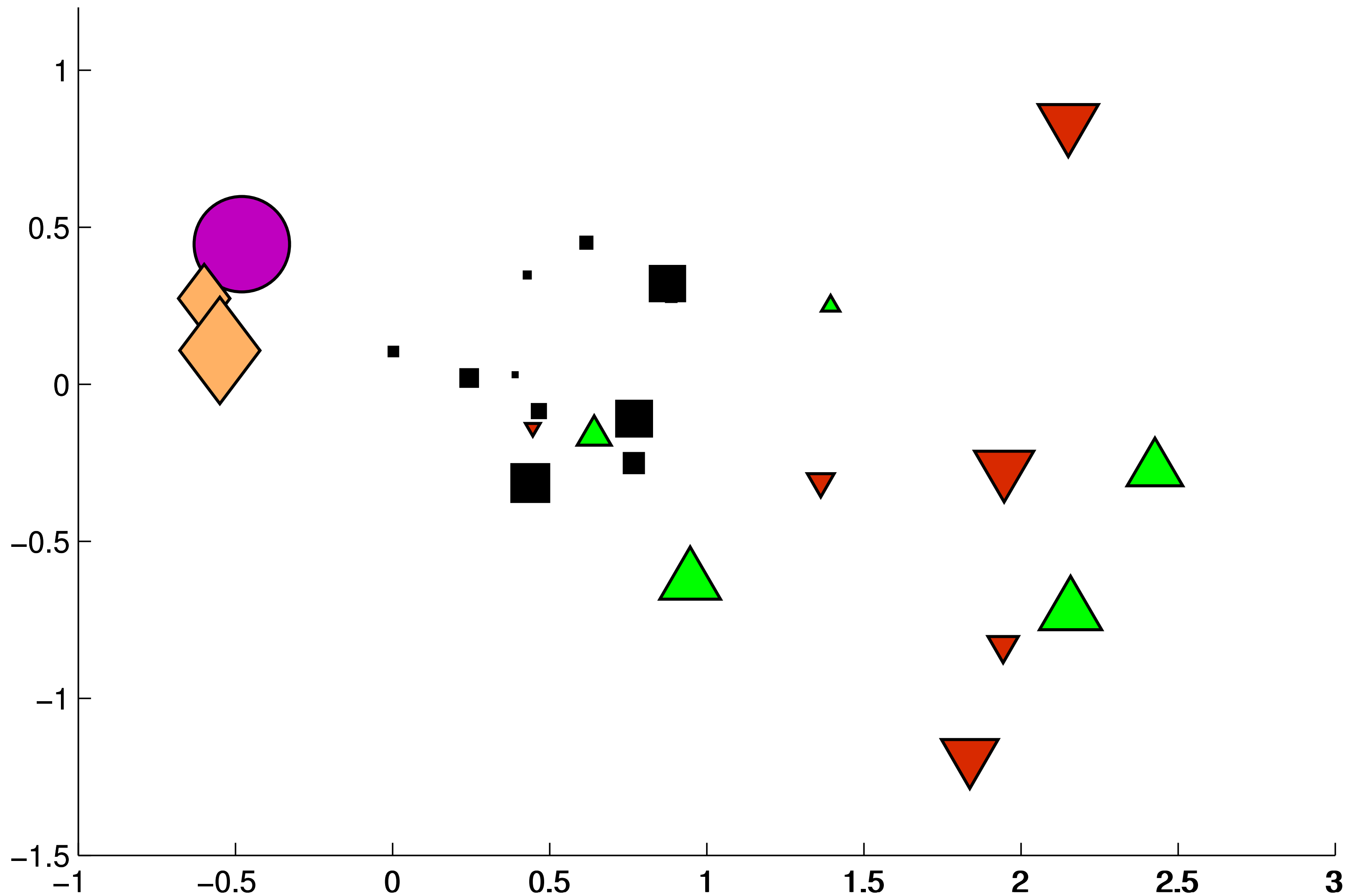
Multimarginal Formulation

- Exact solution (W_2) using MM-OT. **[Agueh'11]**



Multimarginal Formulation

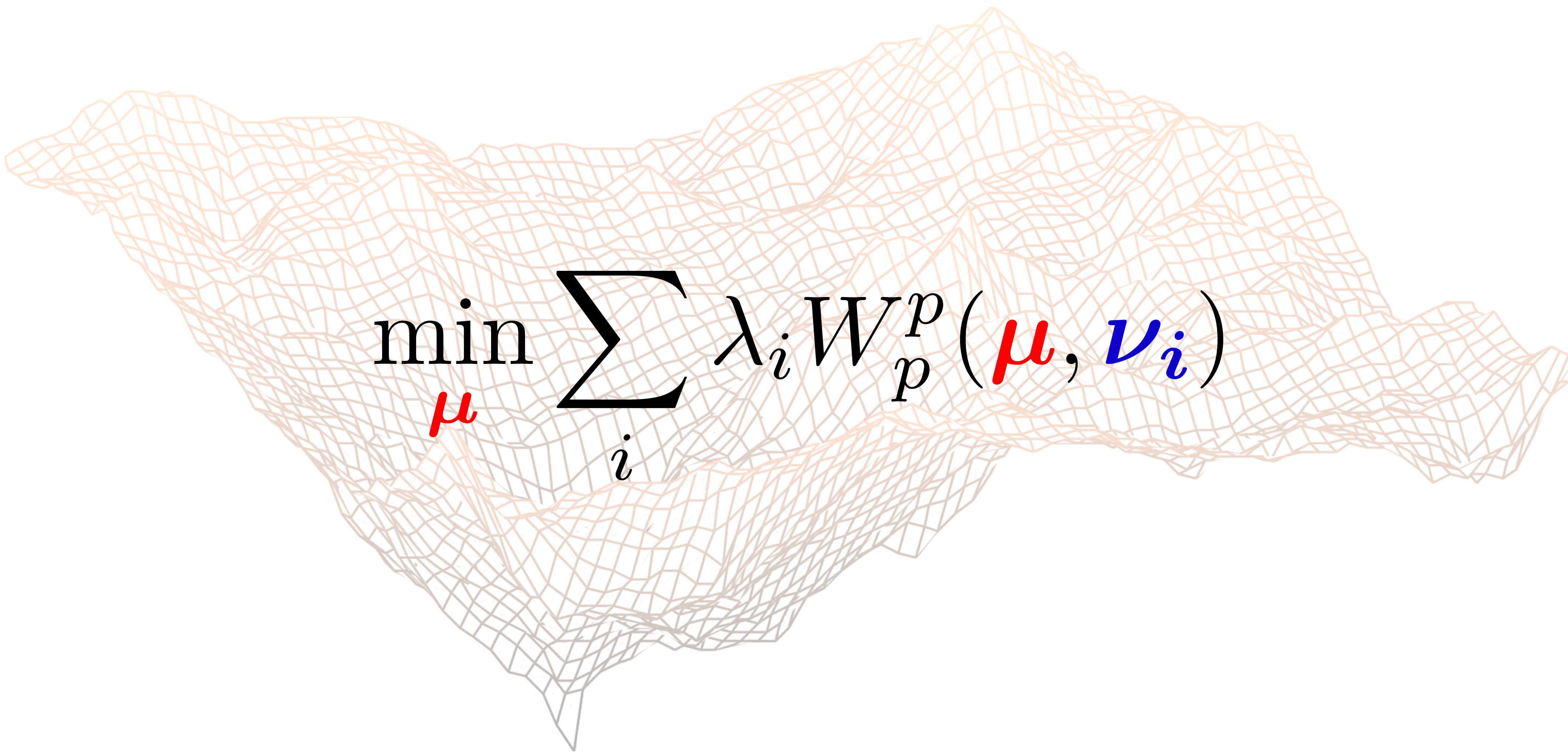
- Exact solution (W_2) using MM-OT. **[Agueh'11]**



$$\text{If } |\operatorname{supp} \boldsymbol{\nu_i}| = \boldsymbol{n_i}, \text{ LP of size } \left(\prod_i \boldsymbol{n_i}, \sum_i \boldsymbol{n_i}\right)$$

Finite Case, LP Formulation

- When Ω is a finite set, metric M, another LP.



$$\min_{\boldsymbol{\mu}} \sum_i \lambda_i W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$

Finite Case, LP Formulation

- When Ω is a finite set, metric M, another LP.

$$\min_{\boldsymbol{P_1},\cdots,\boldsymbol{P_N},\boldsymbol{a}} \sum_{i=1}^{N} \lambda_i \langle \boldsymbol{P_i}, M \rangle$$

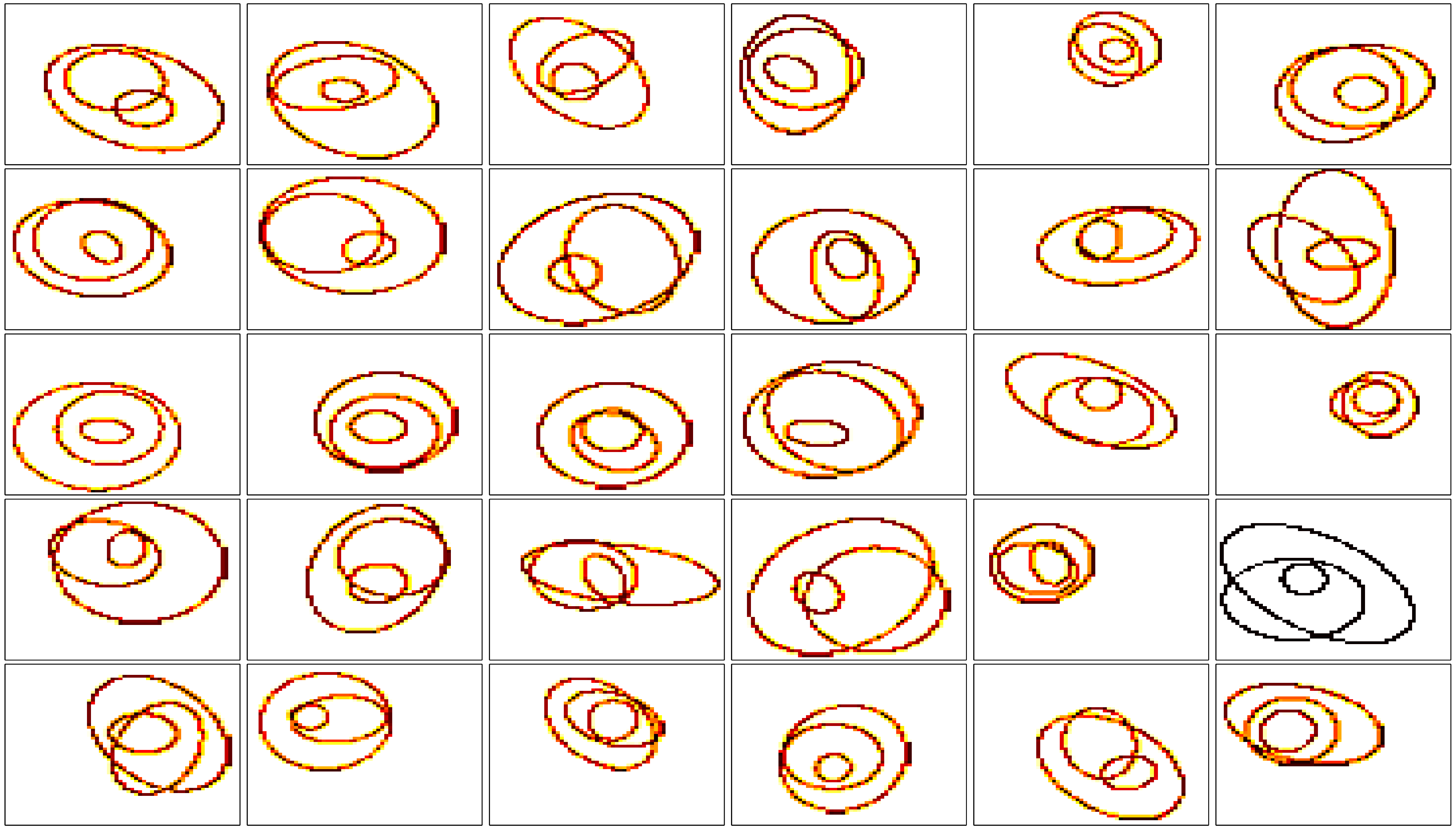
$$\text{s.t.} \quad \boldsymbol{P_i}^T \mathbf{1}_n = \boldsymbol{b_i}, \forall i \leq N,$$

$$\boldsymbol{P_1} \mathbf{1}_n = \cdots = \boldsymbol{P_N} \mathbf{1}_d = \boldsymbol{a}.$$

If $|\Omega| = n$, LP of size $(Nn^2, (2N-1)n)$; unstable

Primal Descent on Regularized W

$$\min_{\mu \in Q \subset \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_{\gamma}(\mu, \nu_i)$$

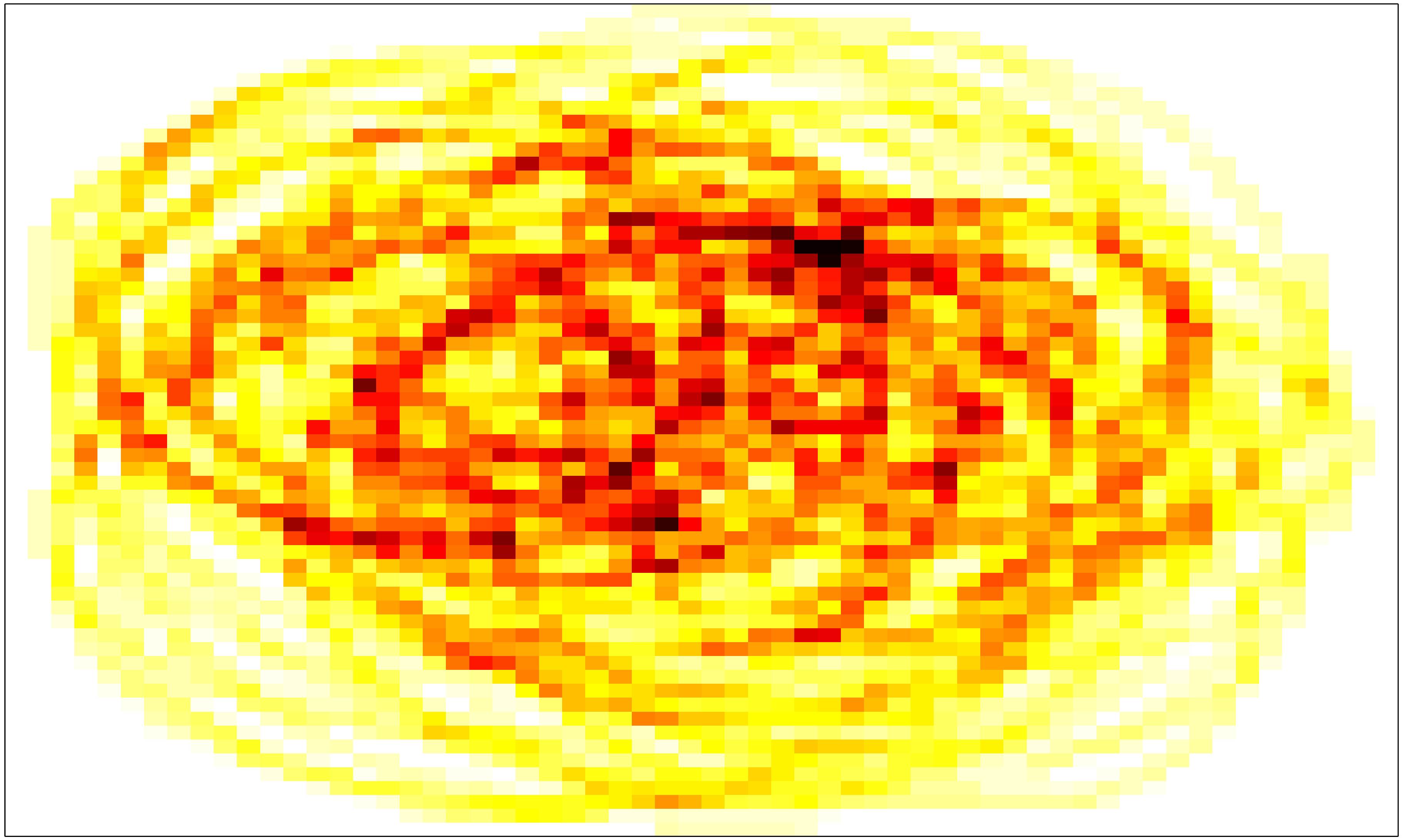


Fast Computation of Wasserstein Barycenters
International Conference on Machine Learning 2014

[CD'14]

Primal Descent on Regularized W

$$\min_{\boldsymbol{\mu} \in \boldsymbol{Q} \subset \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_{\gamma}(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$

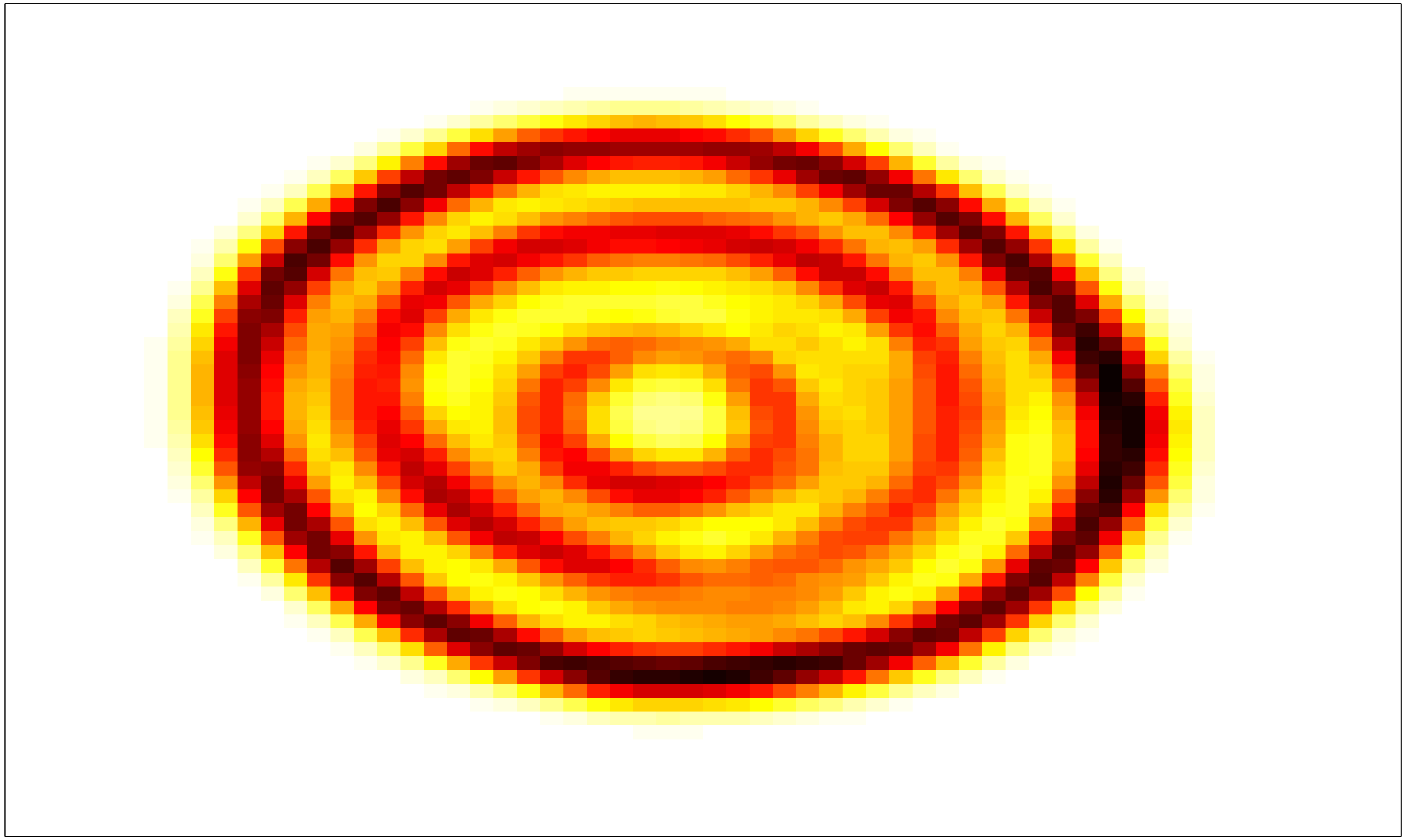


Fast Computation of Wasserstein Barycenters
International Conference on Machine Learning 2014

[CD'14]

Primal Descent on Regularized W

$$\min_{\mu \in Q \subset \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_{\gamma}(\mu, \nu_i)$$



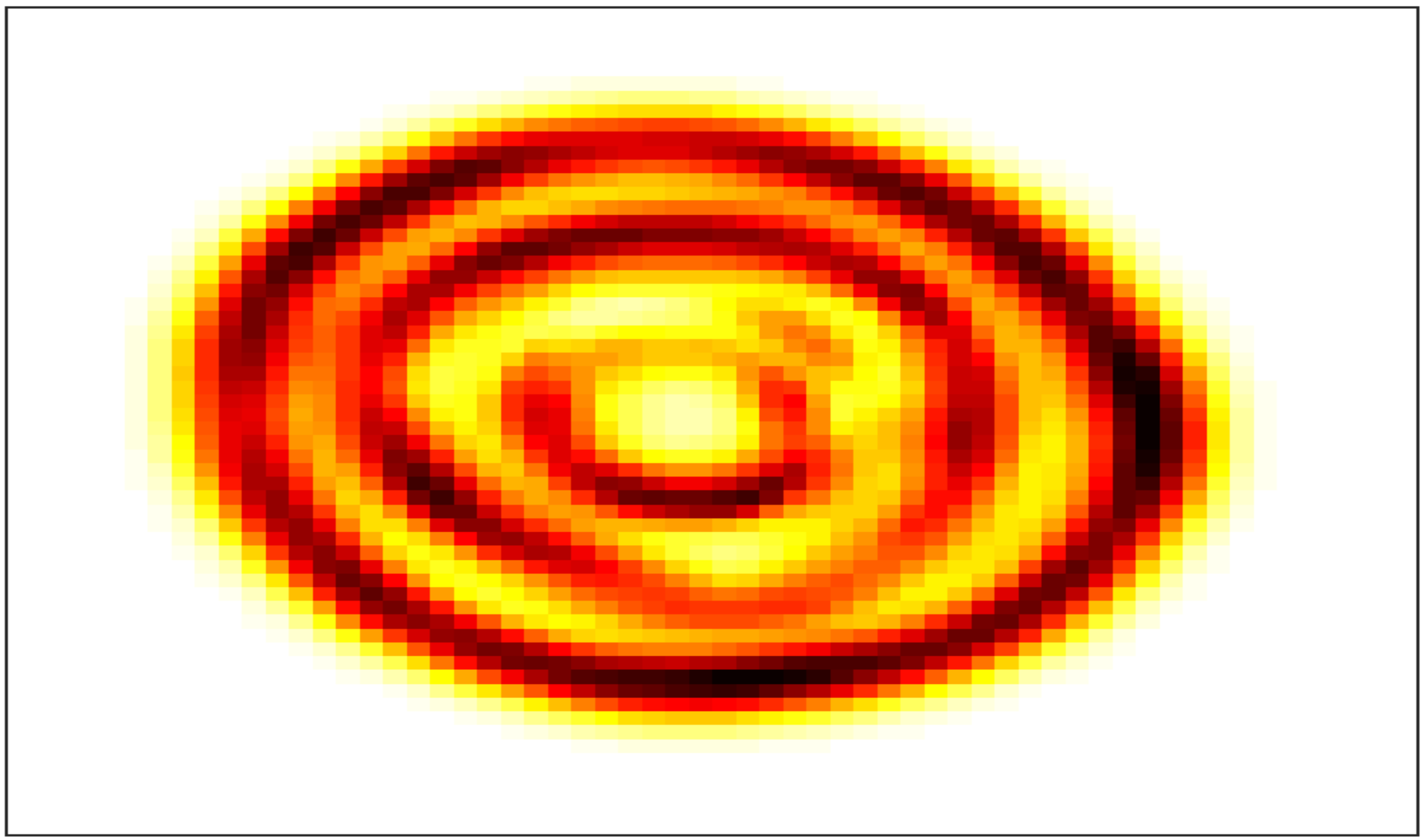
Fast Computation of Wasserstein Barycenters
International Conference on Machine Learning 2014 **[CD'14]**

Primal Descent on Algorithmic W

$$\min_{\boldsymbol{\mu} \in \boldsymbol{Q} \subset \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_{\boldsymbol{L}}(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$

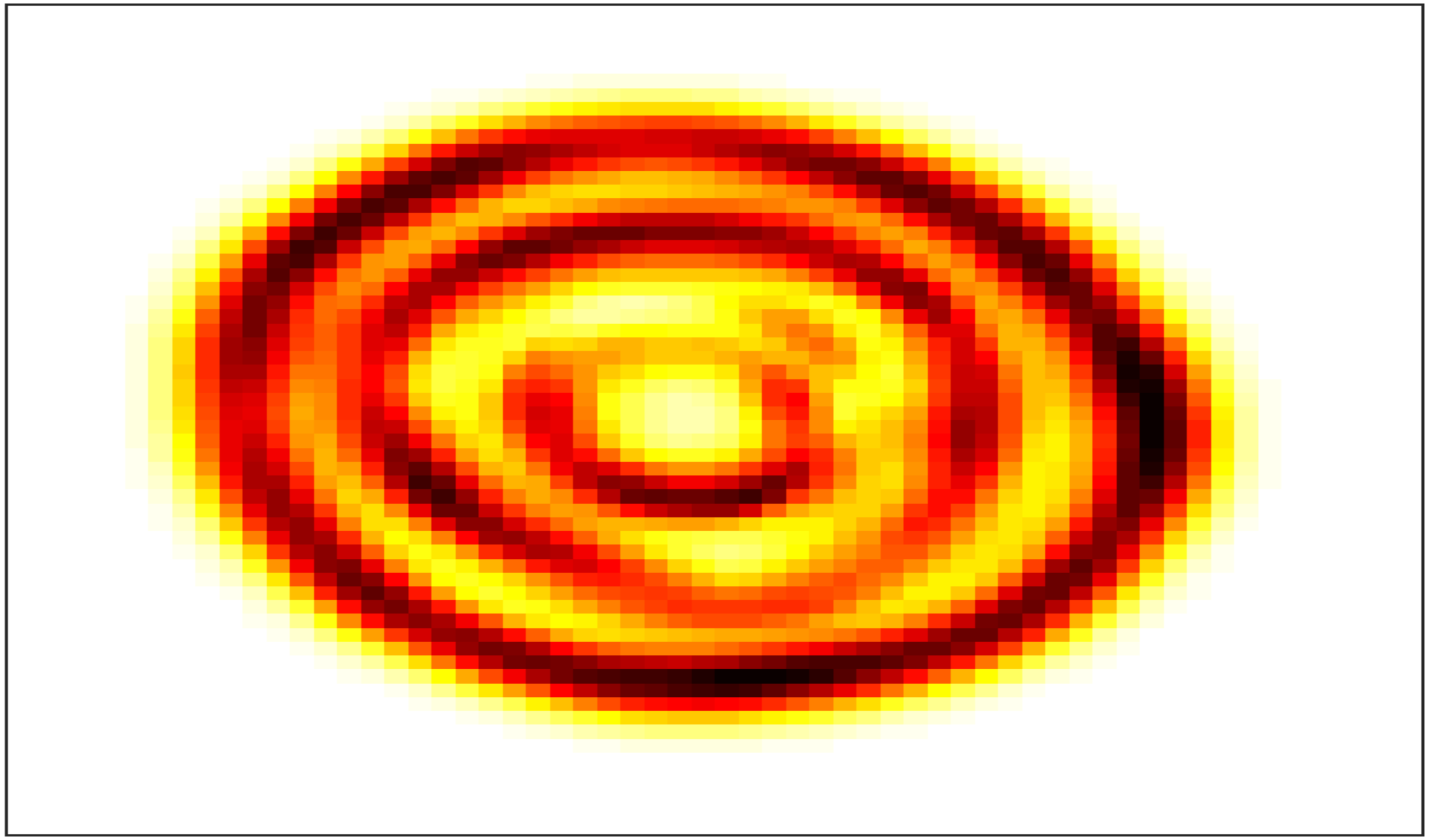
Primal Descent on Algorithmic W

$$\min_{\mu \in Q \subset \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_{L}(\mu, \nu_i)$$



Primal Descent on Algorithmic W

$$\min_{\mu \in Q \subset \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_{L}(\mu, \nu_i)$$



not a convex problem

Inverse Wasserstein Problems

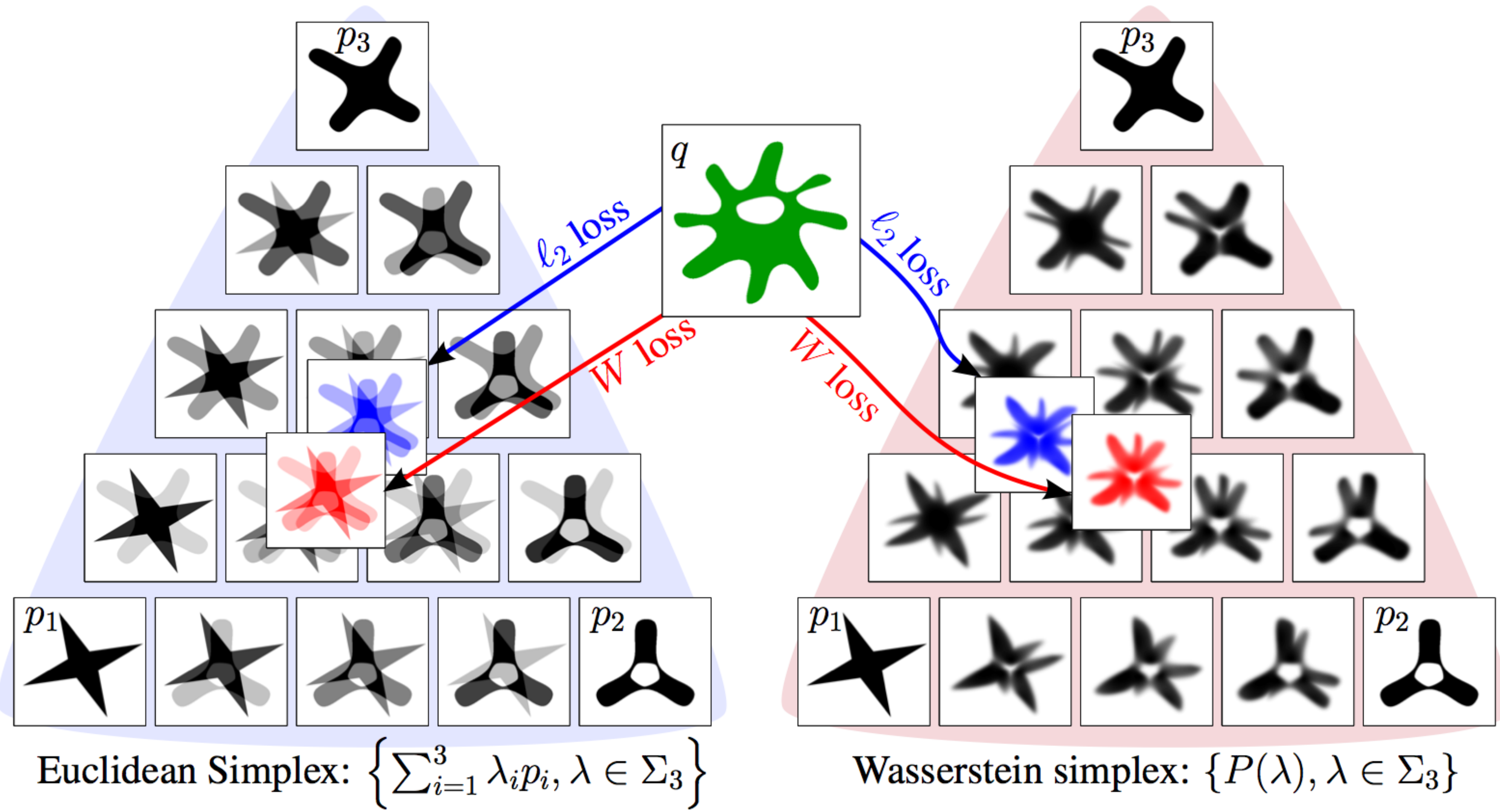
- consider Barycenter operator:

$$\boldsymbol{b}(\lambda) \stackrel{\text{def}}{=} \underset{\boldsymbol{a}}{\operatorname{argmin}} \sum_{i=1}^{N} \lambda_i W_\gamma(\boldsymbol{a}, \boldsymbol{b_i})$$

- address now **Wasserstein inverse problems**:

$$\text{Given } \boldsymbol{a}, \text{ find } \underset{\lambda \in \Sigma_N}{\operatorname{argmin}} \mathcal{E}(\lambda) \stackrel{\text{def}}{=} \operatorname{Loss}(\boldsymbol{a}, \boldsymbol{b}(\lambda))$$

The Wasserstein Simplex



Barycenters = Fixed Points

Prop. [BCCNP'15] Consider $\boldsymbol{B} \in \Sigma_d^N$ and let $\boldsymbol{U_0} = \mathbf{1}_{d\times N}$, and then for $l \geq 0$:

$$\boldsymbol{b}^l \stackrel{\text{def}}{=} \exp\left(\log\left(K^T \boldsymbol{U_l}\right)\lambda\right); \begin{cases} \boldsymbol{V_{l+1}} \stackrel{\text{def}}{=} \dfrac{\boldsymbol{b}^l \mathbf{1}_N^T}{K^T \boldsymbol{U_l}}, \\ \boldsymbol{U_{l+1}} \stackrel{\text{def}}{=} \dfrac{\boldsymbol{B}}{K \boldsymbol{V_{l+1}}}. \end{cases}$$

Using Truncated Barycenters

- instead of using the exact barycenter

$$\underset{\lambda \in \Sigma_N}{\operatorname{argmin}} \mathcal{E}(\lambda) \stackrel{\text{def}}{=} \text{Loss}(\boldsymbol{a}, \boldsymbol{b}(\lambda))$$

- use instead the L-iterate barycenter

$$\underset{\lambda \in \Sigma_N}{\operatorname{argmin}} \mathcal{E}^{(L)}(\lambda) \stackrel{\text{def}}{=} \text{Loss}(\boldsymbol{a}, \boldsymbol{b}^{(L)}(\lambda))$$

- Differente using **the chain rule.**

$$\nabla \mathcal{E}^{(L)}(\lambda) = [\partial \boldsymbol{b}^{(L)}]^T(\boldsymbol{g}), \;\; \boldsymbol{g} \stackrel{\text{def}}{=} \nabla \text{Loss}(\boldsymbol{a}, \cdot)|_{\boldsymbol{b}^{(L)}(\lambda)}.$$

Gradient / Barycenter Computation

function $\textsc{Sinkhorn-Differentiate}((p_s)_{s=1}^S, q, \lambda)$

$\forall\, s, b_s^{(0)} \leftarrow \mathbb{1}$

$(w, r) \leftarrow (0^S, 0^{S\times N})$

for $\ell = 1, 2, \ldots, L$ *// Sinkhorn loop*

$\forall\, s, \varphi_s^{(\ell)} \leftarrow K^\top \frac{p_s}{K b_s^{(\ell-1)}}$

$p \leftarrow \prod_s \left(\varphi_s^{(\ell)}\right)^{\lambda_s}$

$\forall\, s, b_s^{(\ell)} \leftarrow \frac{p}{\varphi_s^{(\ell)}}$

$g \leftarrow \nabla \mathcal{L}(p, q) \odot p$

for $\ell = L, L-1, \ldots, 1$ *// Reverse loop*

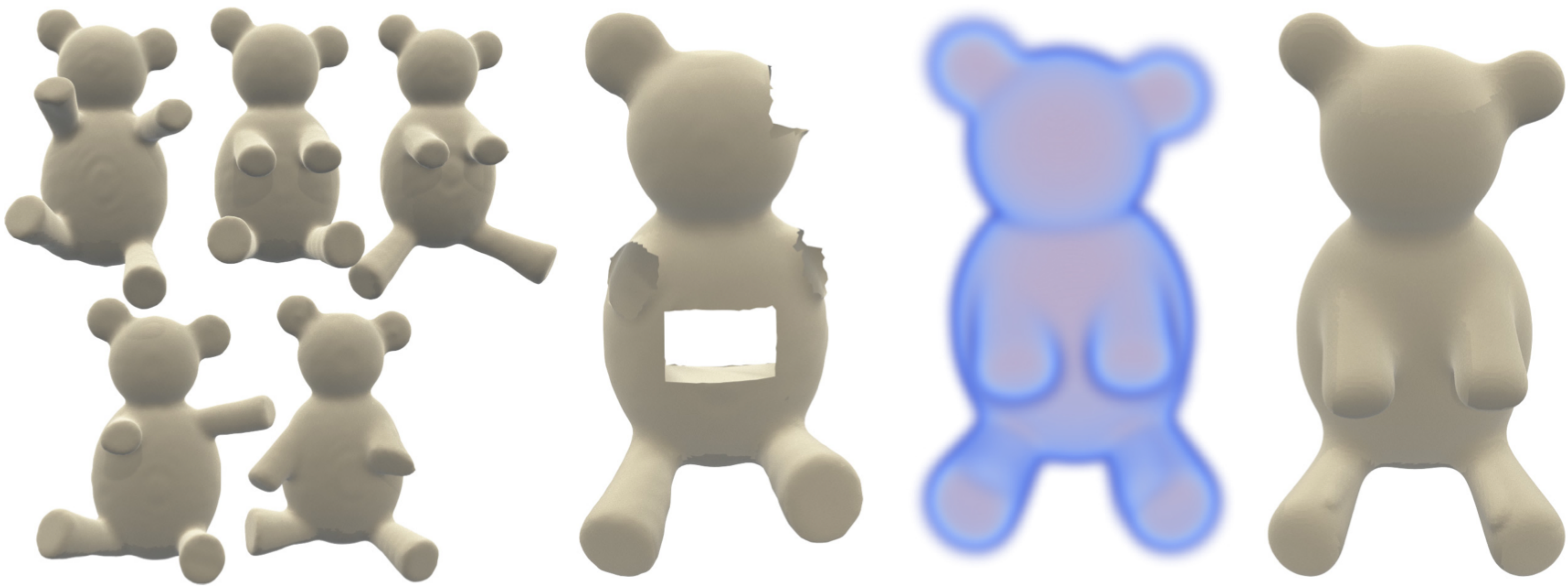
$\forall\, s, w_s \leftarrow w_s + \langle \log \varphi_s^{(\ell)}, g \rangle$

$\forall\, s, r_s \leftarrow -K^\top \left(K\left(\frac{\lambda_s g - r_s}{\varphi_s^{(\ell)}}\right) \odot \frac{p_s}{(K b_s^{(\ell-1)})^2}\right) \odot b_s^{(\ell-1)}$

$g \leftarrow \sum_s r_s$

return $P^{(L)}(\lambda) \leftarrow p, \nabla \mathcal{E}_L(\lambda) \leftarrow w$

Application: Volume Reconstruction



Shape database $(p_1, \ldots, p_5)$ | Input shape q | Projection $P(\lambda)$ | Iso-surface

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

Application: Color Grading

Application: Color Grading

$\lambda_0 = 0.03$

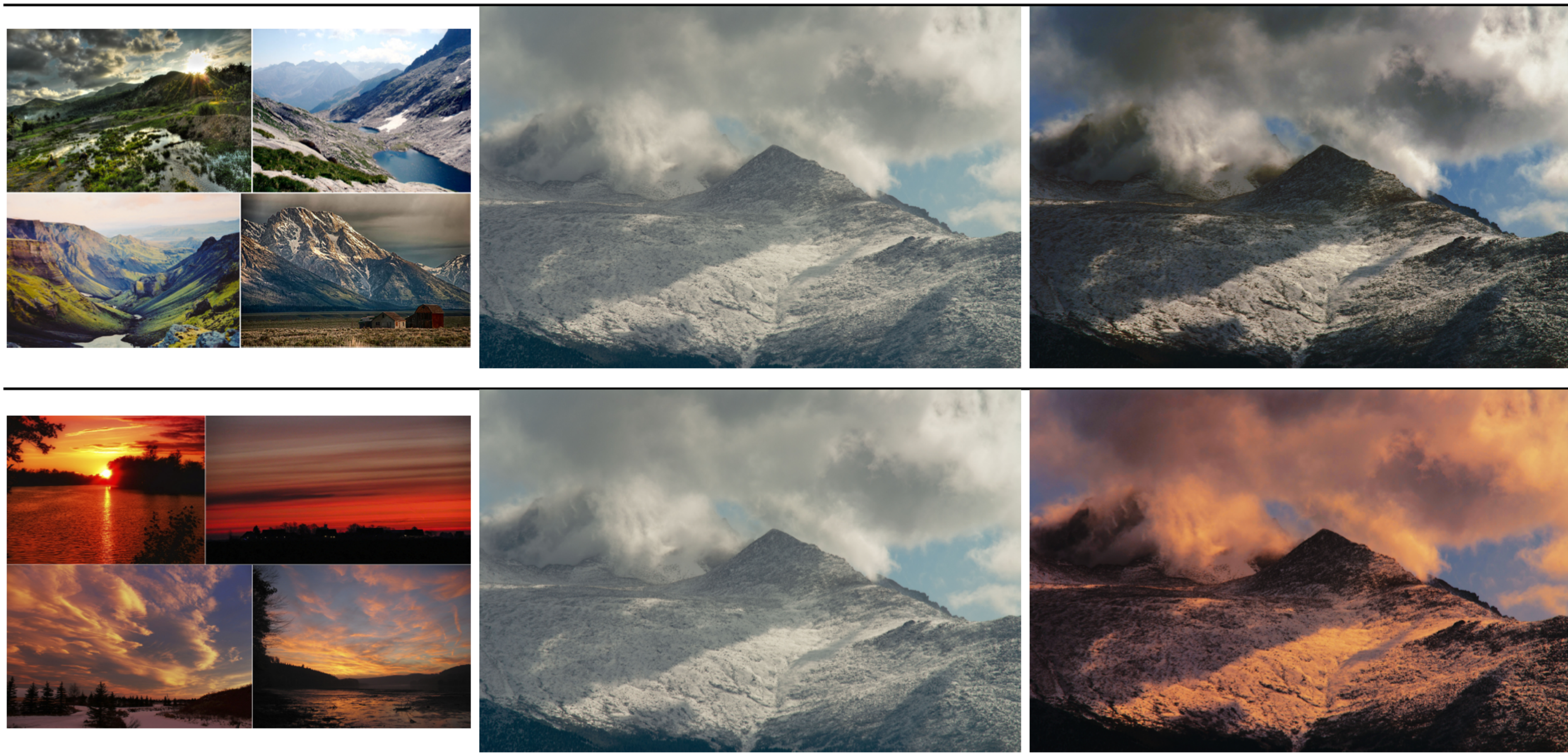
$\lambda_1 = 0.12$

$\lambda_2 = 0.40$

$\lambda_3 = 0.43$

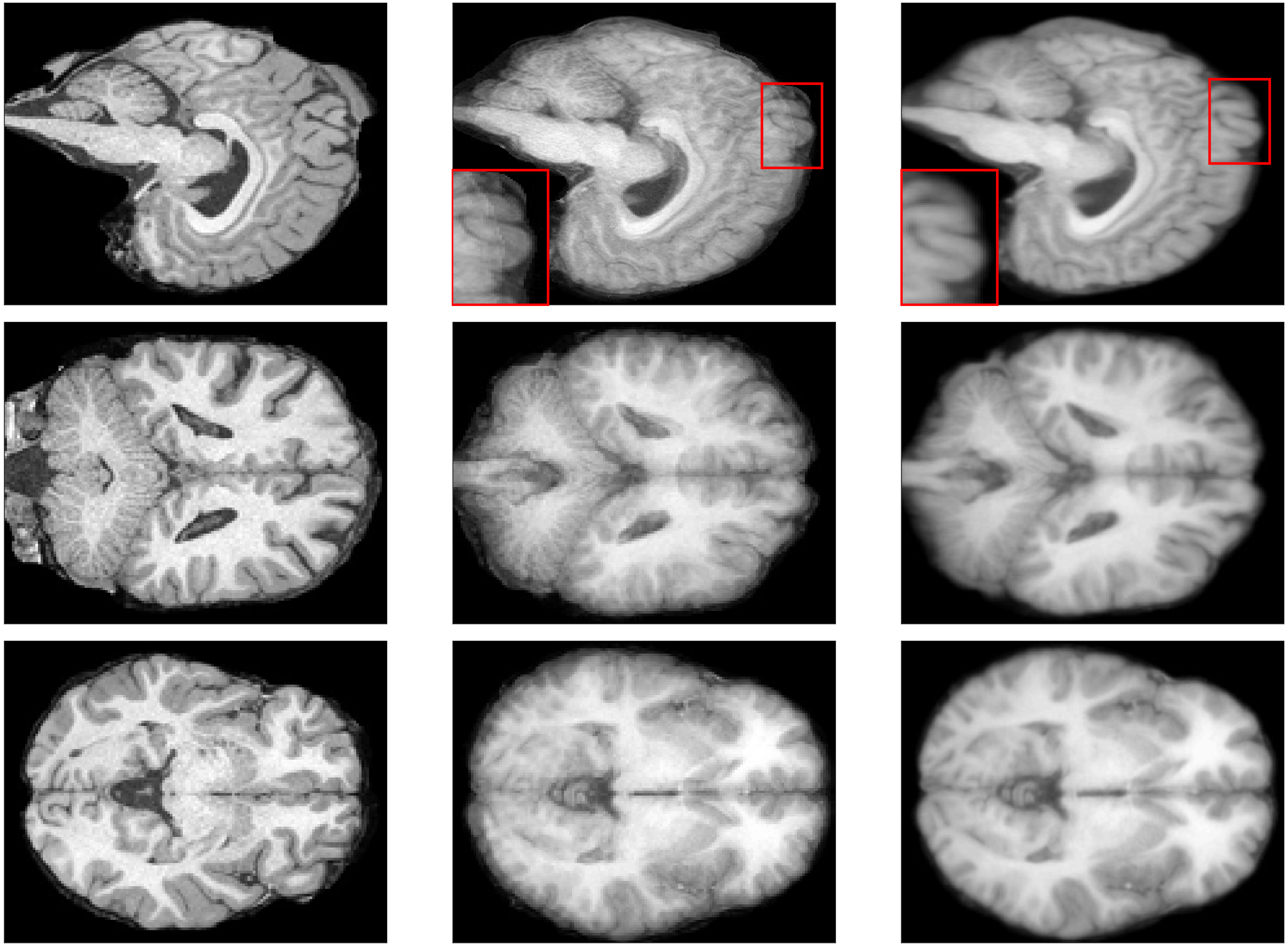
Application: Color Grading

Application: Color Grading



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

Application: Brain Mapping

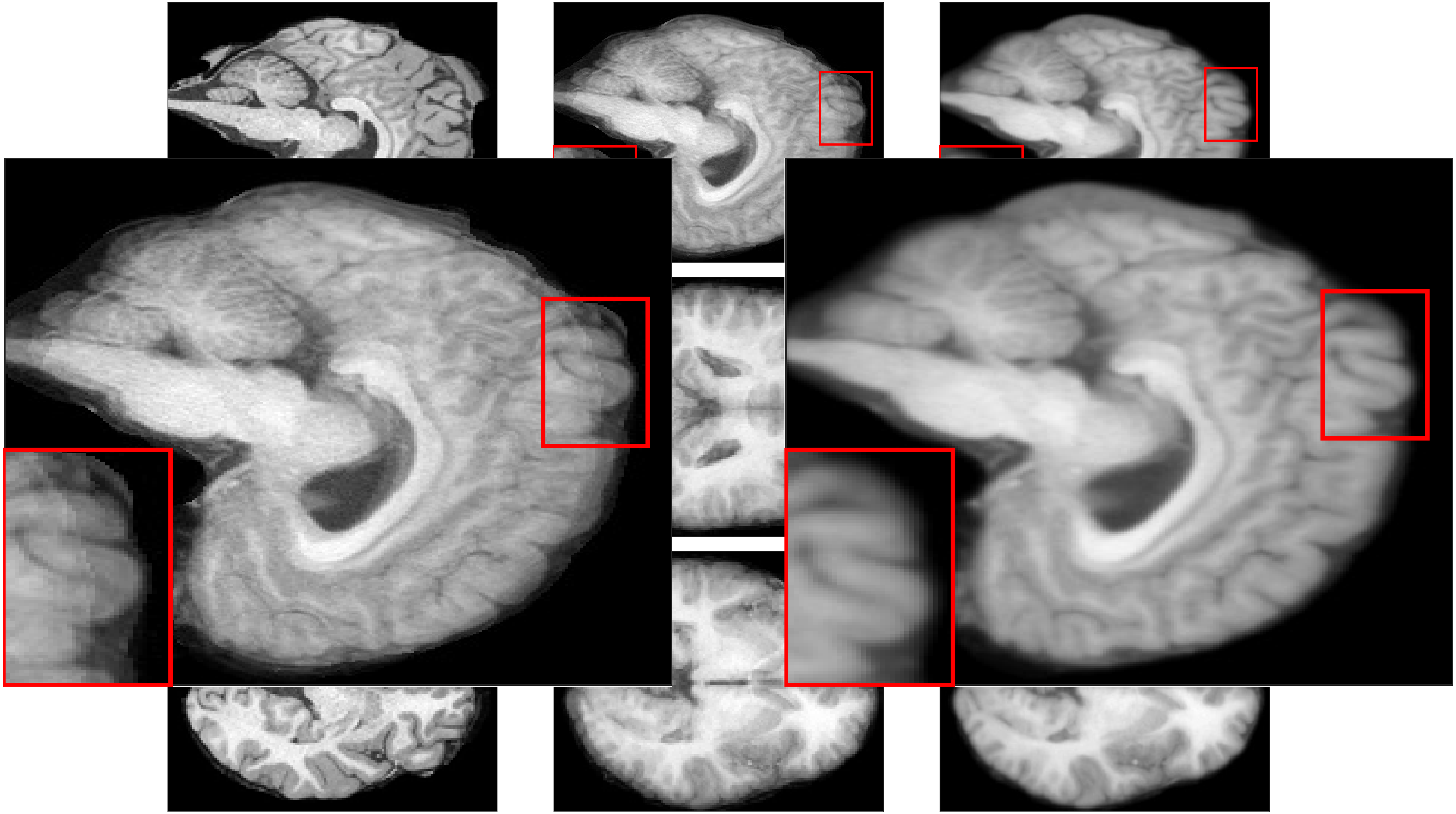


Original

Euclidean projection

Wasserstein projection

Application: Brain Mapping



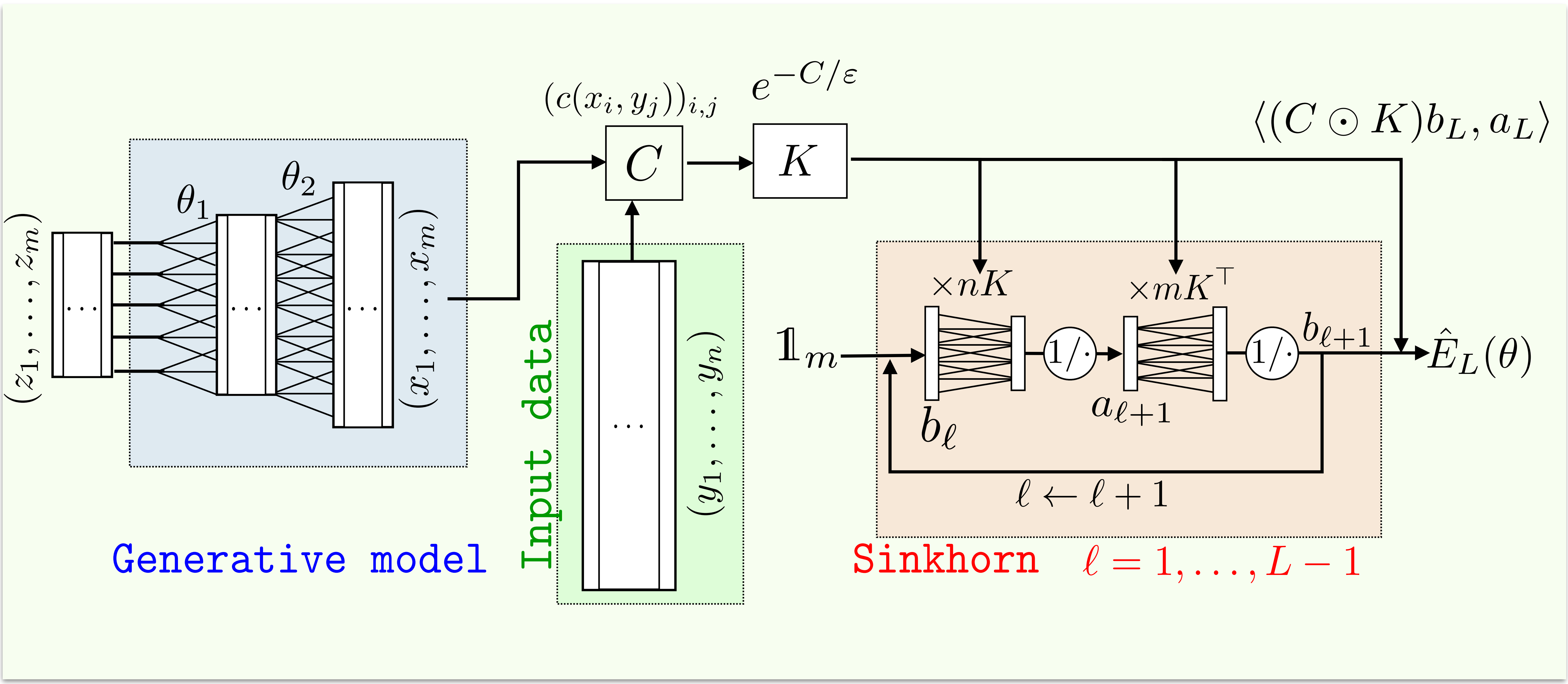
Original

Euclidean projection

Wasserstein projection

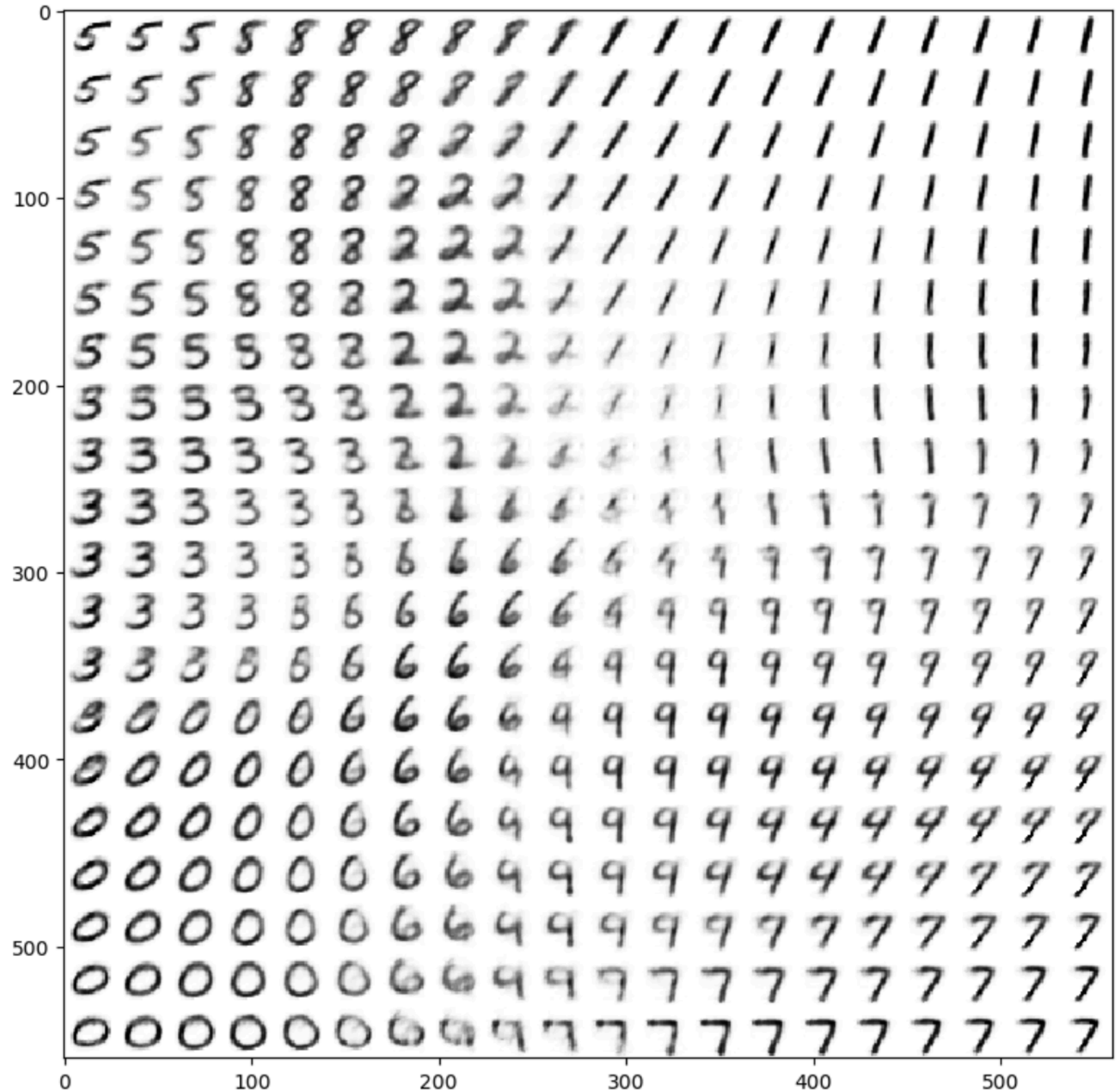
At Last: Application to Generative Models

Approximate W loss by the transport cost $\bar{W}_L$ after L Sinkhorn iterations.

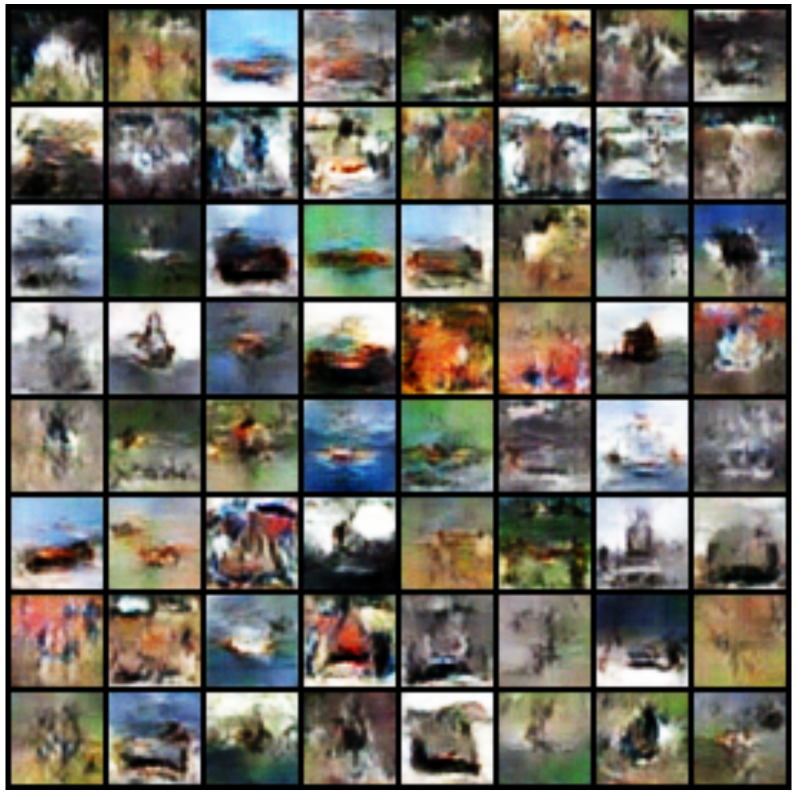


[GPC'17]

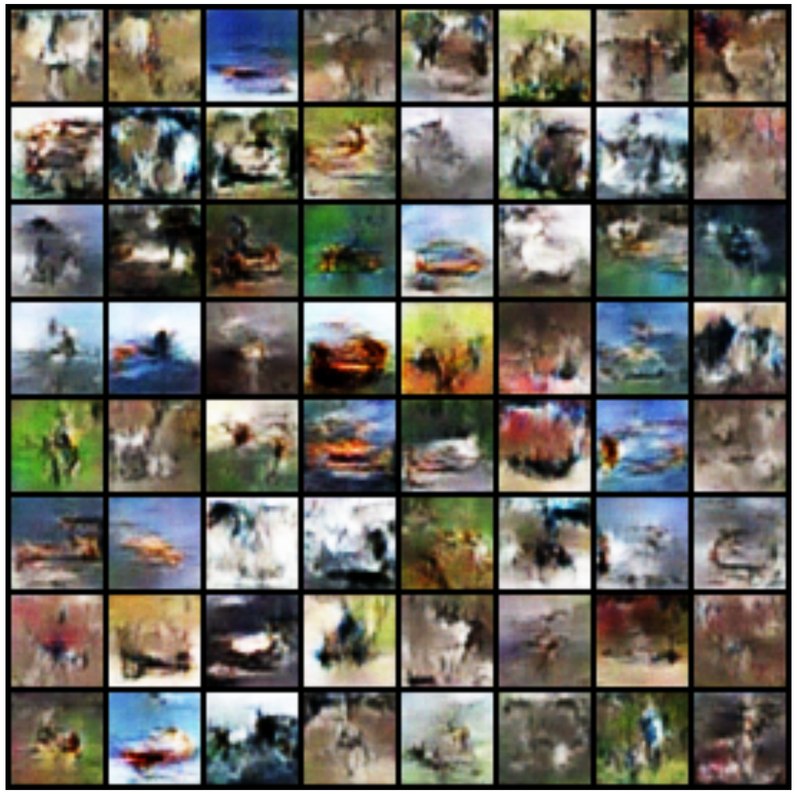
Example: MNIST, Learning f_{θ}



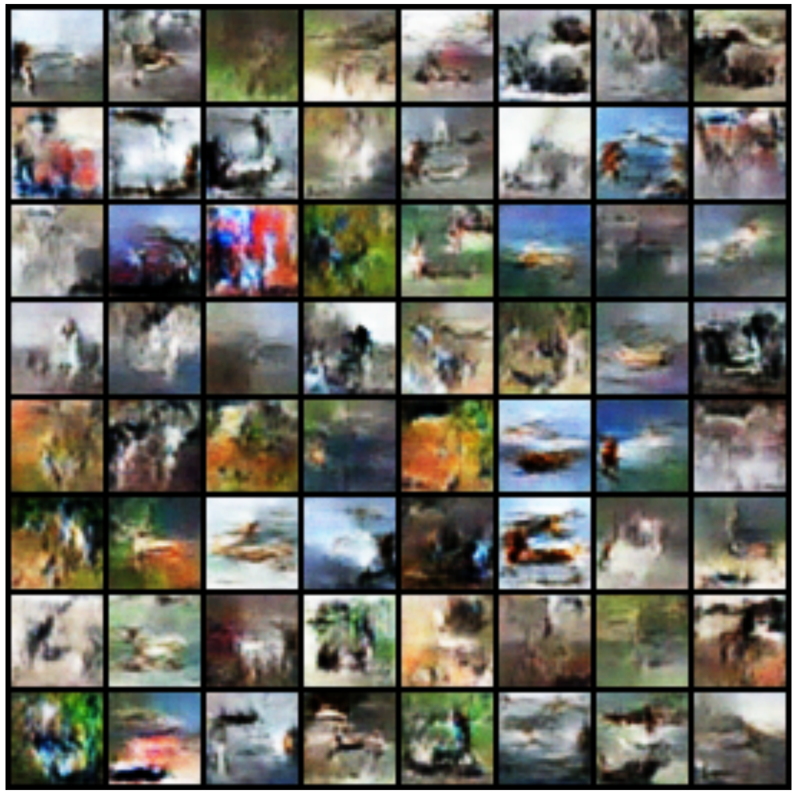
Example: Generation of Images



MMD-GAN



gamma = 1000



gamma=10

- CIFAR-10 images
- In these examples the cost function is also learned adversarially, as a NN mapping onto feature vectors.

Concluding Remarks

- *Regularized OT* is much faster than OT.
- *Regularized OT* can interpolate between W and the *MMD / Energy distance* metrics.
- The solution of *regularized OT* is *"auto-differentiable"*.
- **Many open problems remain!**

Sat Dec 9th 08:00 AM -- 06:30 PM @ None — Workshop

Optimal Transport and Machine Learning

Olivier Bousquet · Marco Cuturi · Gabriel Peyré · Fei Sha · Justin Solomon

NIPS'17 WORKSHOP

Dates n/a. @ TBA — Tutorial

A Primer on Optimal Transport

Marco Cuturi · Justin M Solomon

NIPS'17 TUTORIAL