A Primer on

Optimal Transport

Marco Cuturi

ENSAE ParisTech

École nationale de la statistique et de l'administration économique

université PARIS-SACLAY

Justin Solomon

MIT CSAIL

What is Optimal Transport?

The natural geometry for probability measures


Monge


Kantorovich


Koopmans


Dantzig


Brenier


Otto


McCann

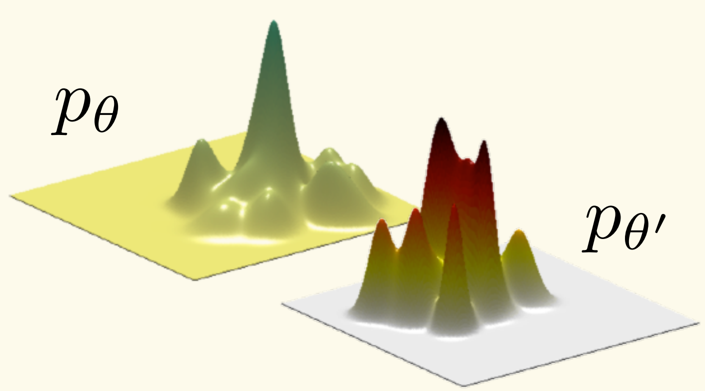

Villani

Nobel ’75

Fields ’10

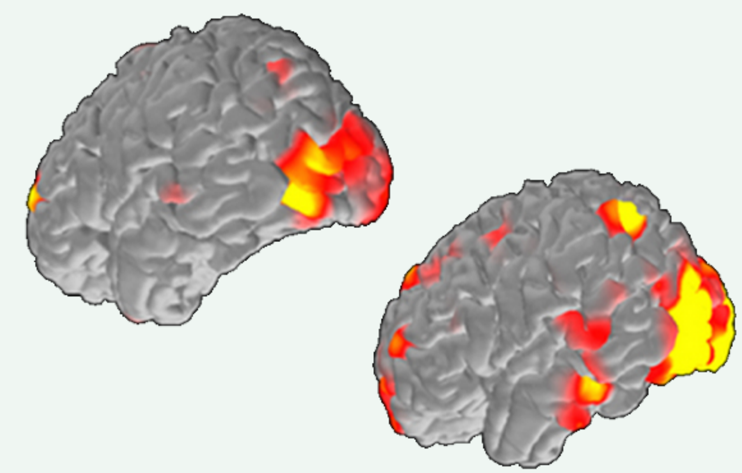
What is Optimal Transport?

The natural geometry for probability measures



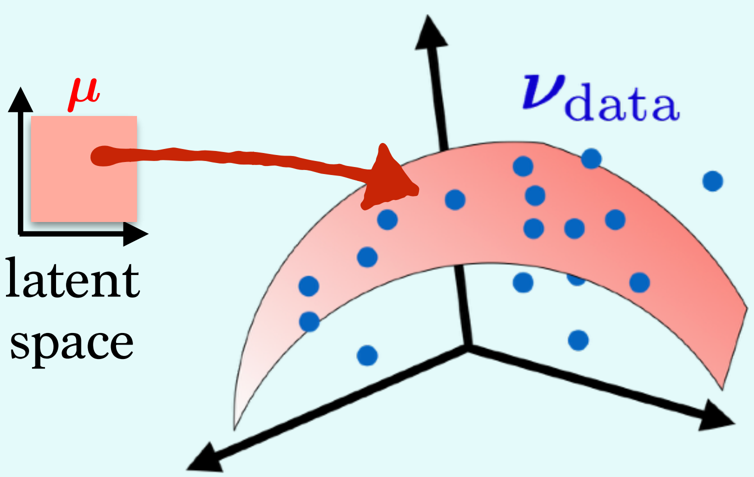
Statistical Models

Bags of features



Brain Activation Maps

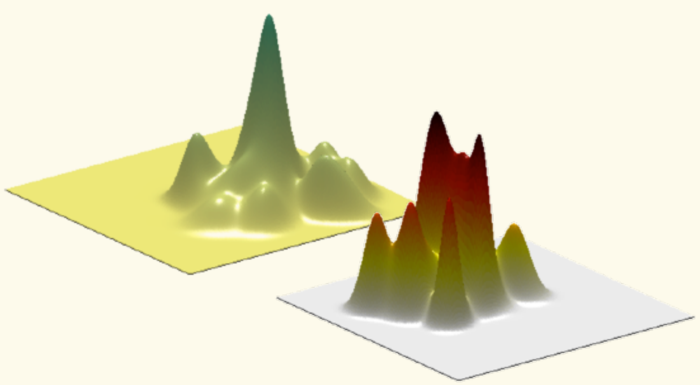
Generative Models vs. data



Color Histograms

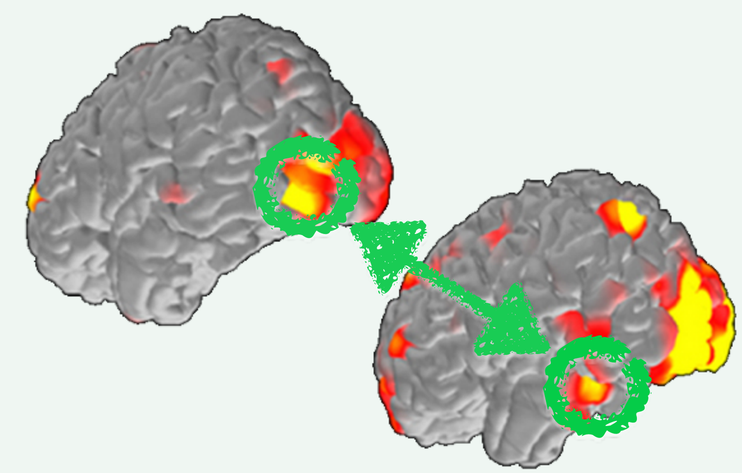
What is Optimal Transport?

The natural geometry for probability measures supported on a geometric space.



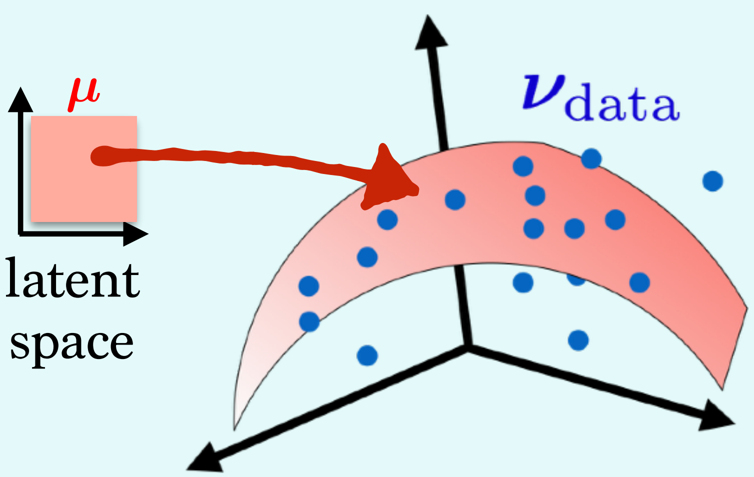
Statistical Models

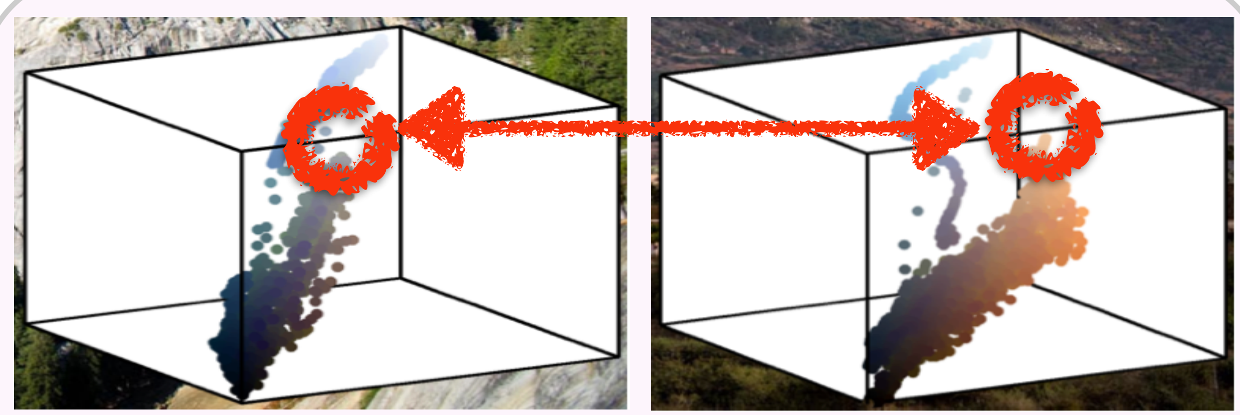
Bags of Features



Brain Activation Maps

Generative Models vs. Data

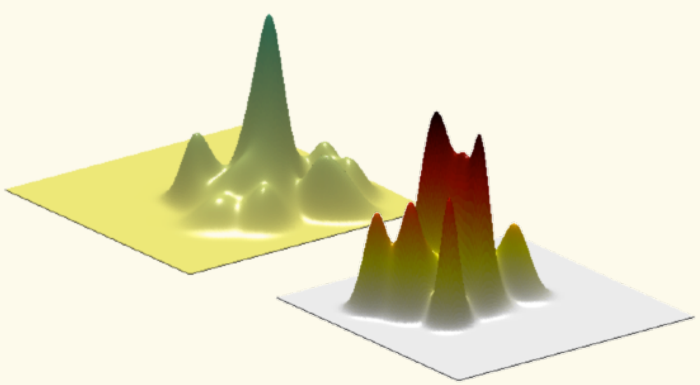




Color Histograms

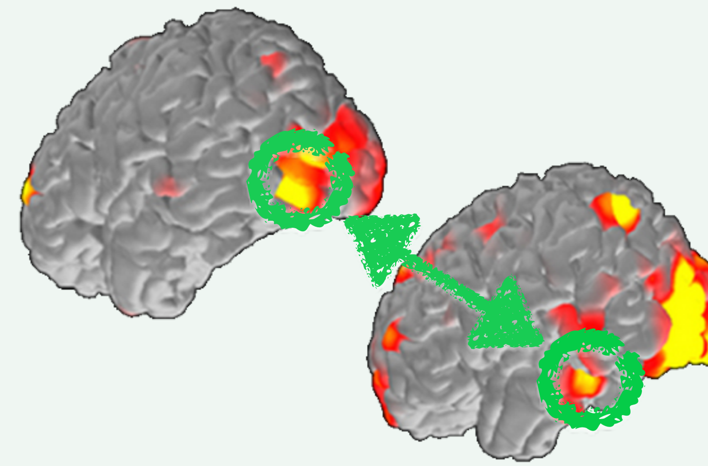
What is Optimal Transport?

The natural geometry for probability measures supported on a geometric space.



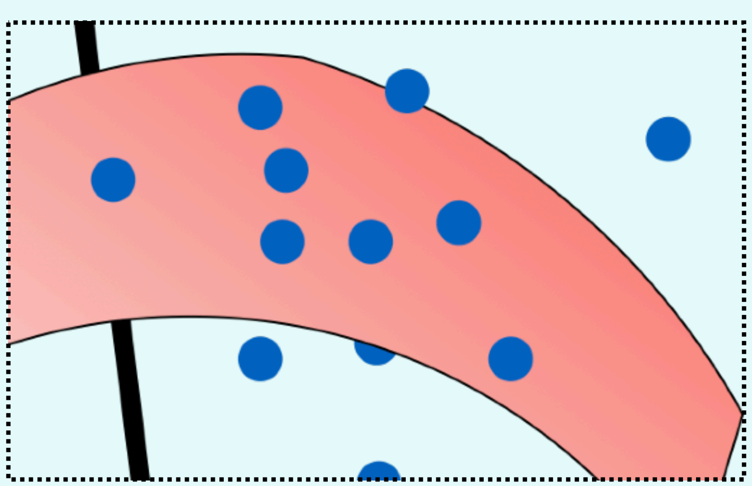
Statistical Models

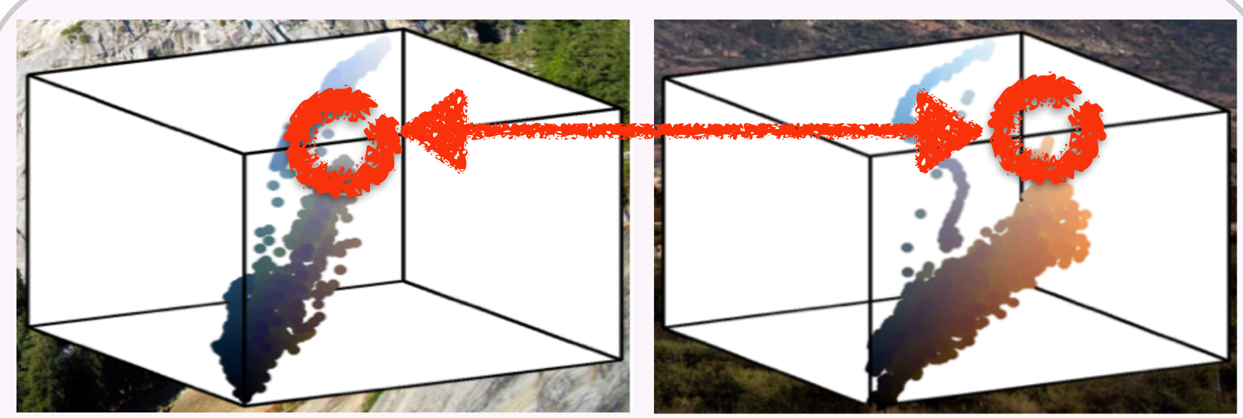
Bags of Features



Brain Activation Maps

Generative Models vs. Data

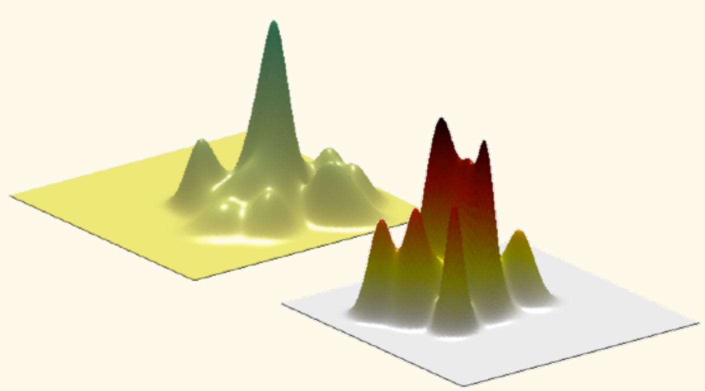




Color Histograms

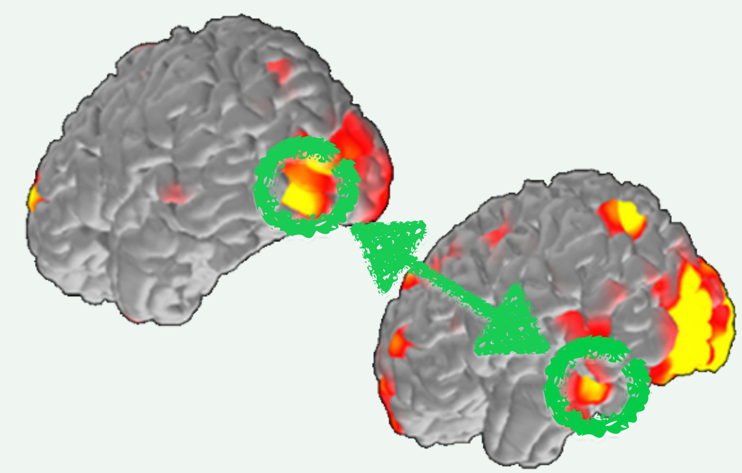
What is Optimal Transport?

The natural geometry for probability measures supported on a geometric space.



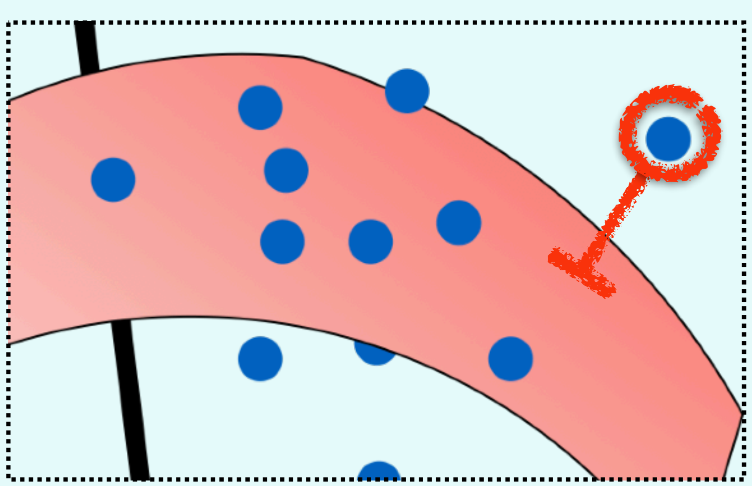
Statistical Models

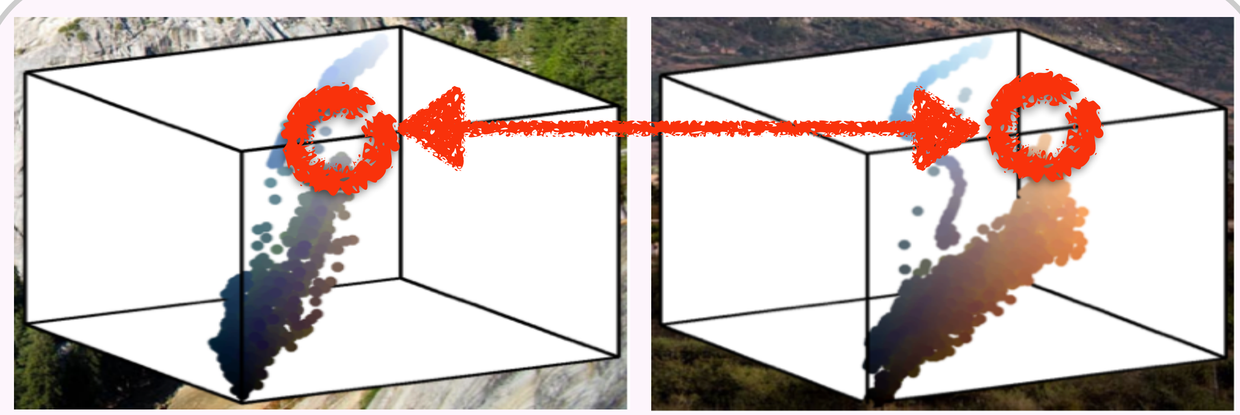
Bags of Features



Brain Activation Maps

Generative Models vs. Data





Color Histograms

Today's Outline

1. Introduction to optimal transport
2. Optimal transport algorithms
3. Applications (W as a loss)
4. Applications (W for estimation)

More...

This Saturday: **OT & ML Workshop**

7 Talks by: Jacob, Kraig, Andoni, Gangbo, Bottou, Flamary, Bach

17 posters and spotlight presentations.

Organizers: Bousquet, Cuturi, Peyré, Sha, Solomon

Survey and slides: **https://optimaltransport.github.io/**

Introduction to OT

- Two examples: moving earth & soldiers
- Monge problem, Kantorovich problem
- OT as geometry, OT as a loss function

Origins: Monge Problem (1781)

6(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 Problem (1781)

MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

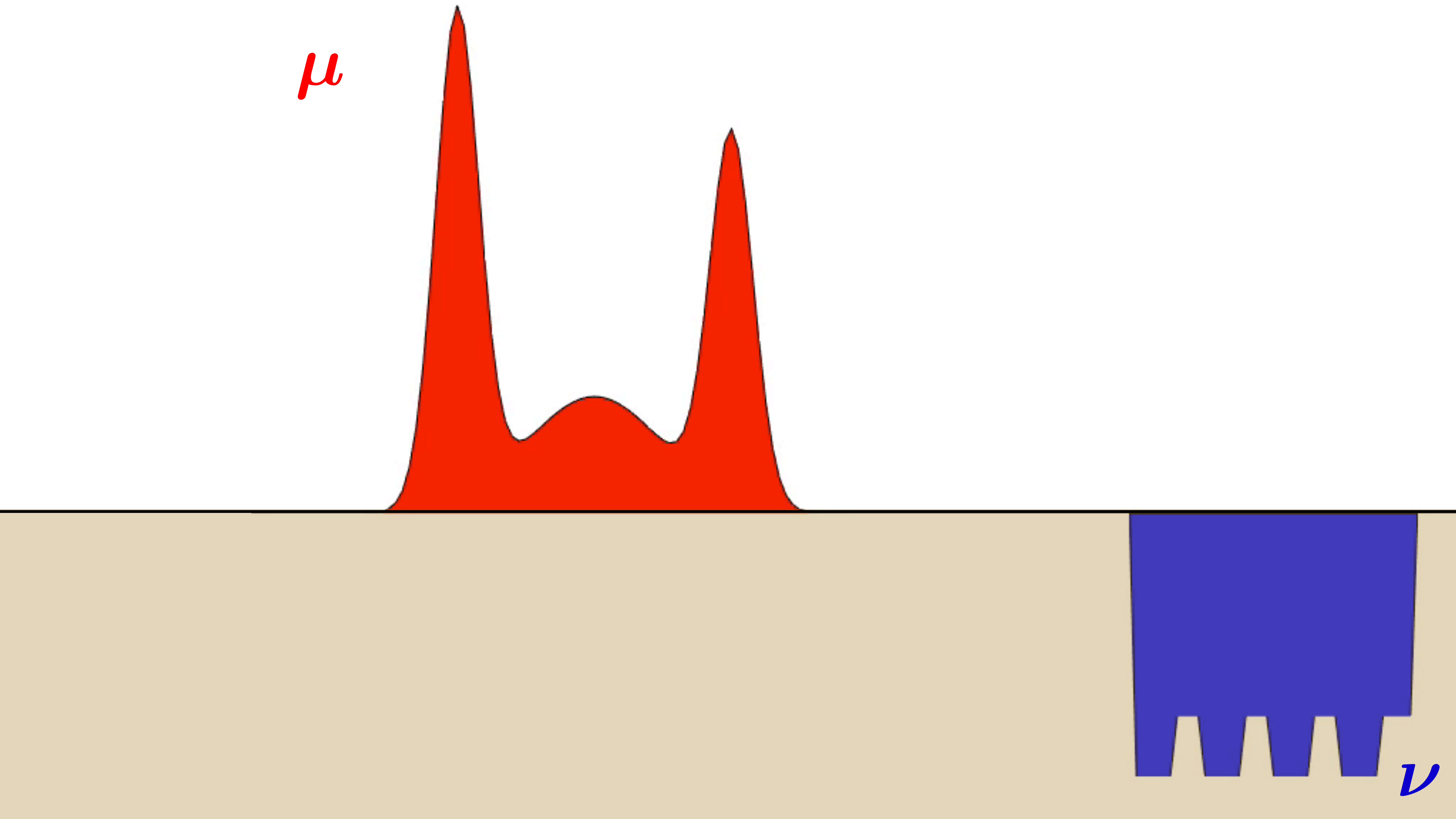
THÉORIE DES DÉBLAIS

ET DES REMBLAIS.

When one has to bring earth from one place to another…

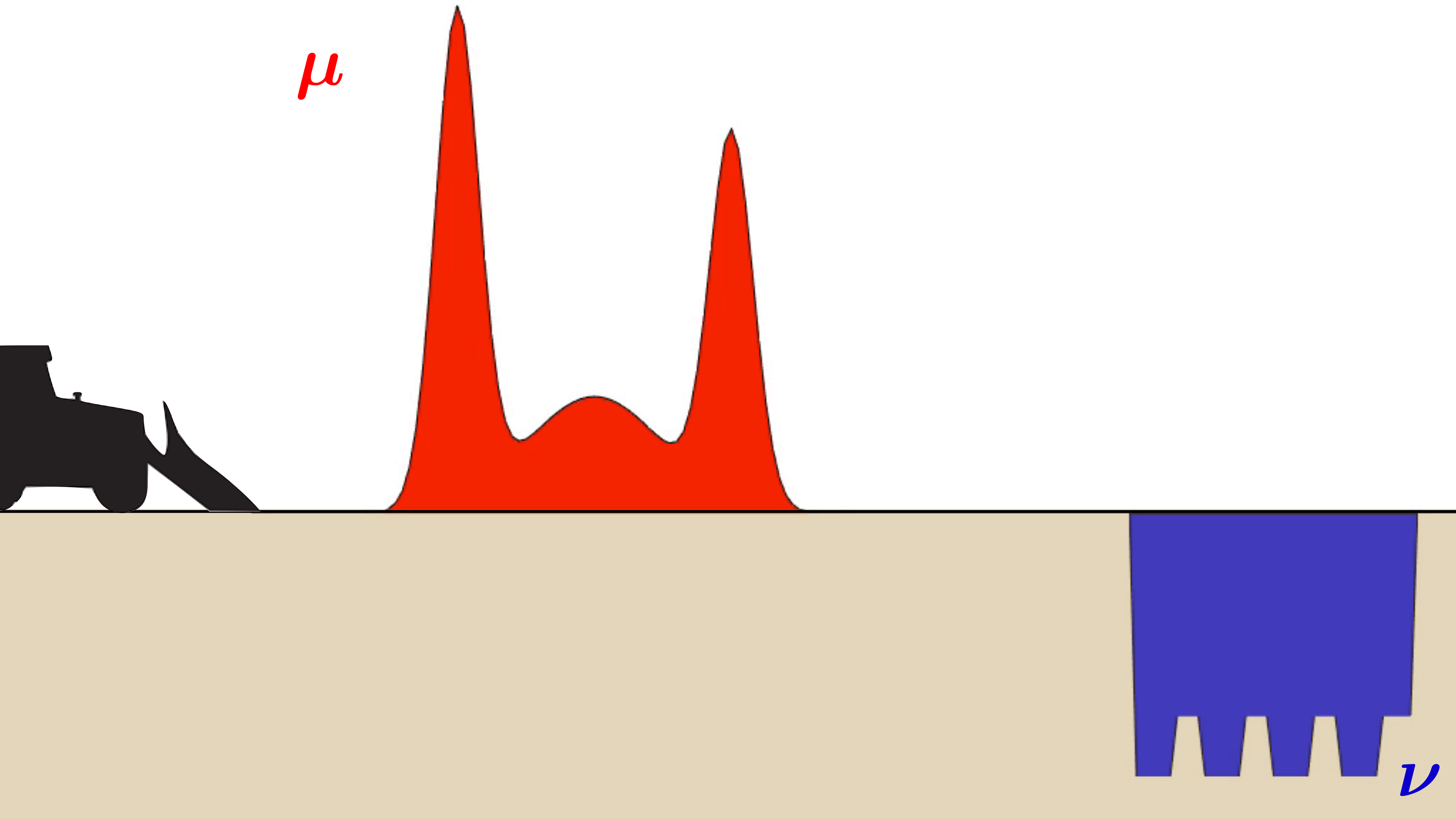
LORSQU'ON doit tranſporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranſporter, & le nom de *Remblai* à l'eſpace qu'elles doivent occuper après le tranſport.

Origins: Monge Problem



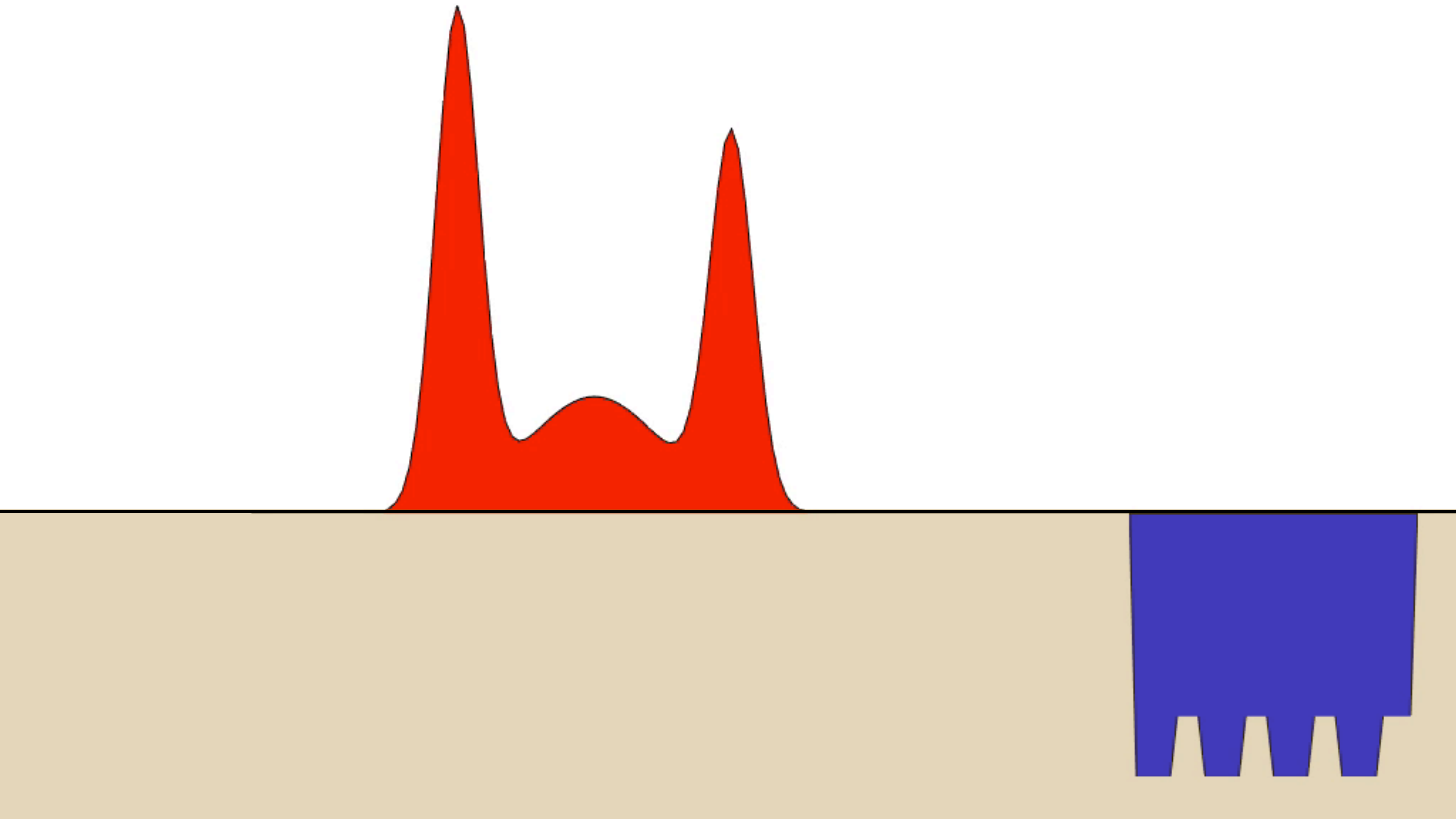
Origins: Monge Problem

In the 21st Century...



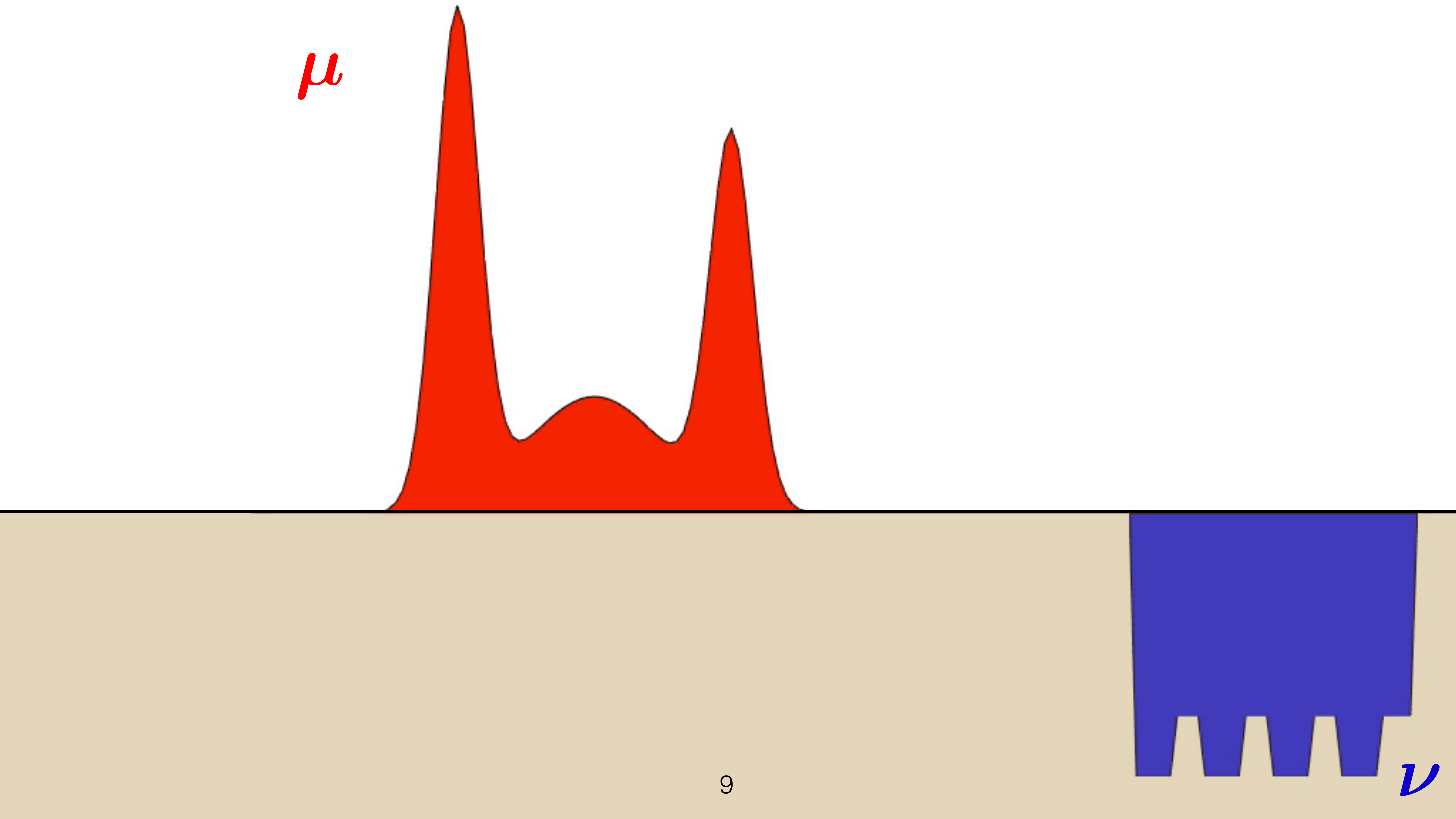
Origins: Monge Problem

In the 21st Century...



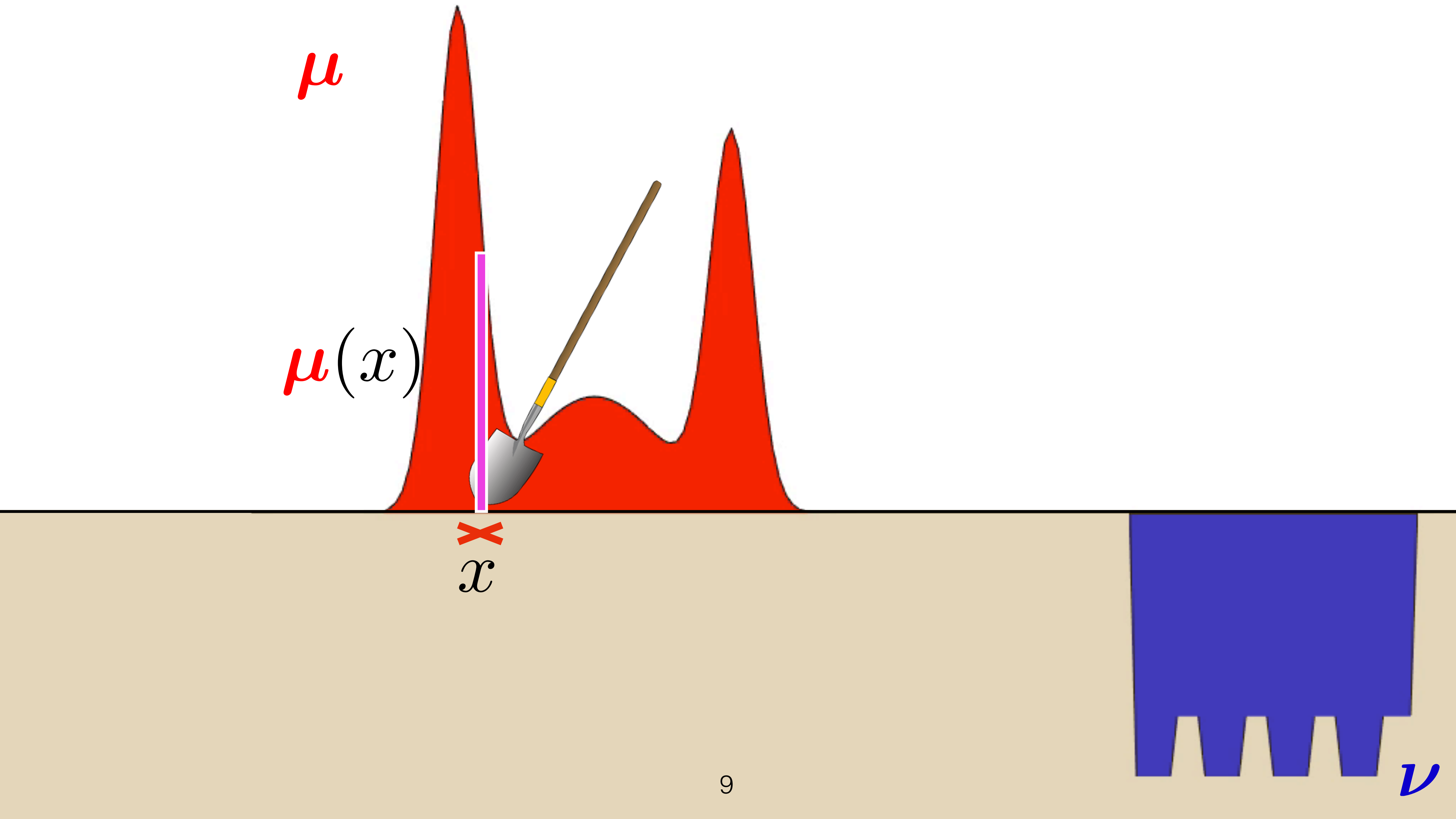
Origins: Monge's Problem

In 1781 however…



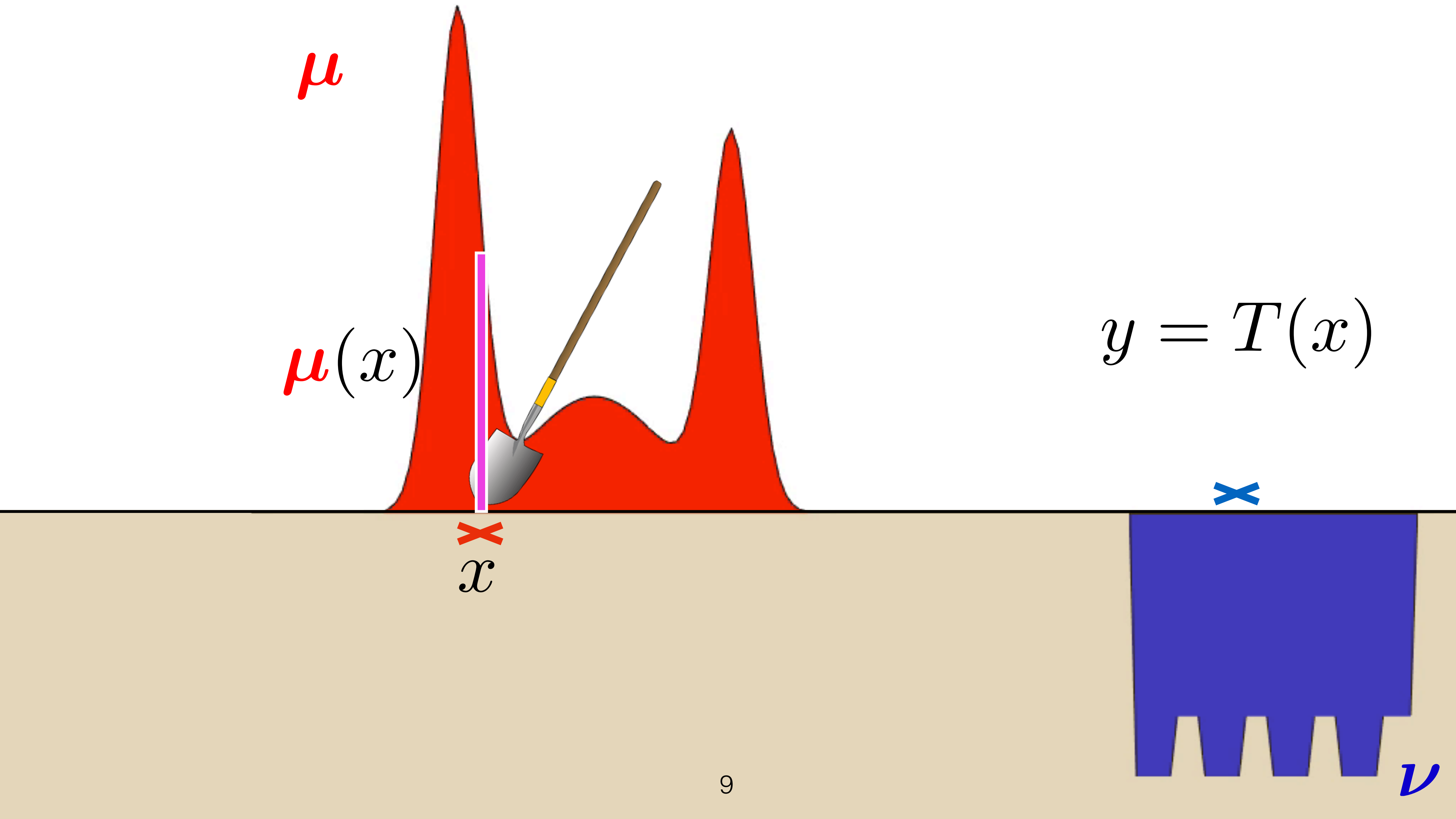
Origins: Monge's Problem

In 1781 however…



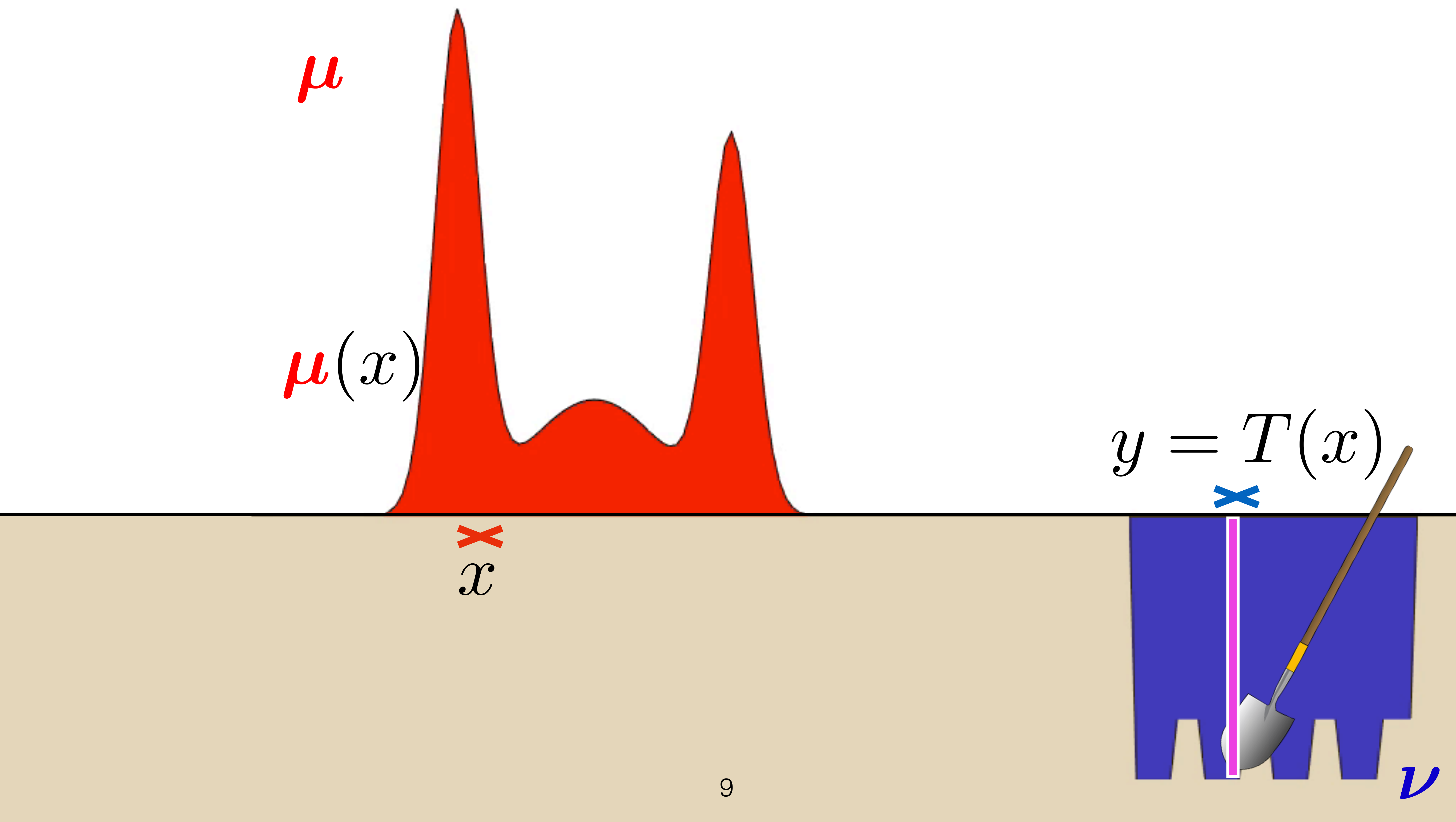
Origins: Monge's Problem

In 1781 however...



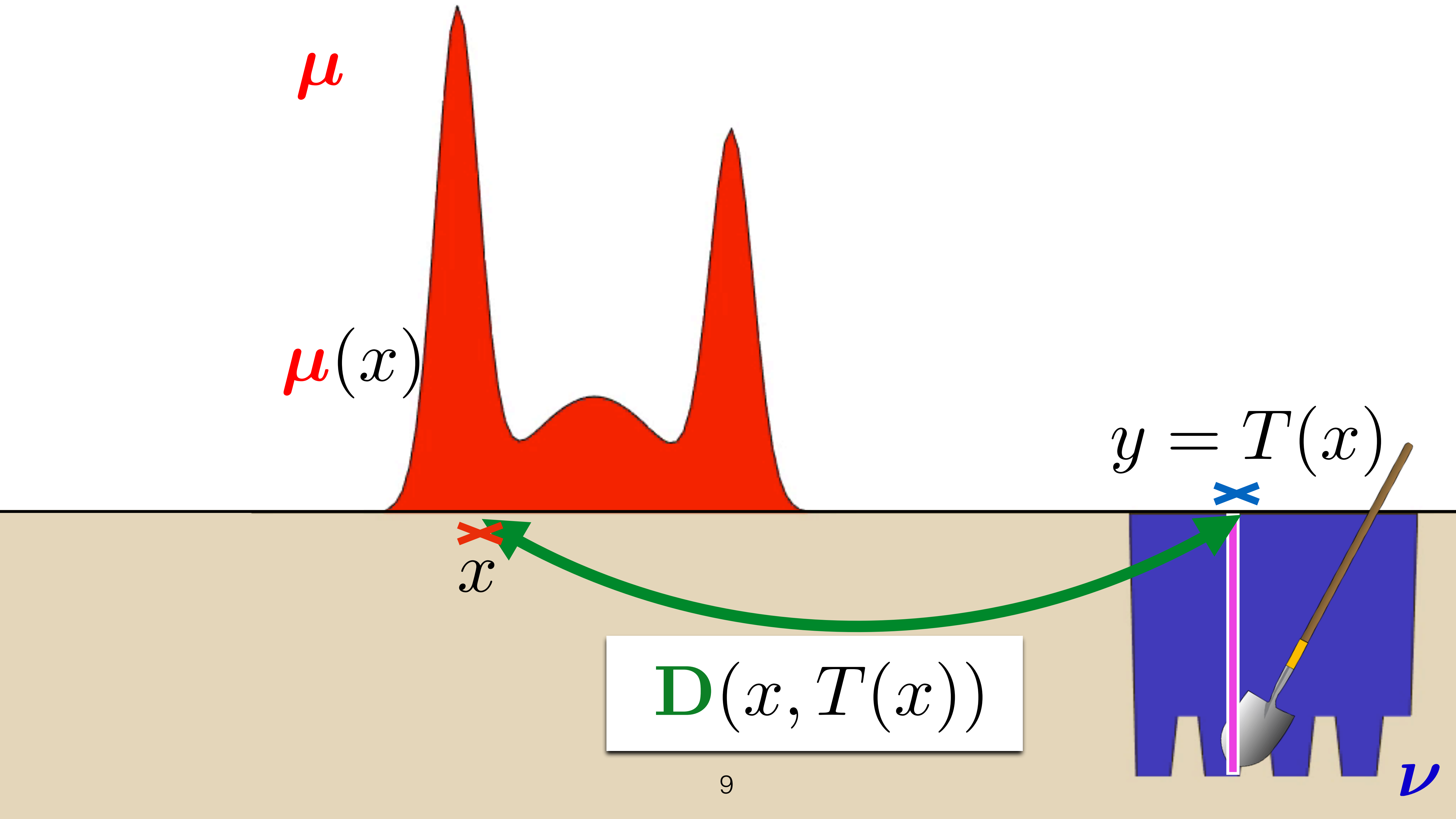
Origins: Monge's Problem

In 1781 however...



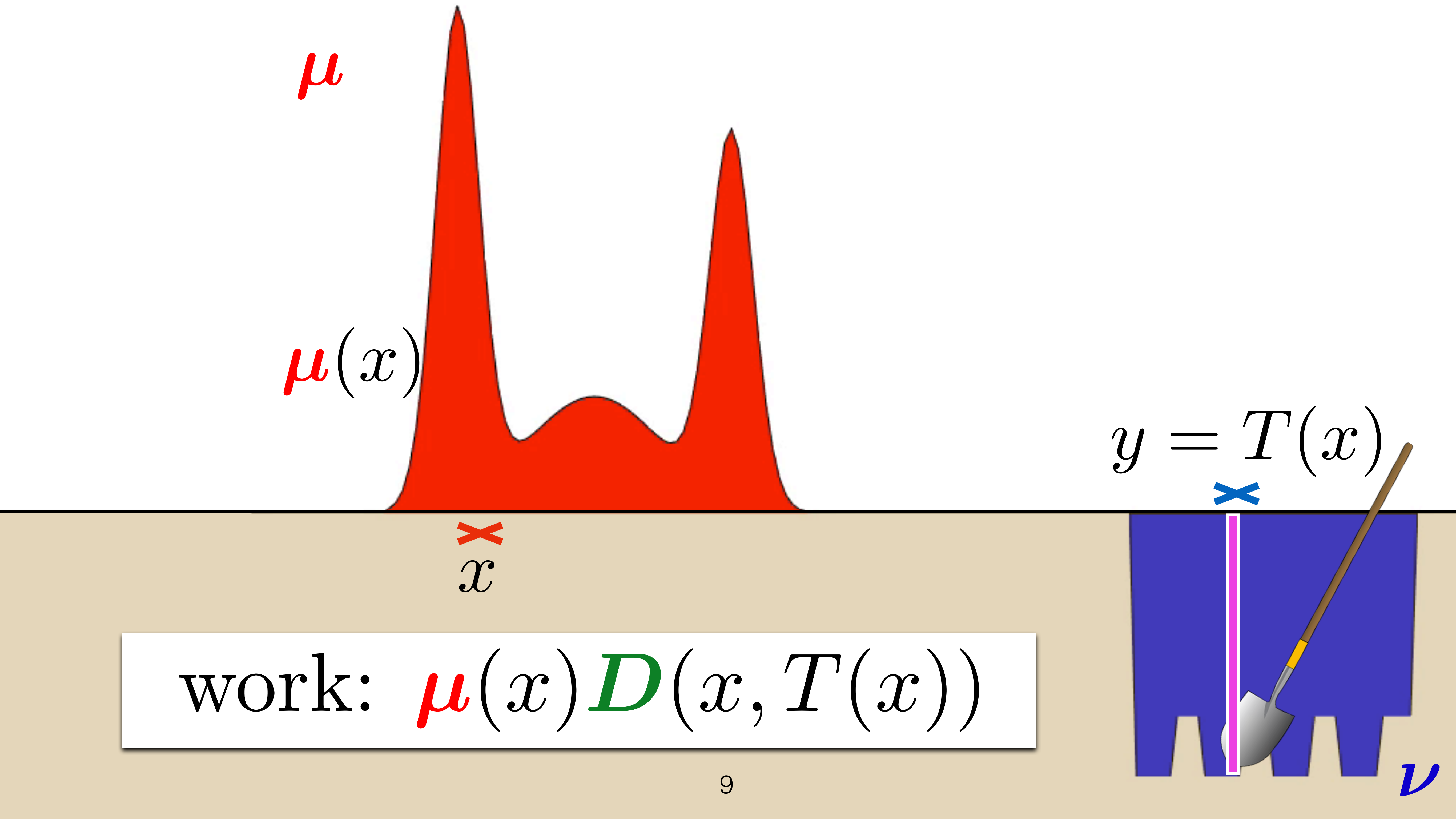
Origins: Monge's Problem

In 1781 however…



Origins: Monge's Problem

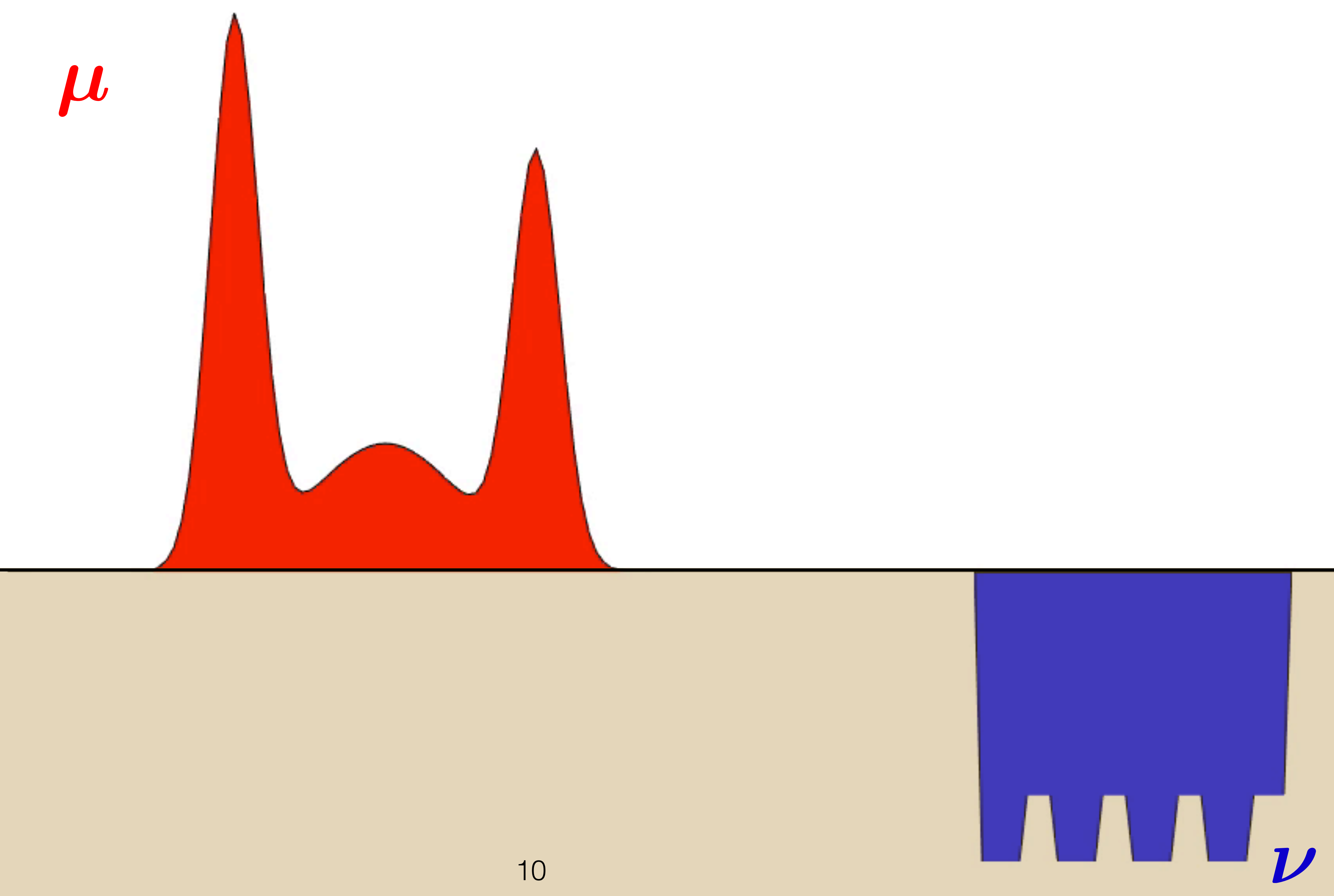
In 1781 however...



work: $\mu(x) D(x, T(x))$

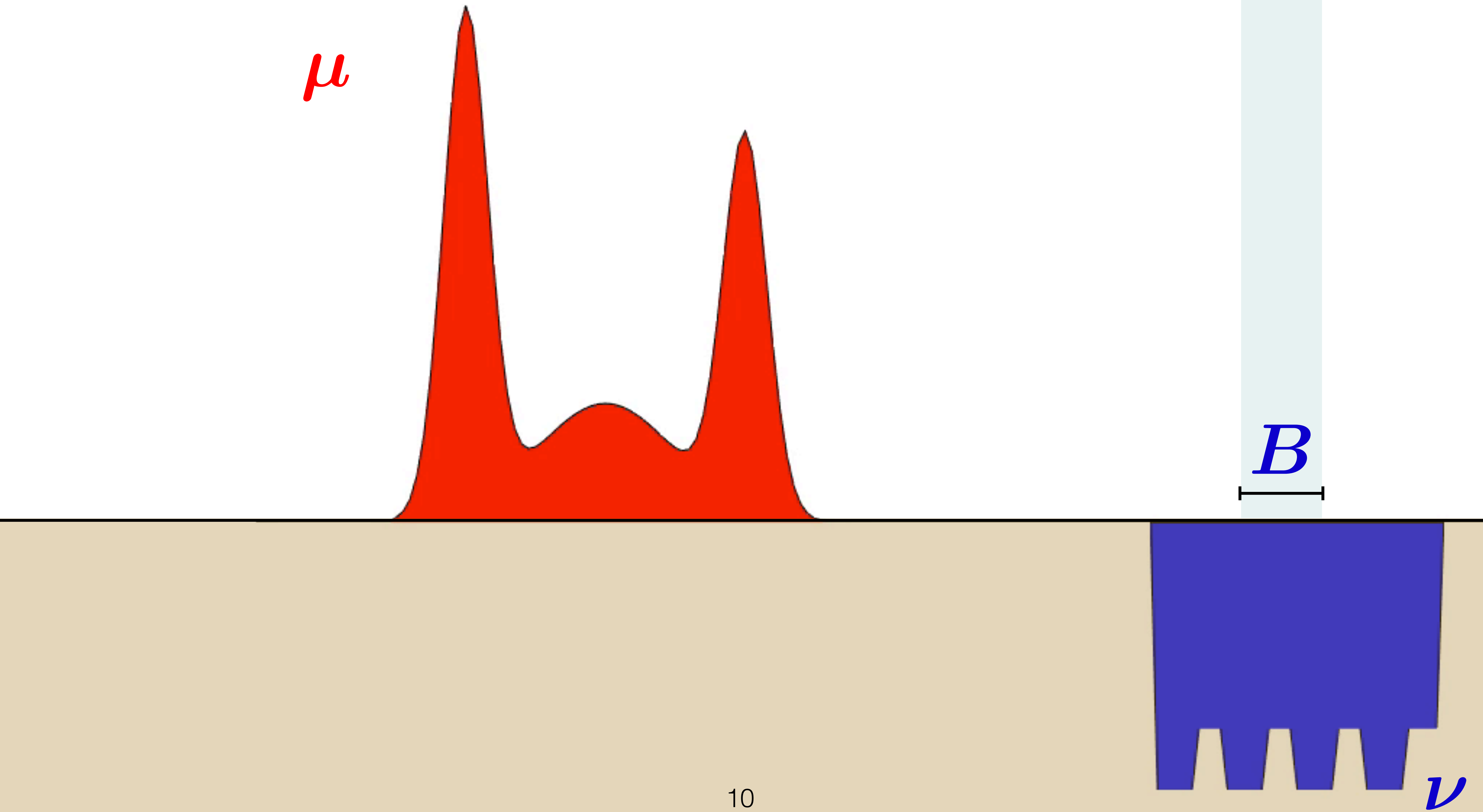
Origins: Monge's Problem

T must map red to blue.



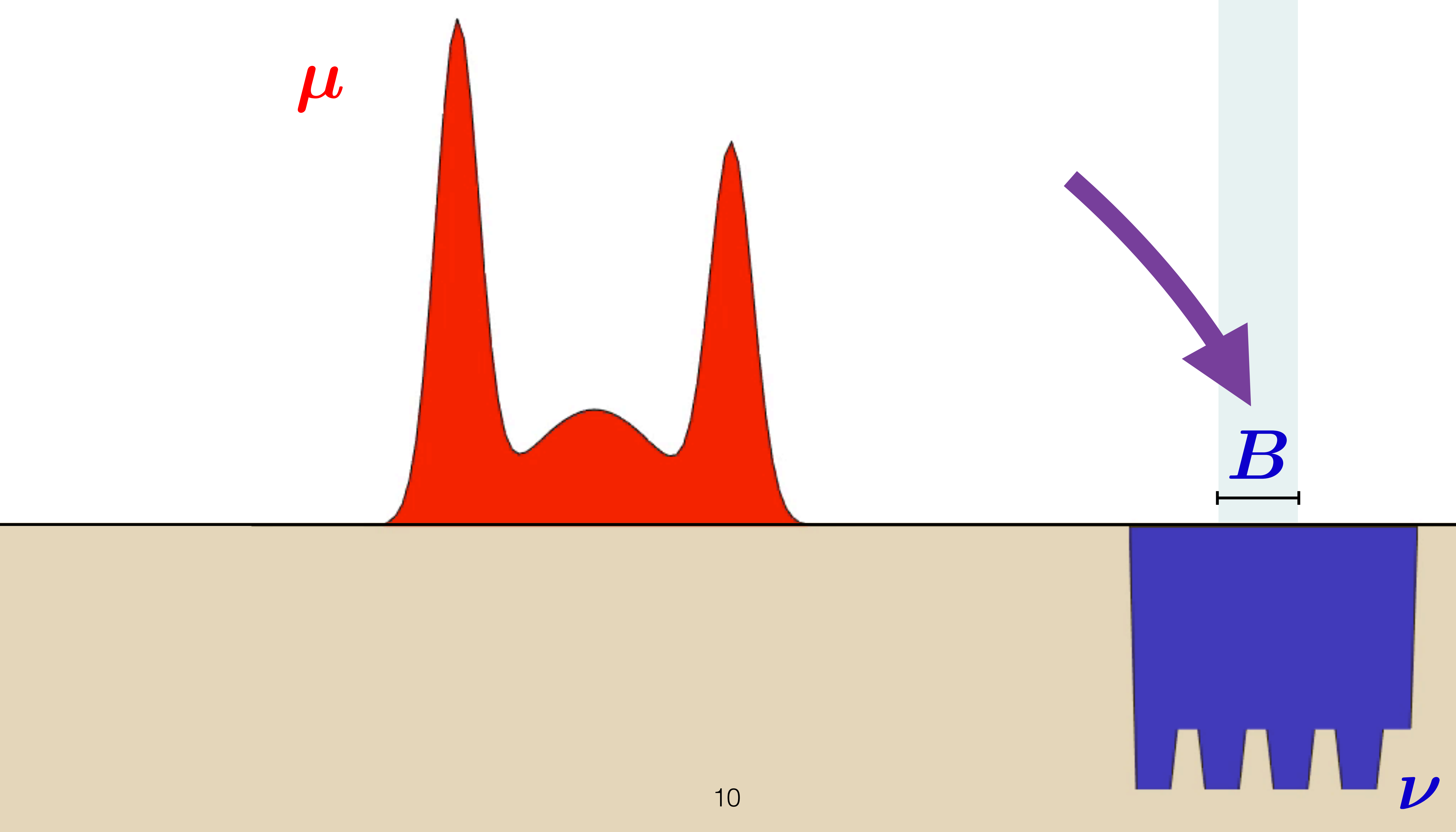
Origins: Monge's Problem

T must map red to blue.



Origins: Monge's Problem

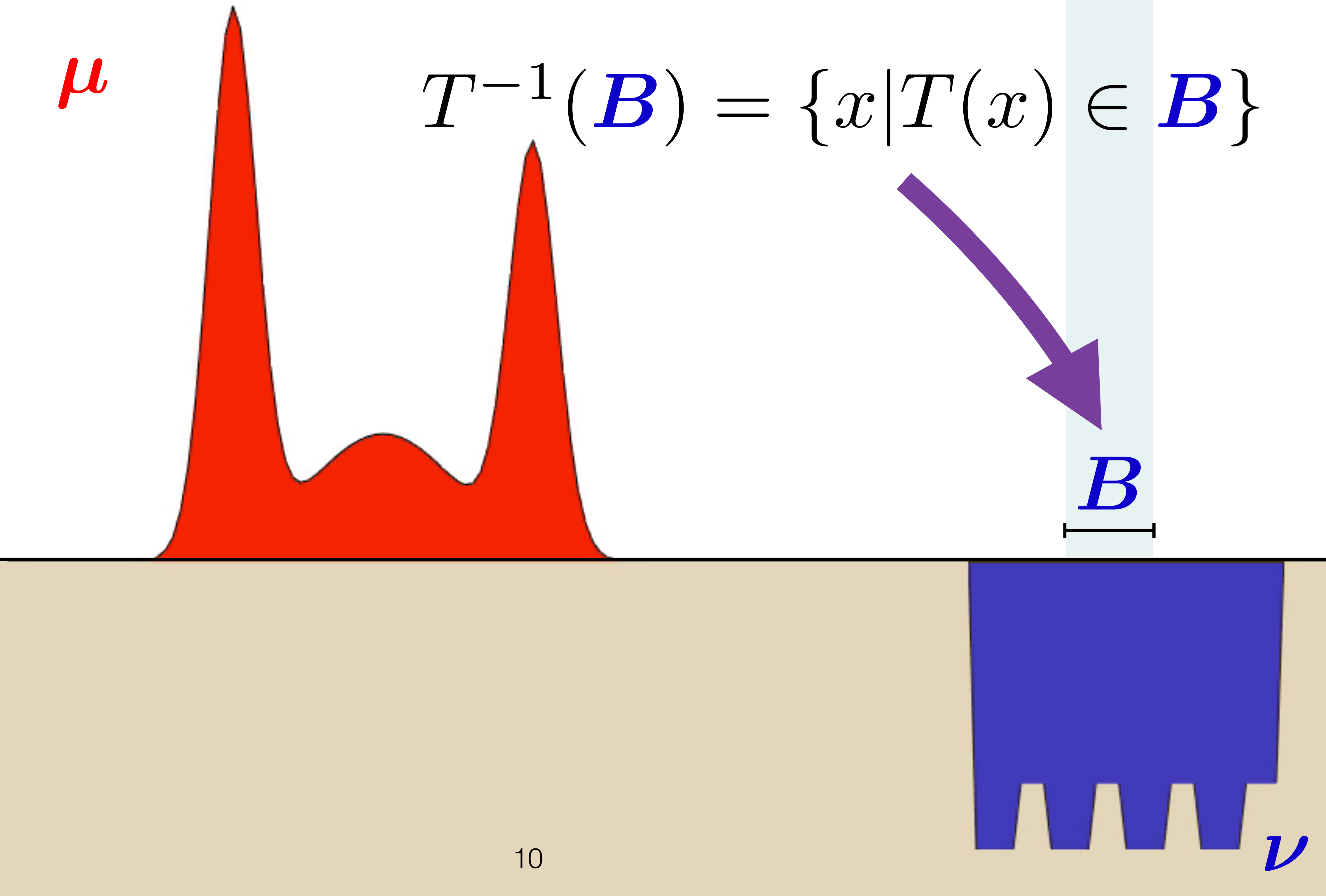
T must map red to blue.



Origins: Monge's Problem

T must map red to blue.

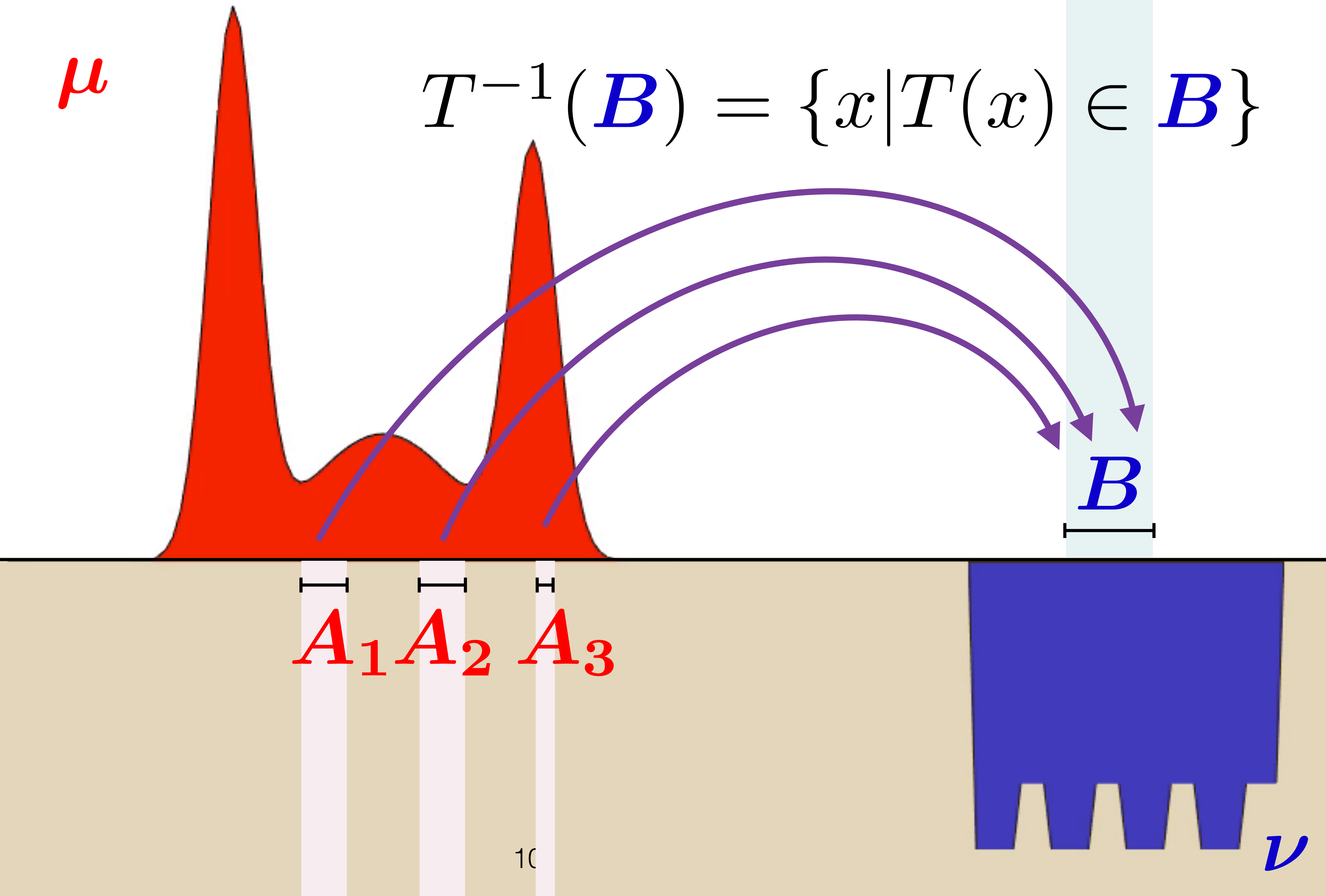
$$T^{-1}(\boldsymbol{B}) = \{x \mid T(x) \in \boldsymbol{B}\}$$



Origins: Monge's Problem

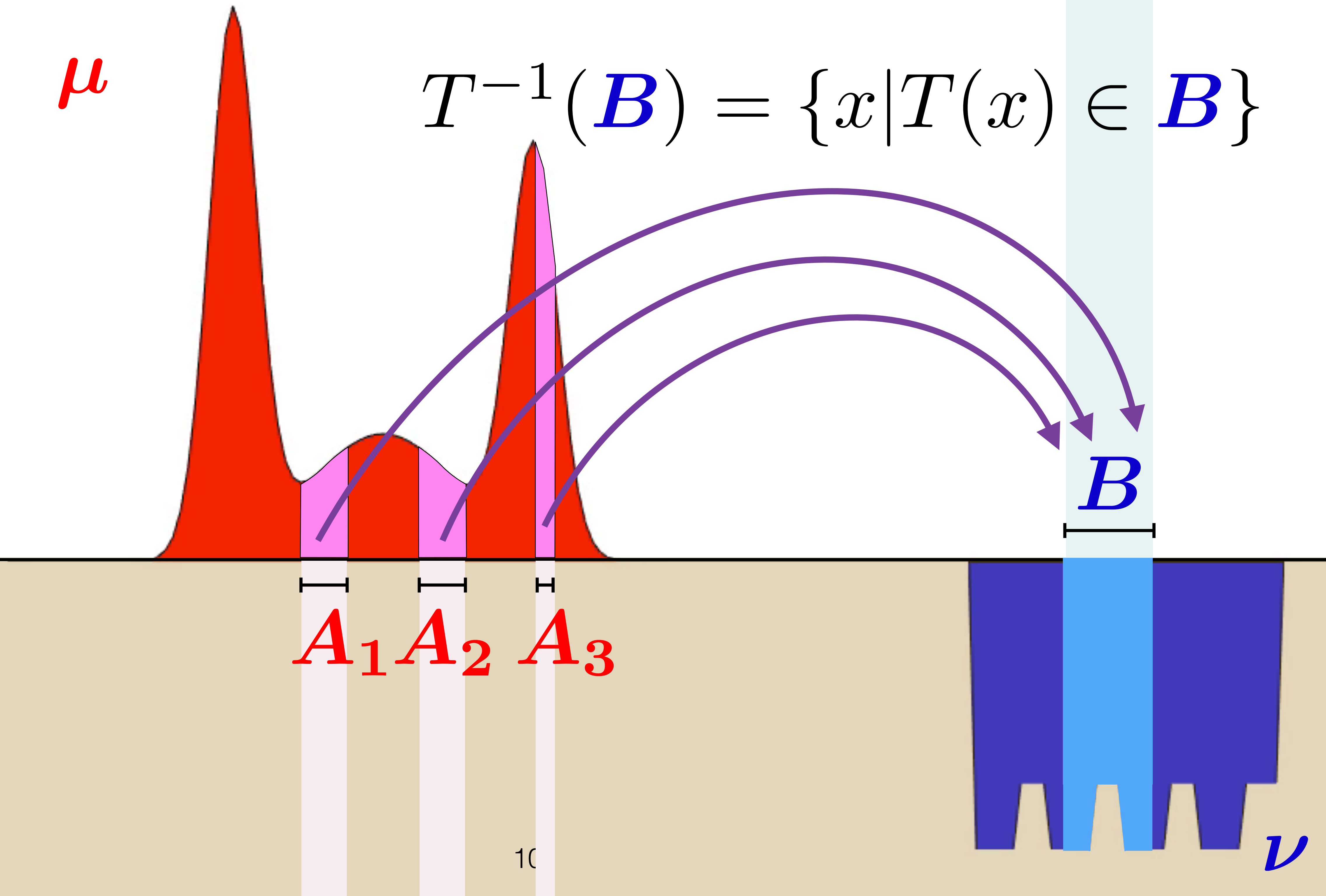
T must map red to blue.

$$T^{-1}(\boldsymbol{B}) = \{x | T(x) \in \boldsymbol{B}\}$$



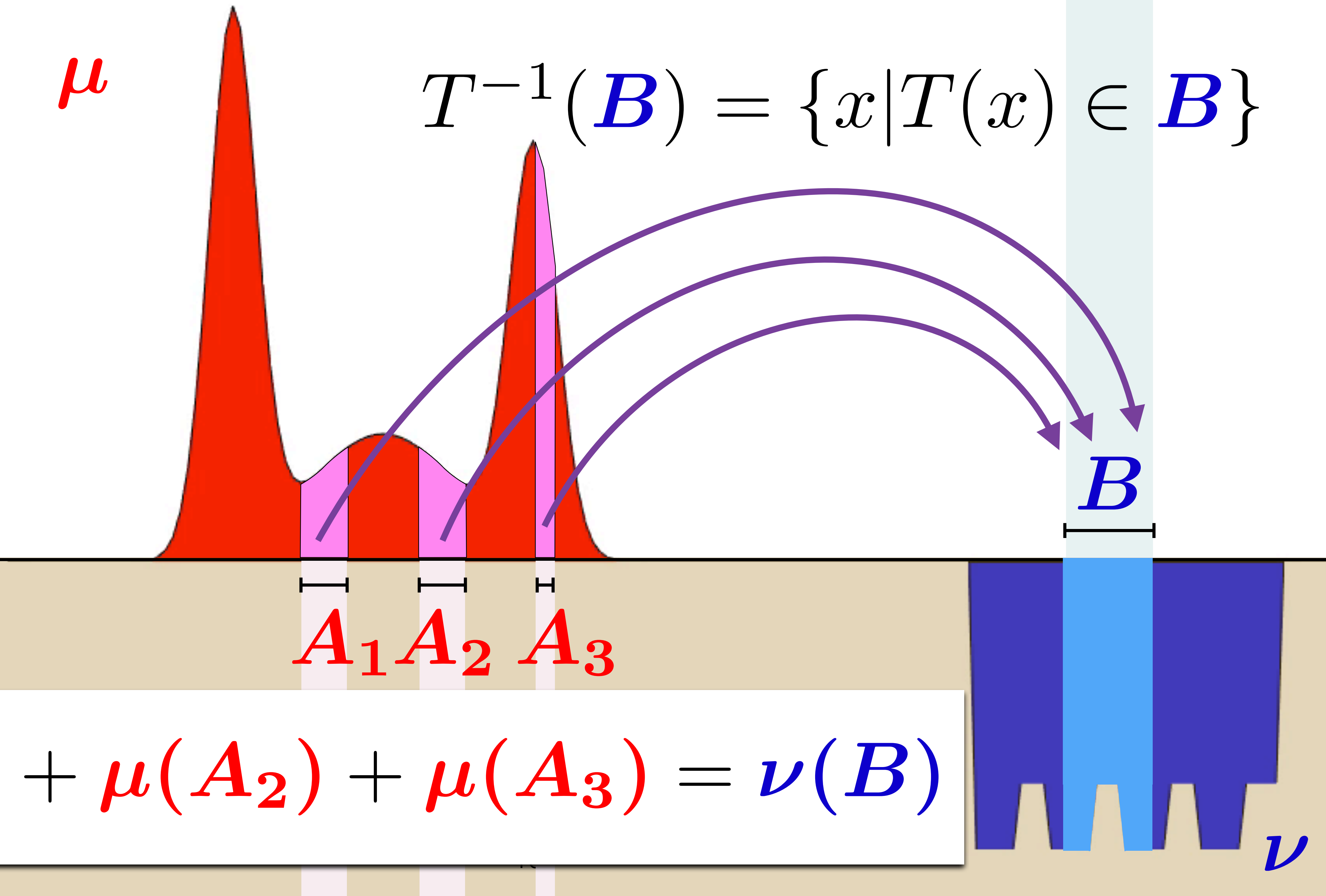
Origins: Monge's Problem

T must map red to blue.



Origins: Monge's Problem

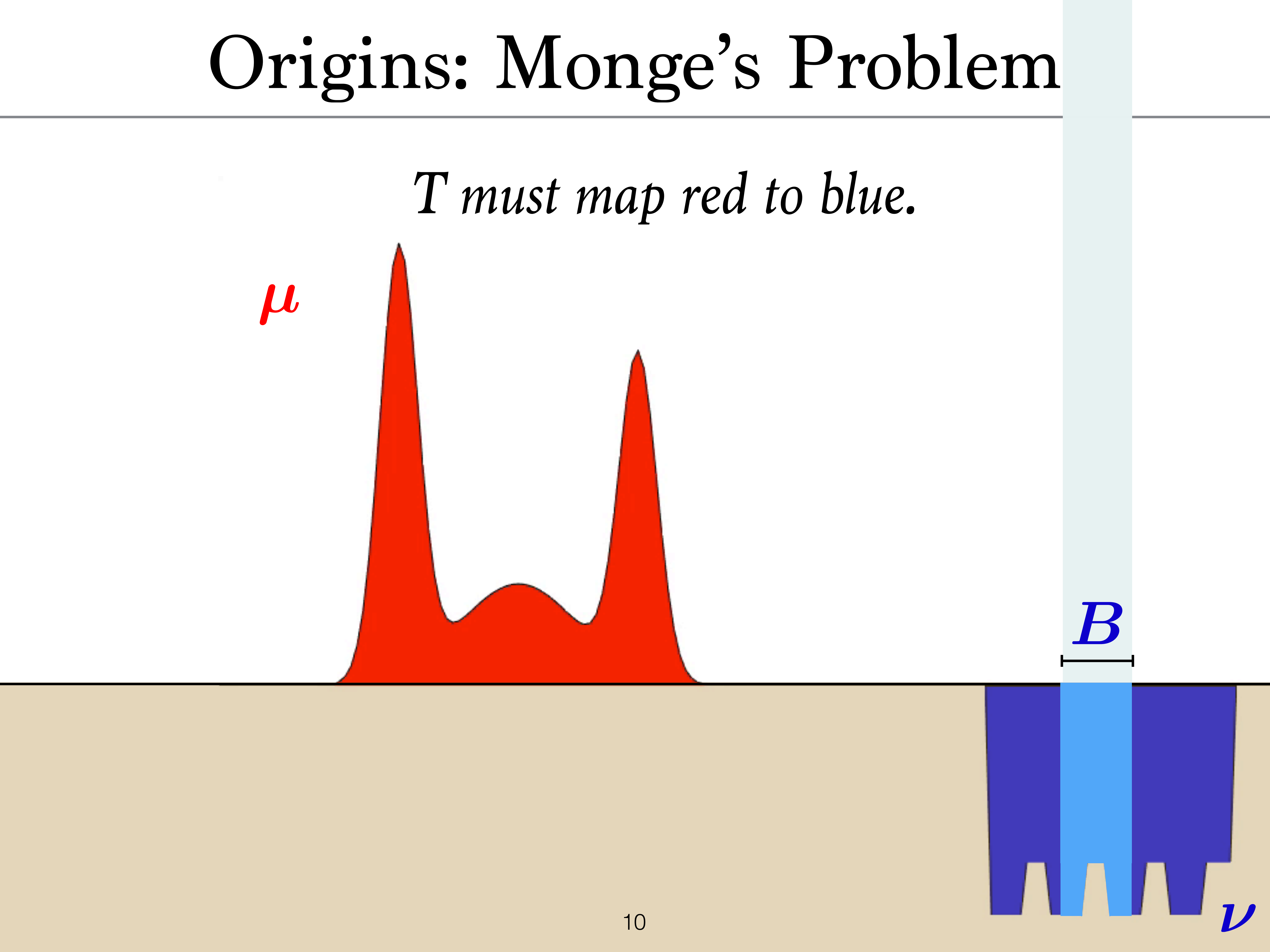
T must map red to blue.



$$\mu(A_1) + \mu(A_2) + \mu(A_3) = \nu(B)$$

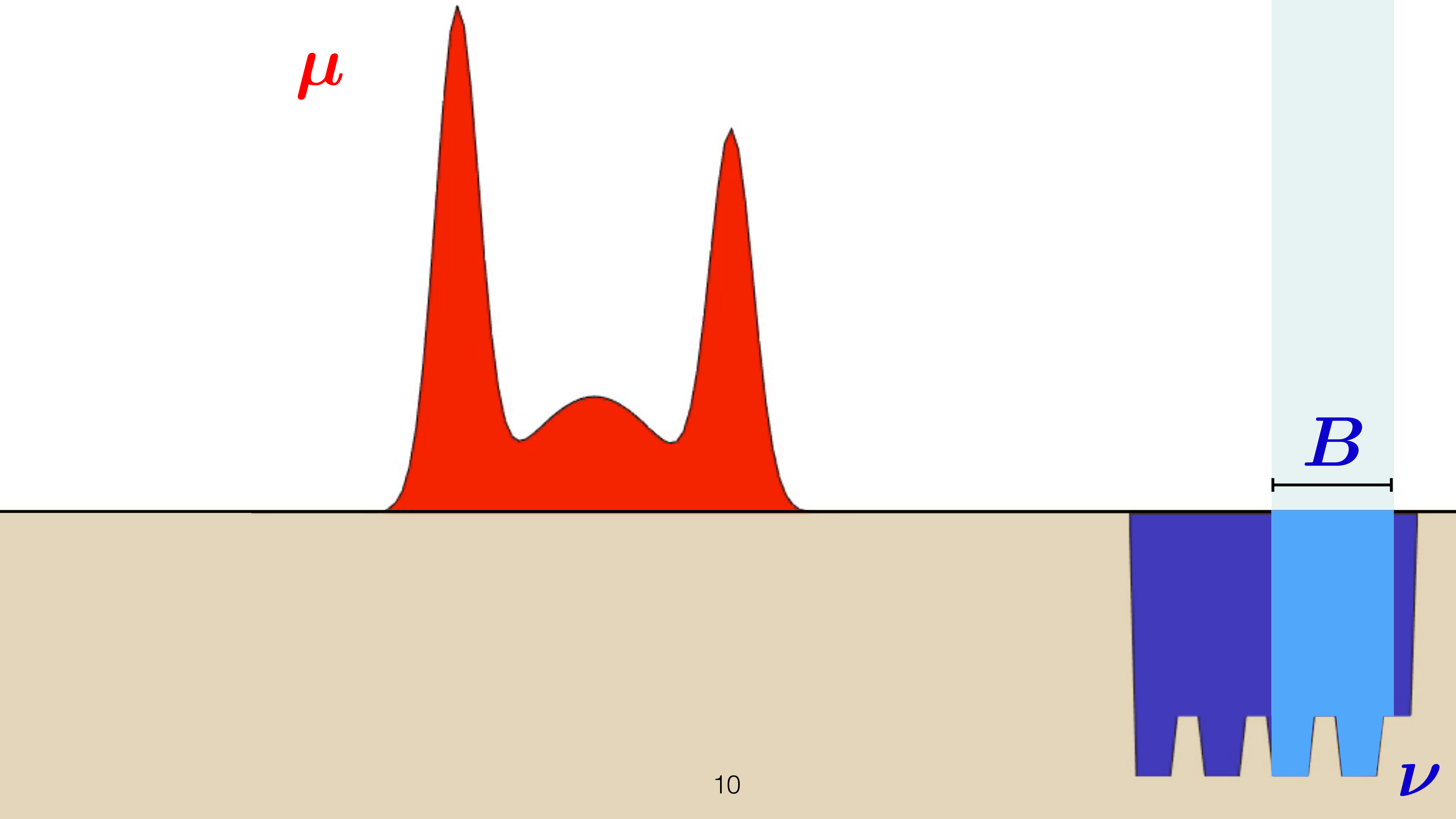
Origins: Monge's Problem

T must map red to blue.



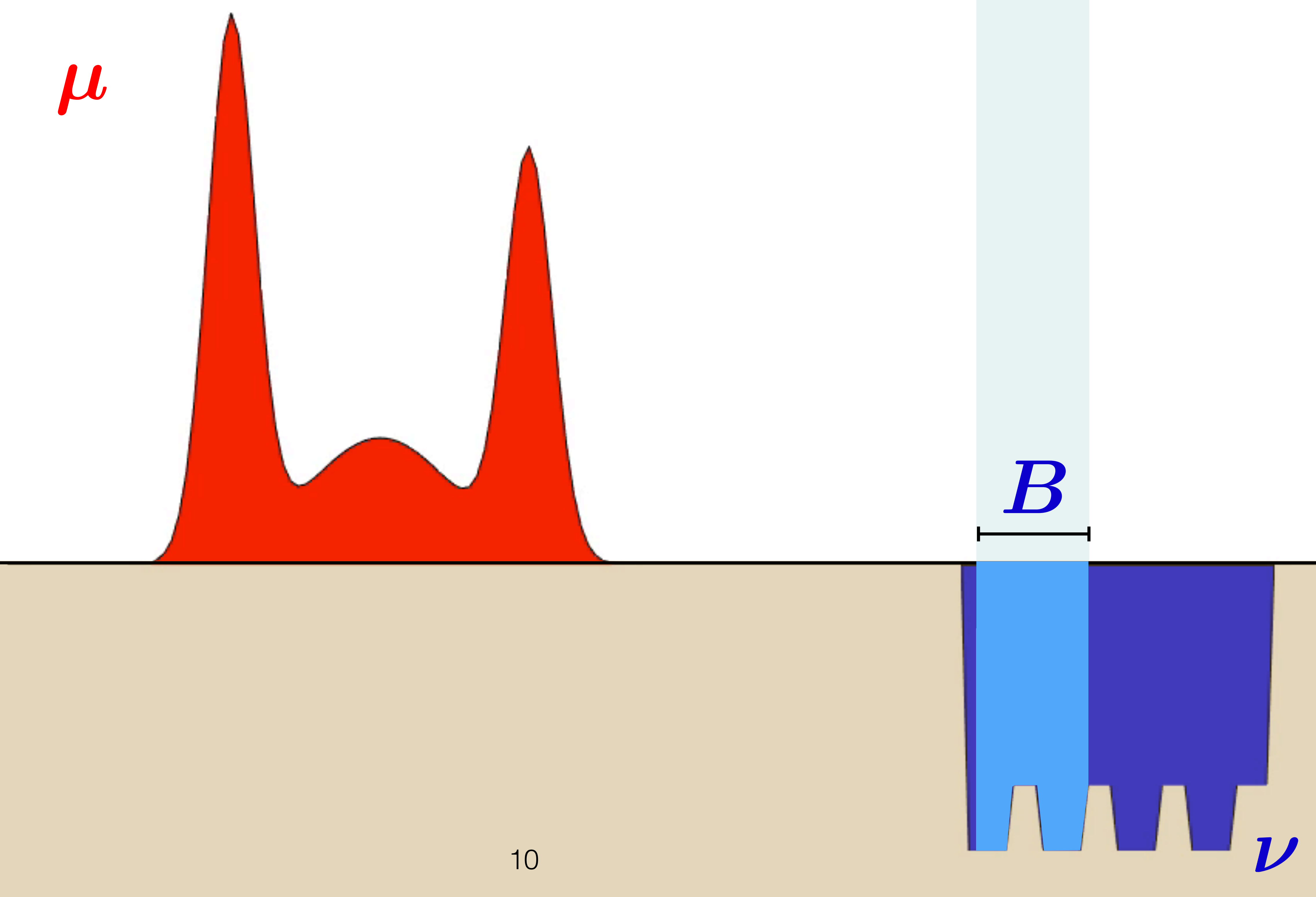
Origins: Monge's Problem

T must map red to blue.



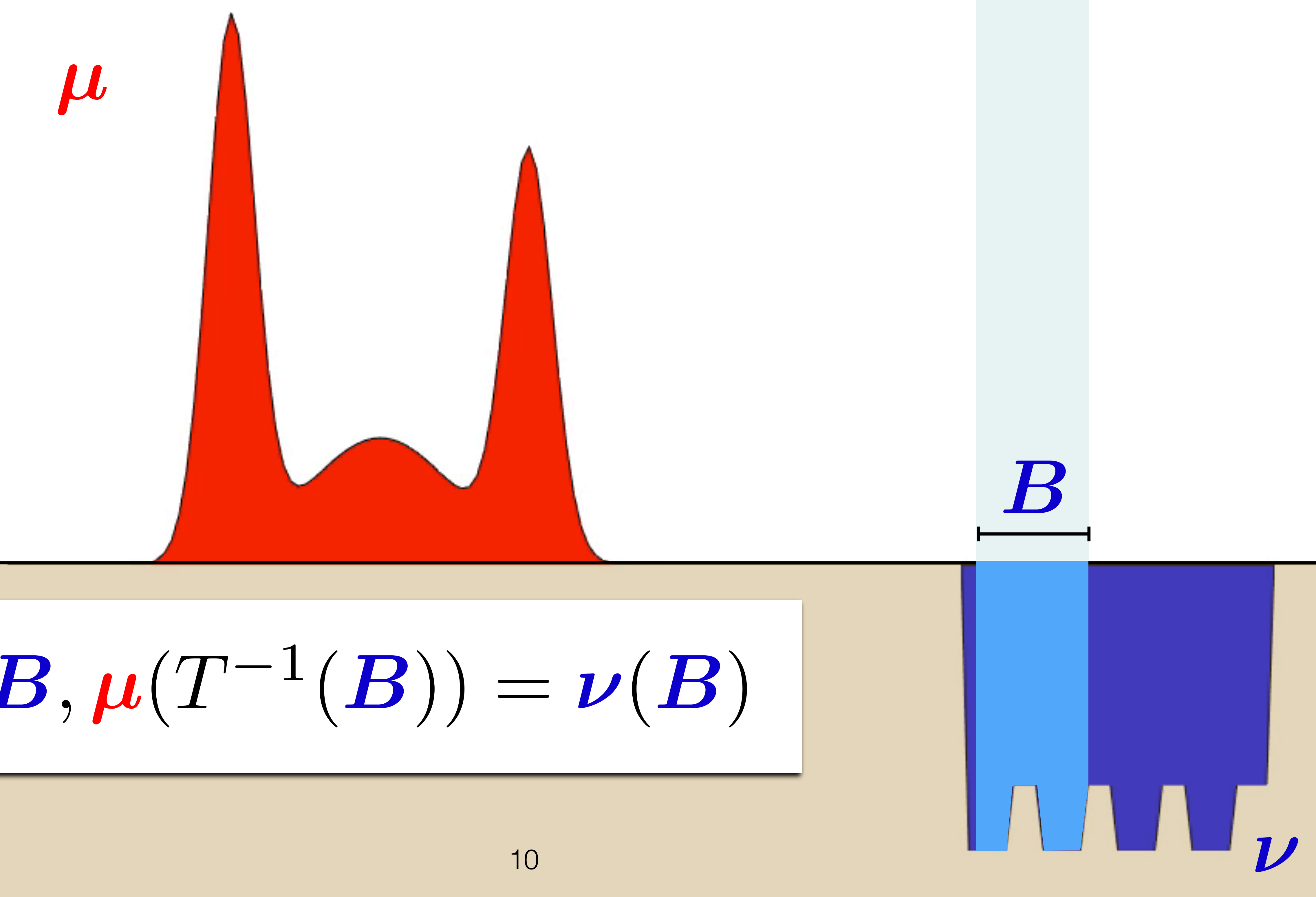
Origins: Monge's Proble

T must map red to blue.



Origins: Monge's Proble

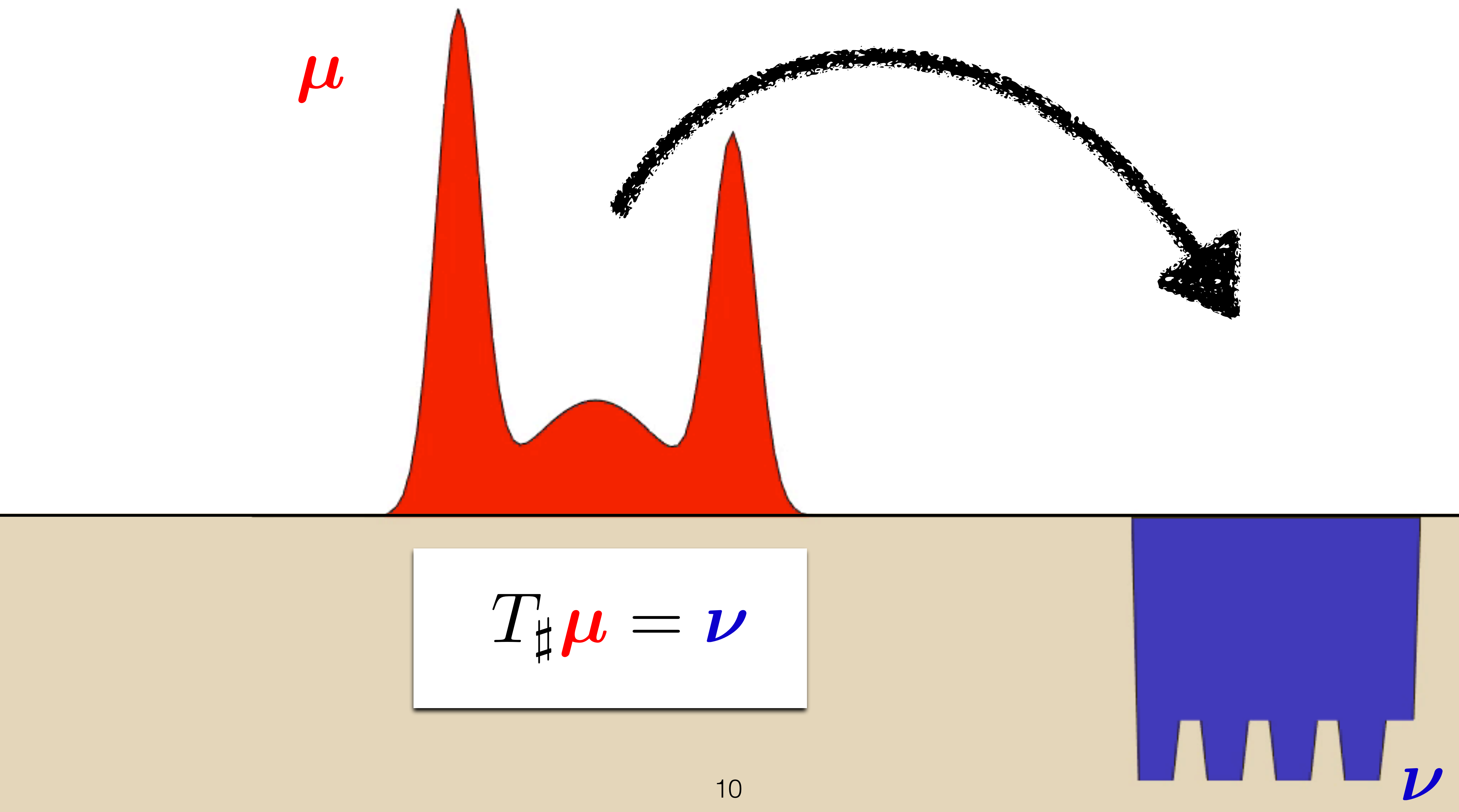
T must map red to blue.



$$\forall B, \mu(T^{-1}(B)) = \nu(B)$$

Origins: Monge's Problem

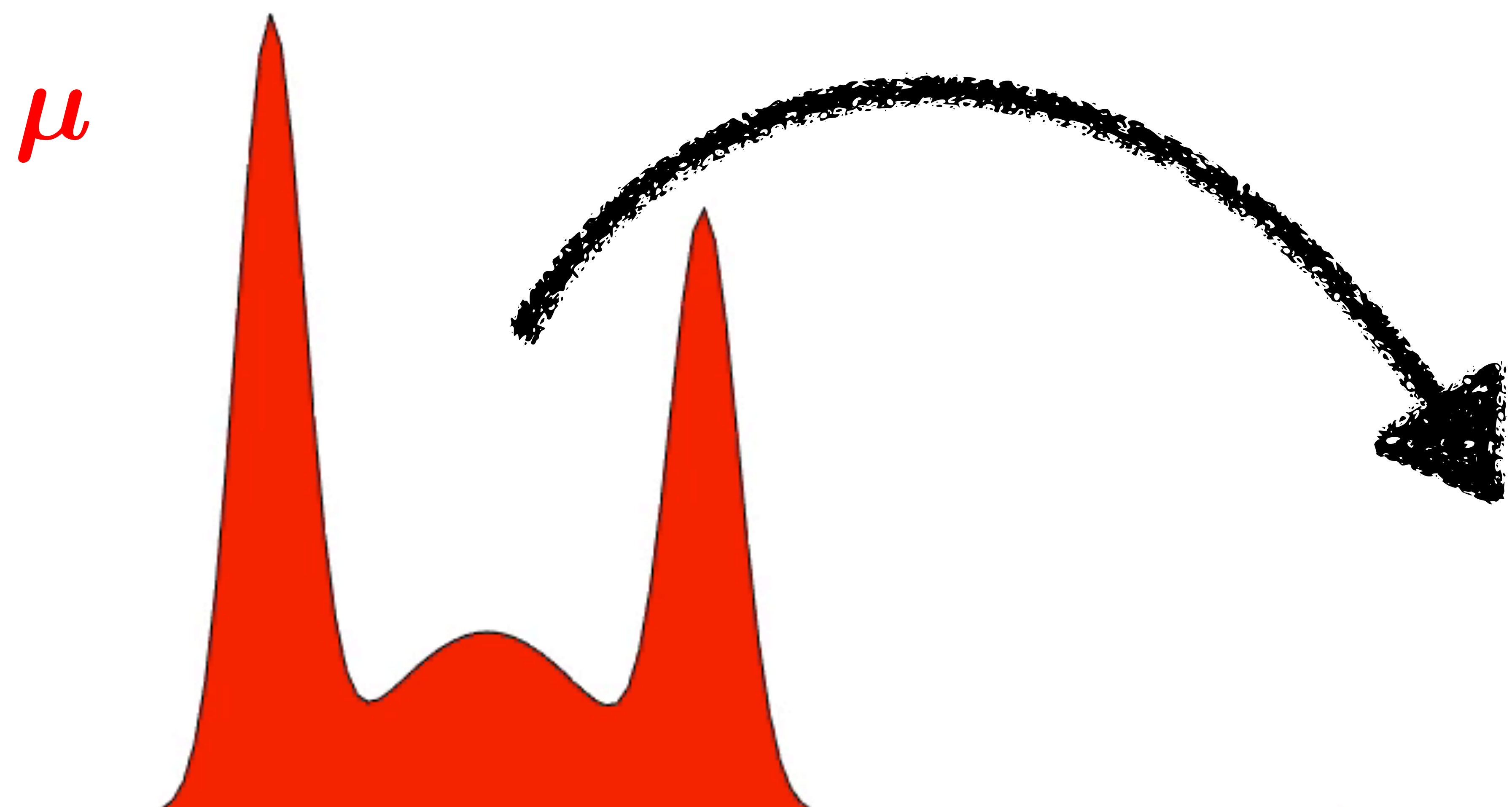
T must **push-forward** *the red measure towards the blue*



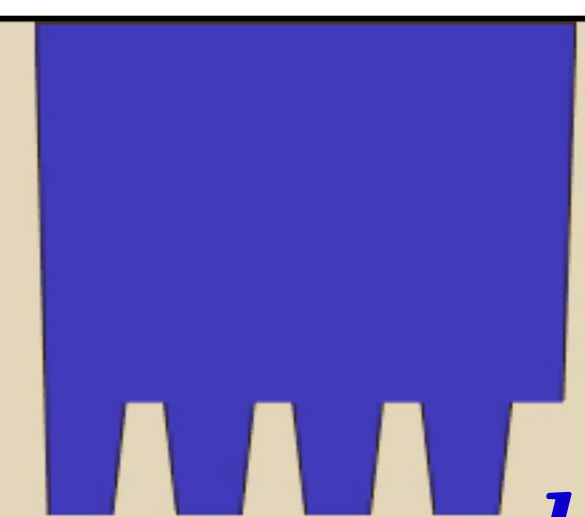
$$T_\sharp \mu = \nu$$

Origins: Monge's Problem

T must **push-forward** *the red measure towards the blue*



What T s.t. $T_\sharp \mu = \nu$
minimizes $\int D(x, T(x))\mu(dx)$?

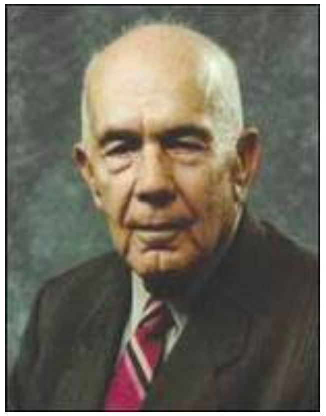


Kantorovich Problem

Kantorovich

Tolstoi

1930



Hitchcock

1939

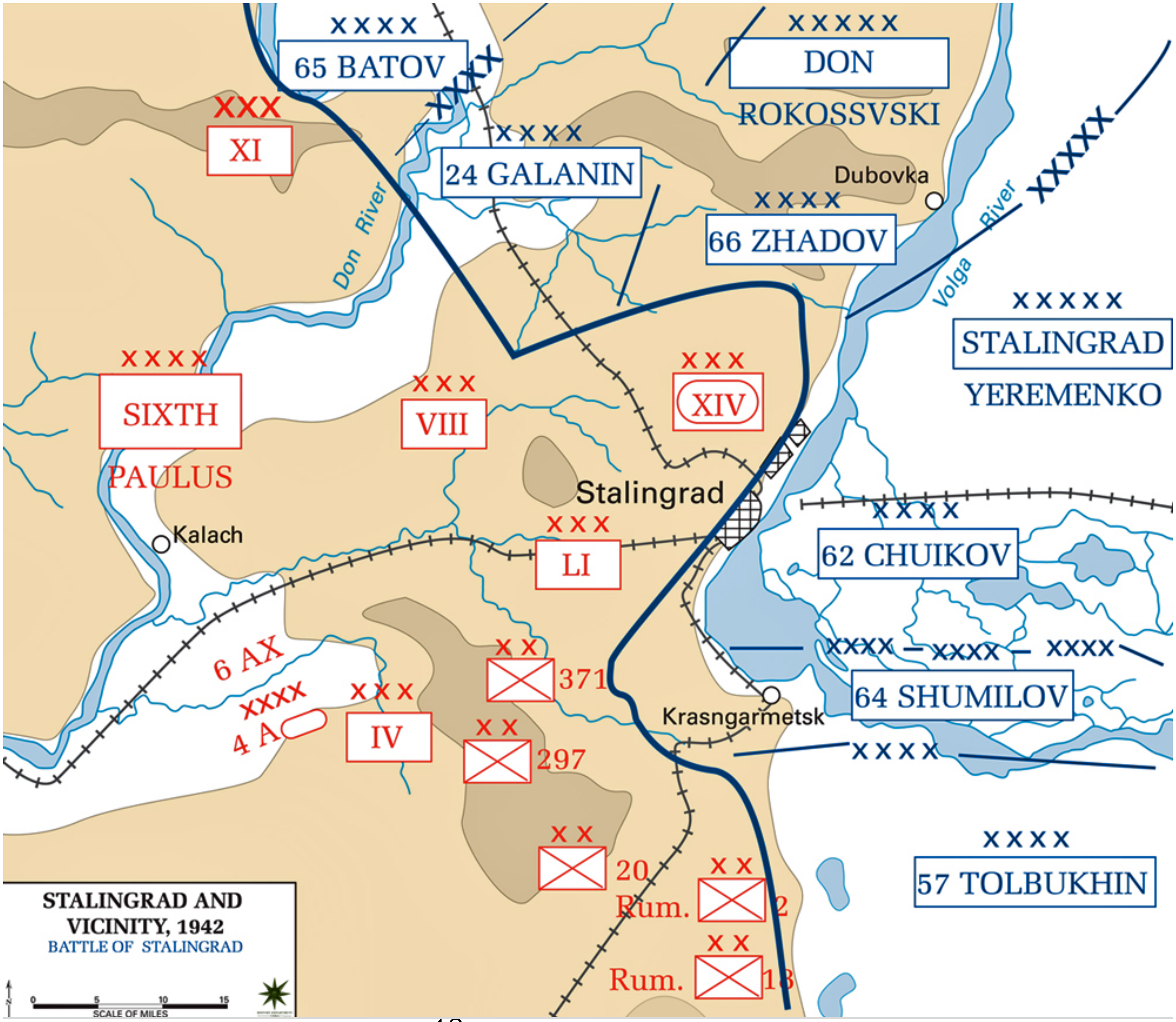
THE DISTRIBUTION OF A PRODUCT FROM SEVERAL SOURCES TO NUMEROUS LOCALITIES

BY FRANK L. HITCHCOCK

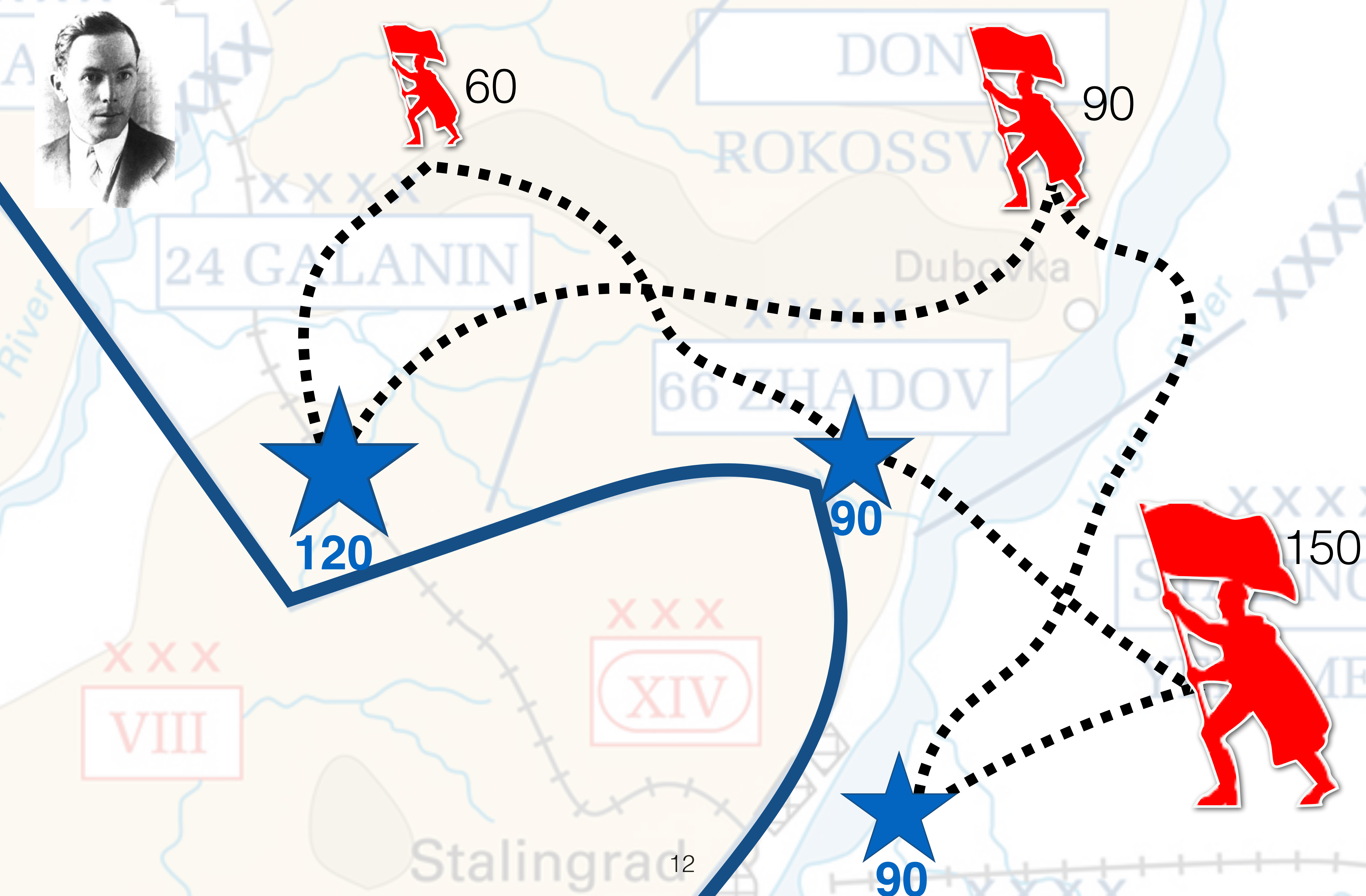
1. Statement of the problem. When several factories supply a product to a number of cities we desire the least costly manner of distribution. Due to freight rates and other matters the cost of a ton of product to a particular city will vary according to which factory supplies it, and will also vary from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

1941

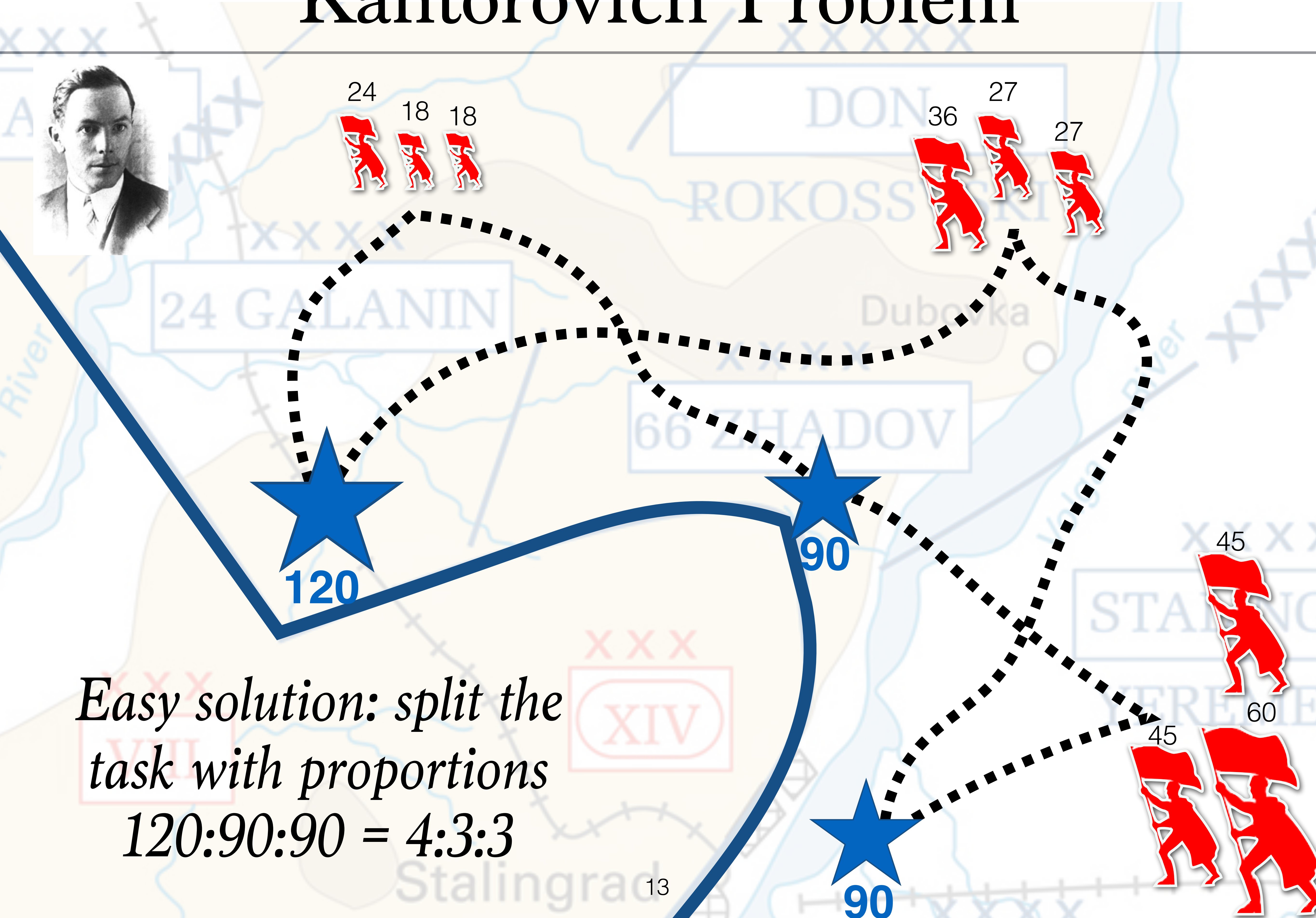
Kantorovich Problem



Kantorovich Problem

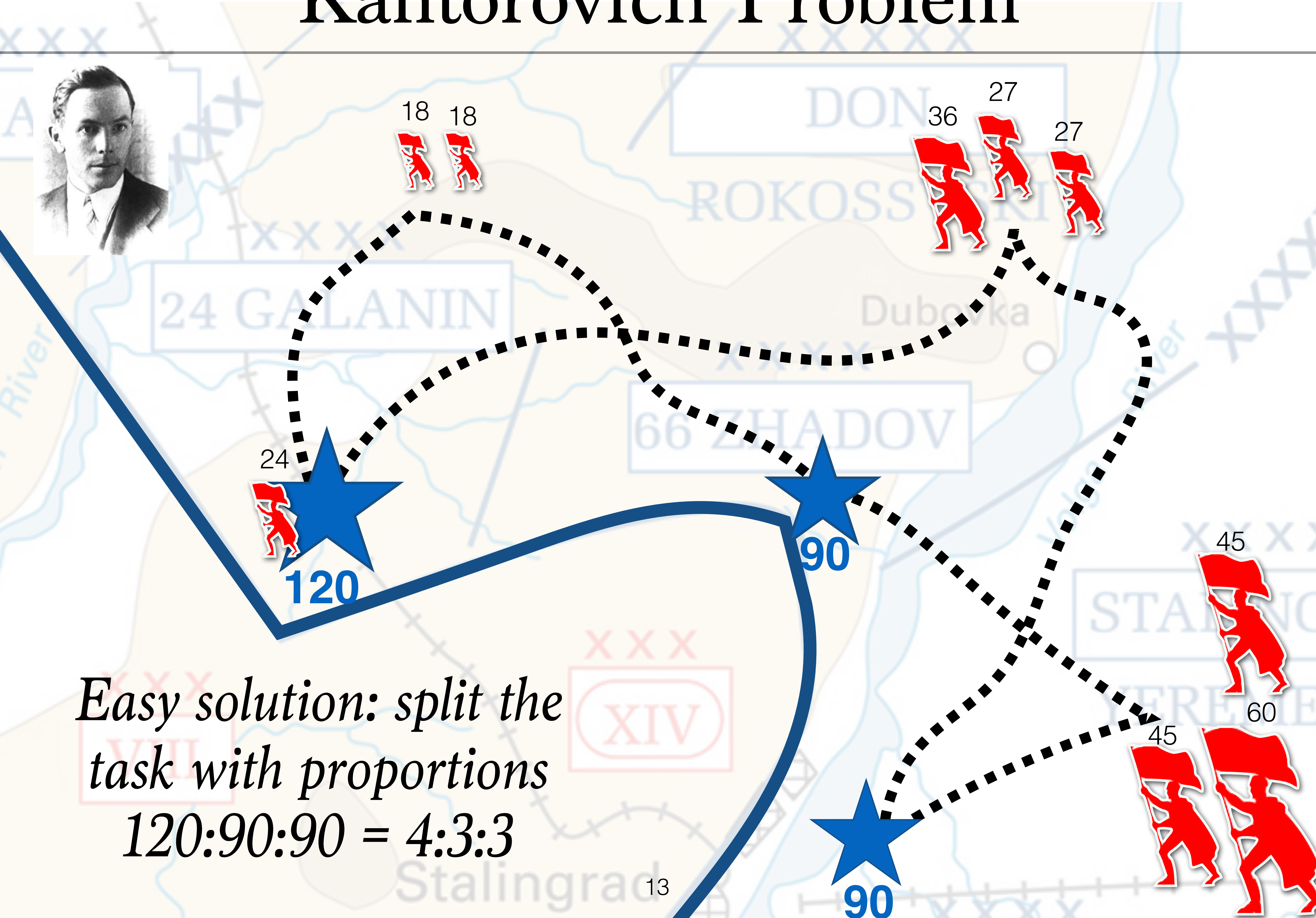


Kantorovich Problem



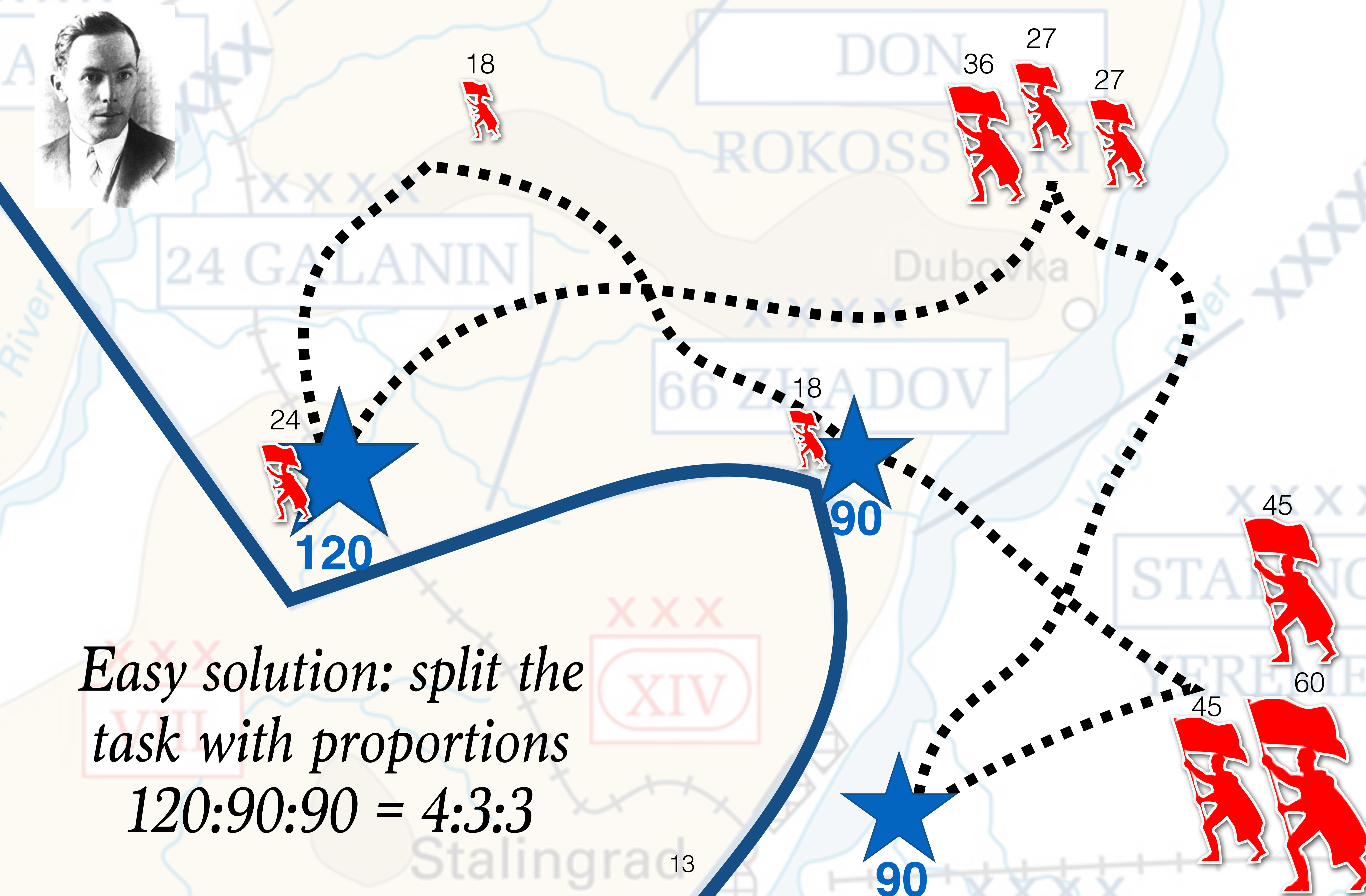
Easy solution: split the task with proportions 120:90:90 = 4:3:3

Kantorovich Problem



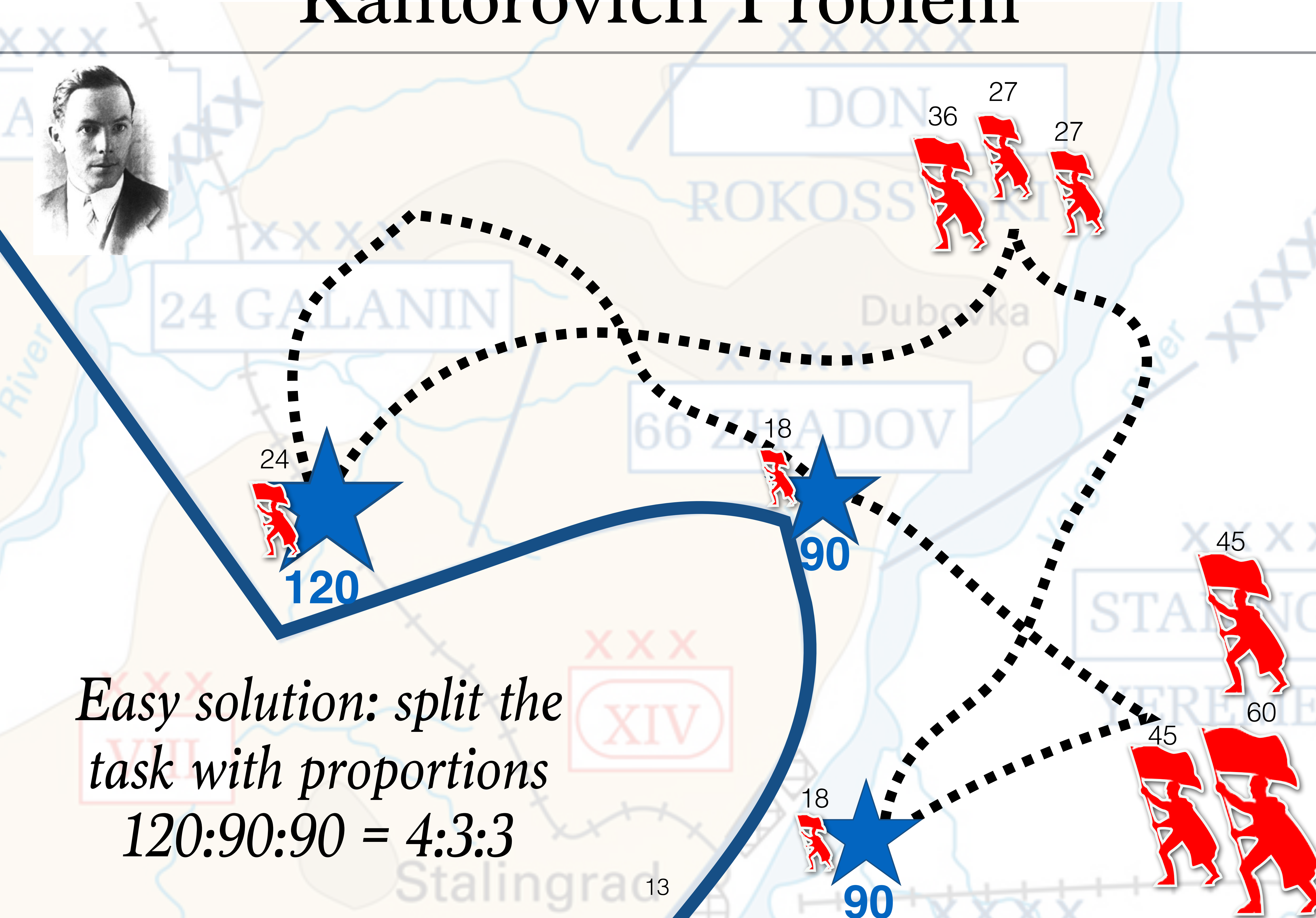
Easy solution: split the task with proportions 120:90:90 = 4:3:3

Kantorovich Problem



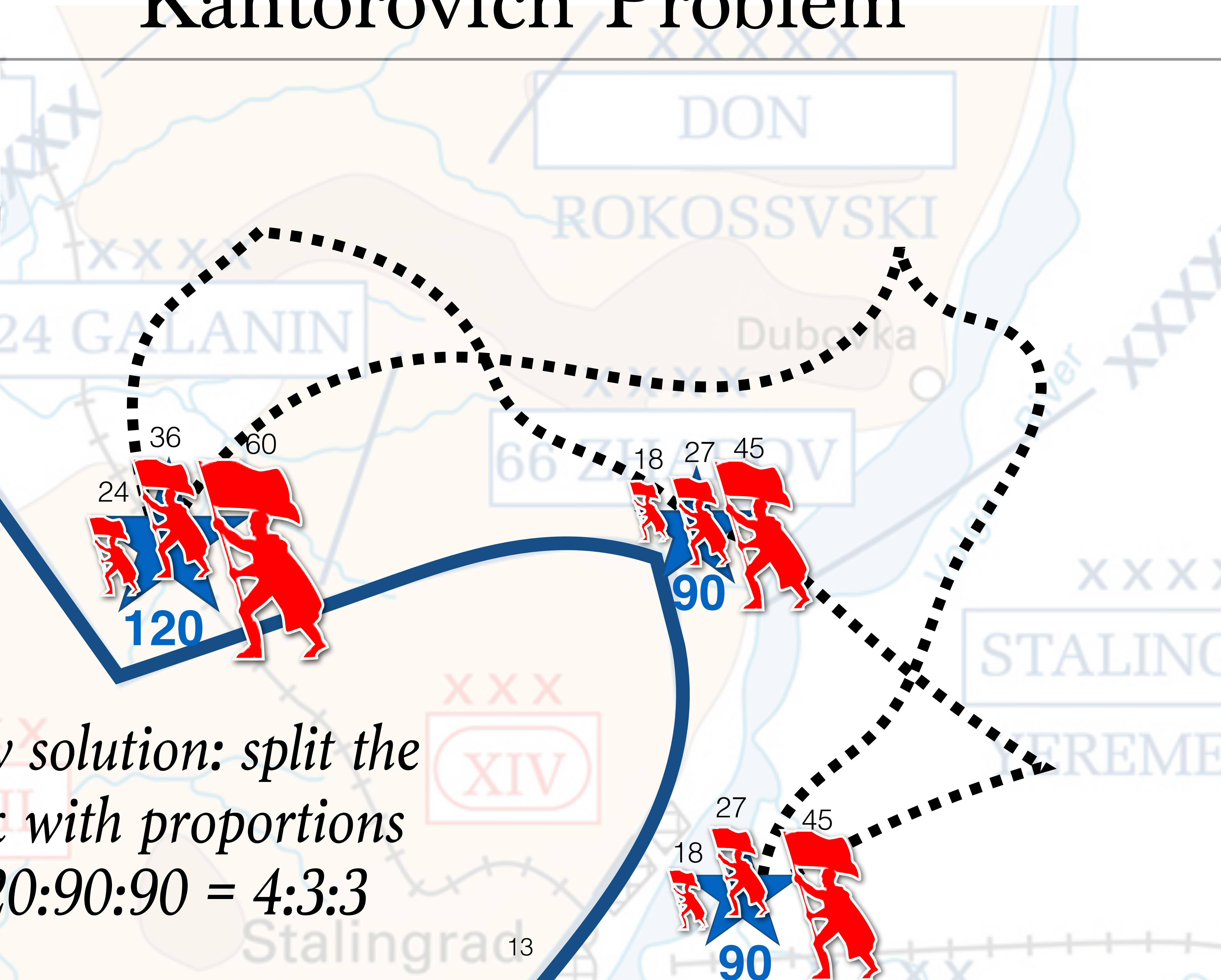
Easy solution: split the task with proportions 120:90:90 = 4:3:3

Kantorovich Problem



Easy solution: split the task with proportions 120:90:90 = 4:3:3

Kantorovich Problem



Easy solution: split the task with proportions 120:90:90 = 4:3:3

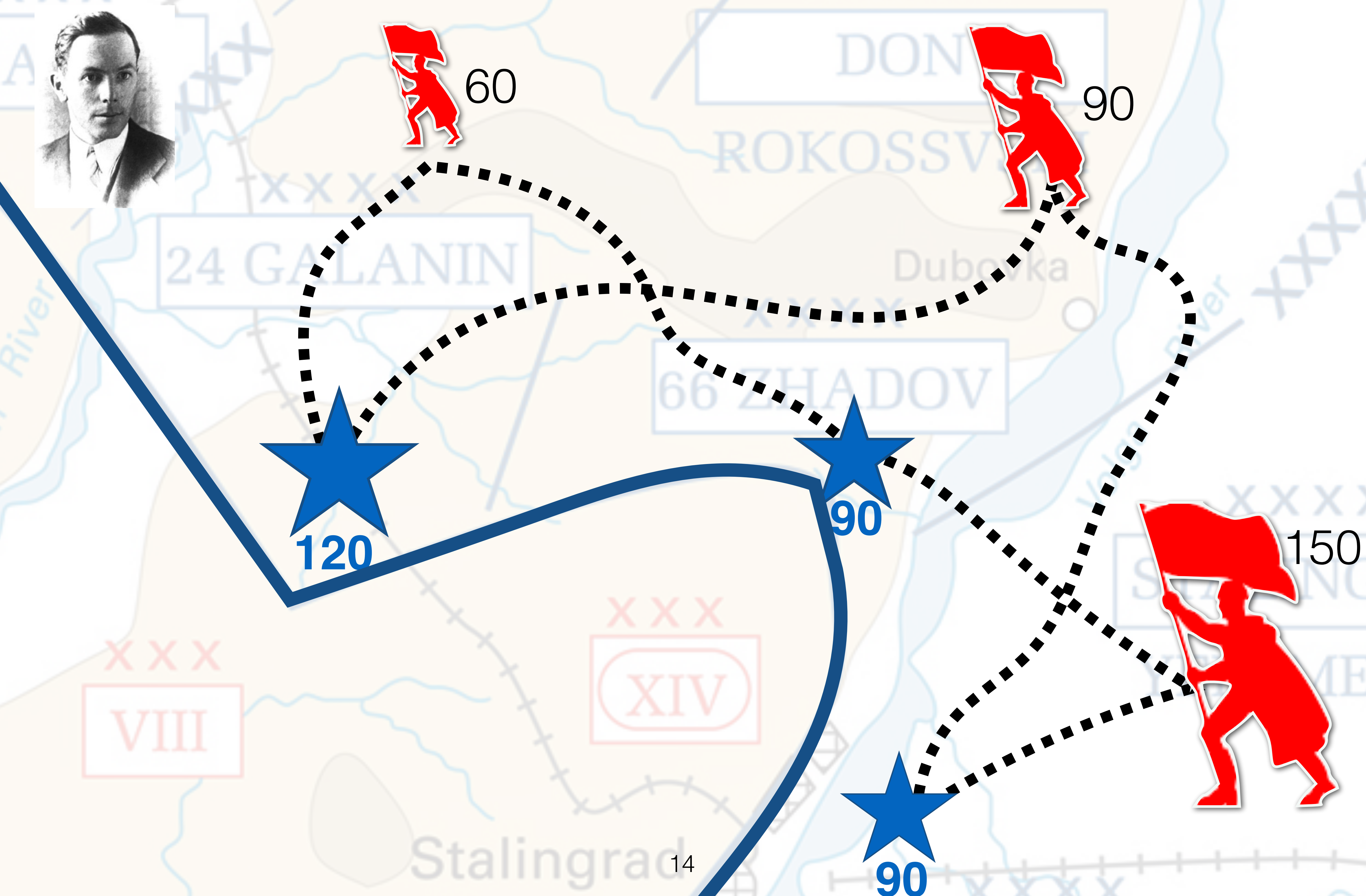
Kantorovich Problem

Naive approach results in too many displacements.

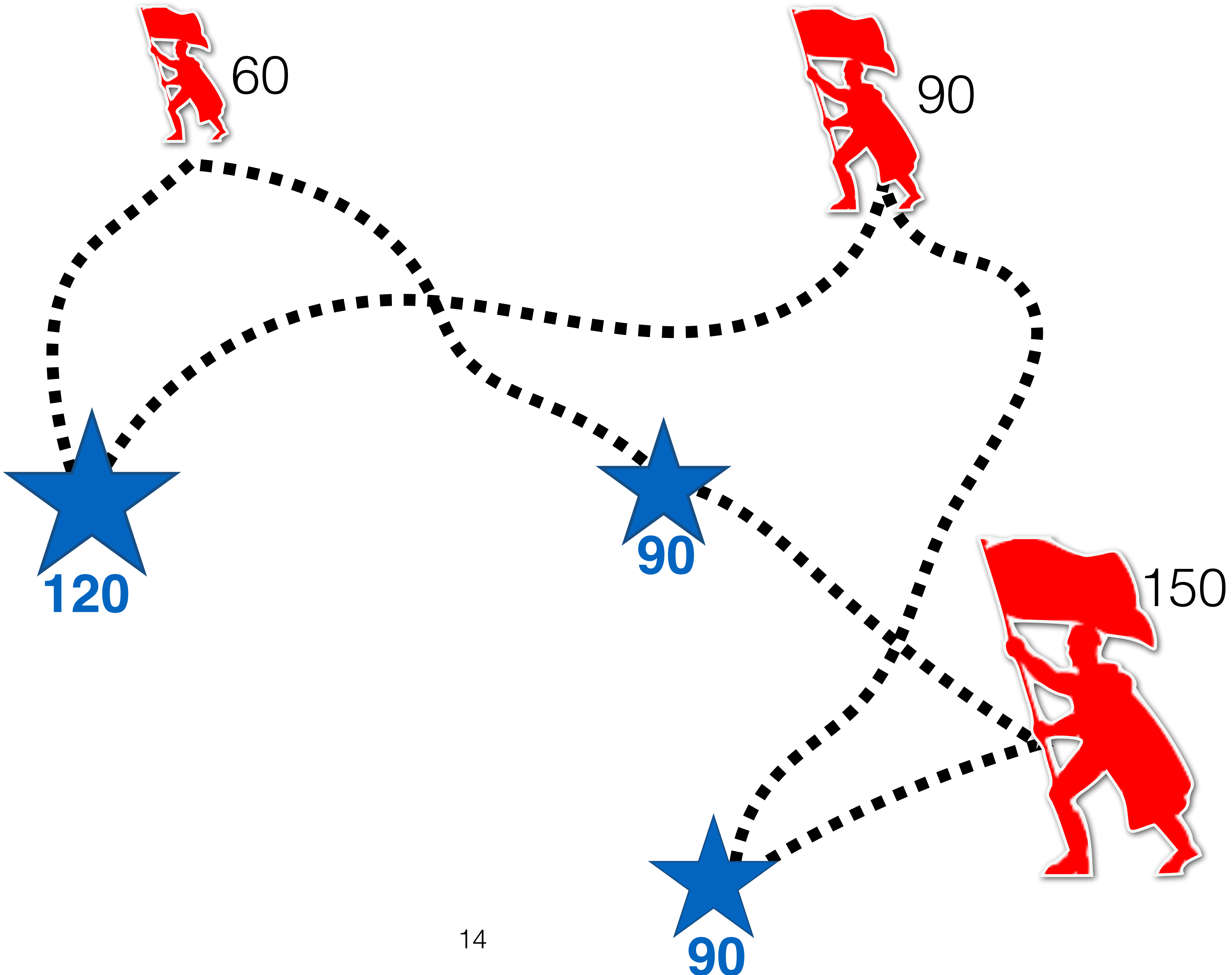
Goal: find a cheaper alternative

Easy solution: split the task with proportions 120:90:90 = 4:3:3

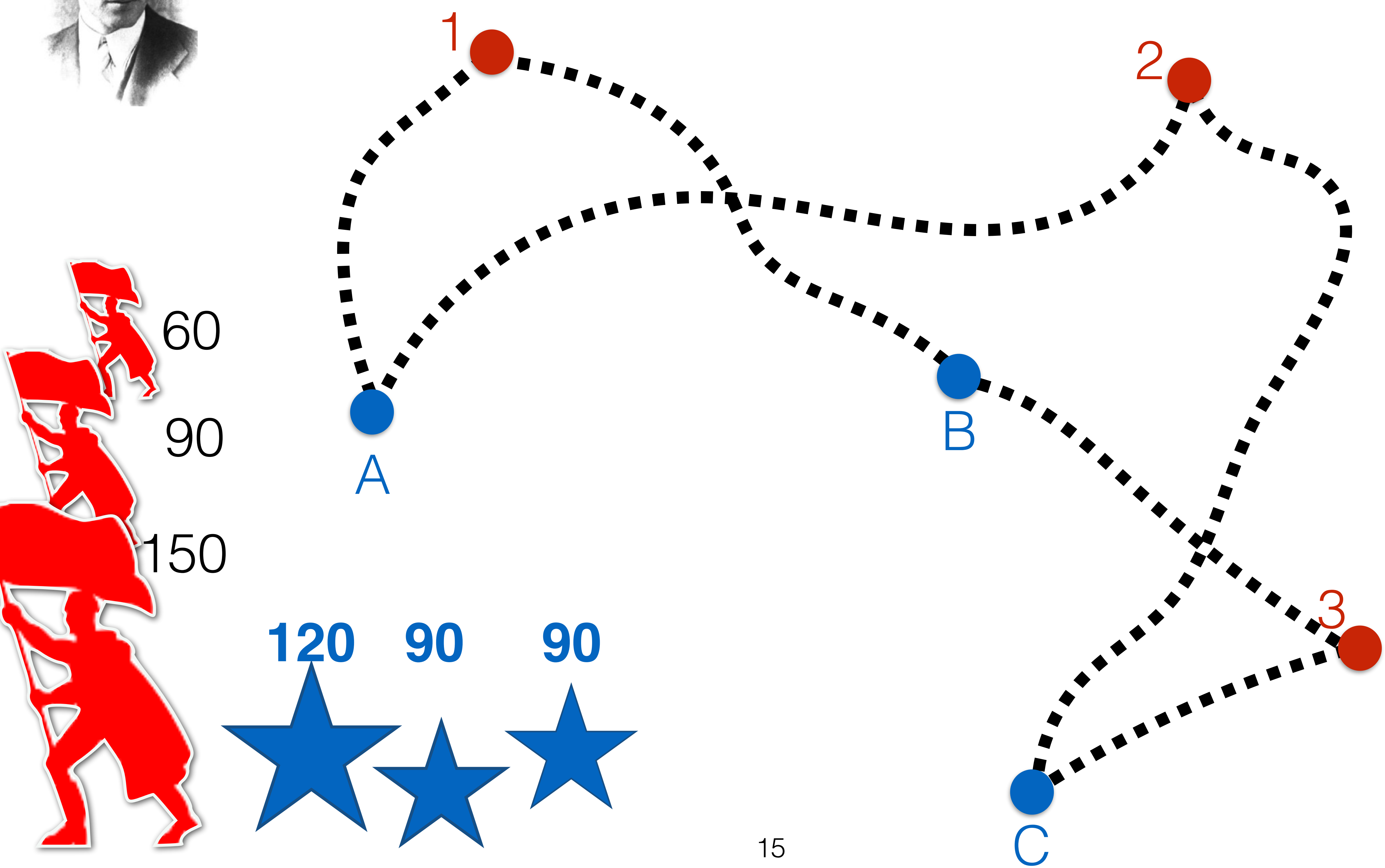
Kantorovich Problem



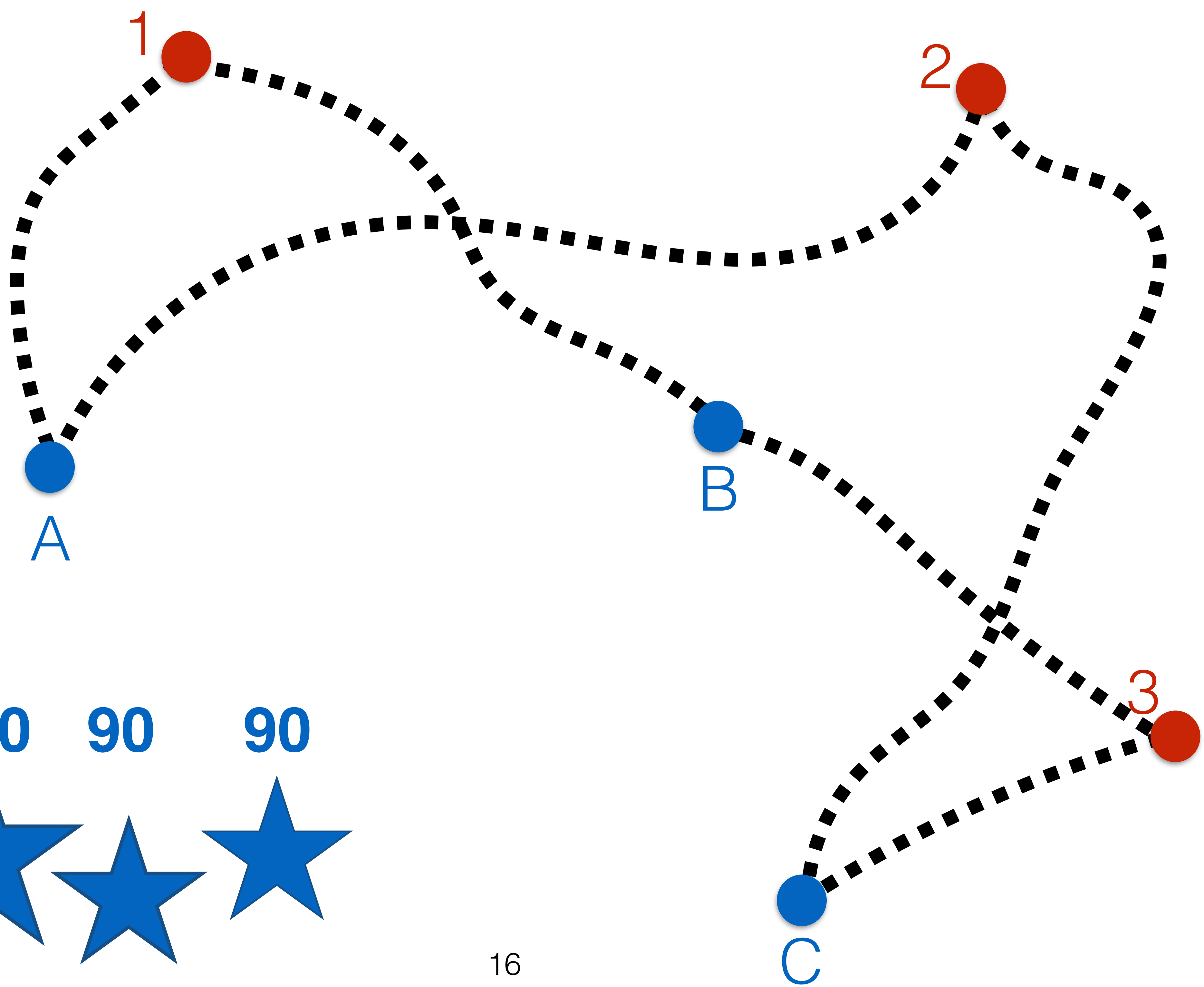
Kantorovich Problem

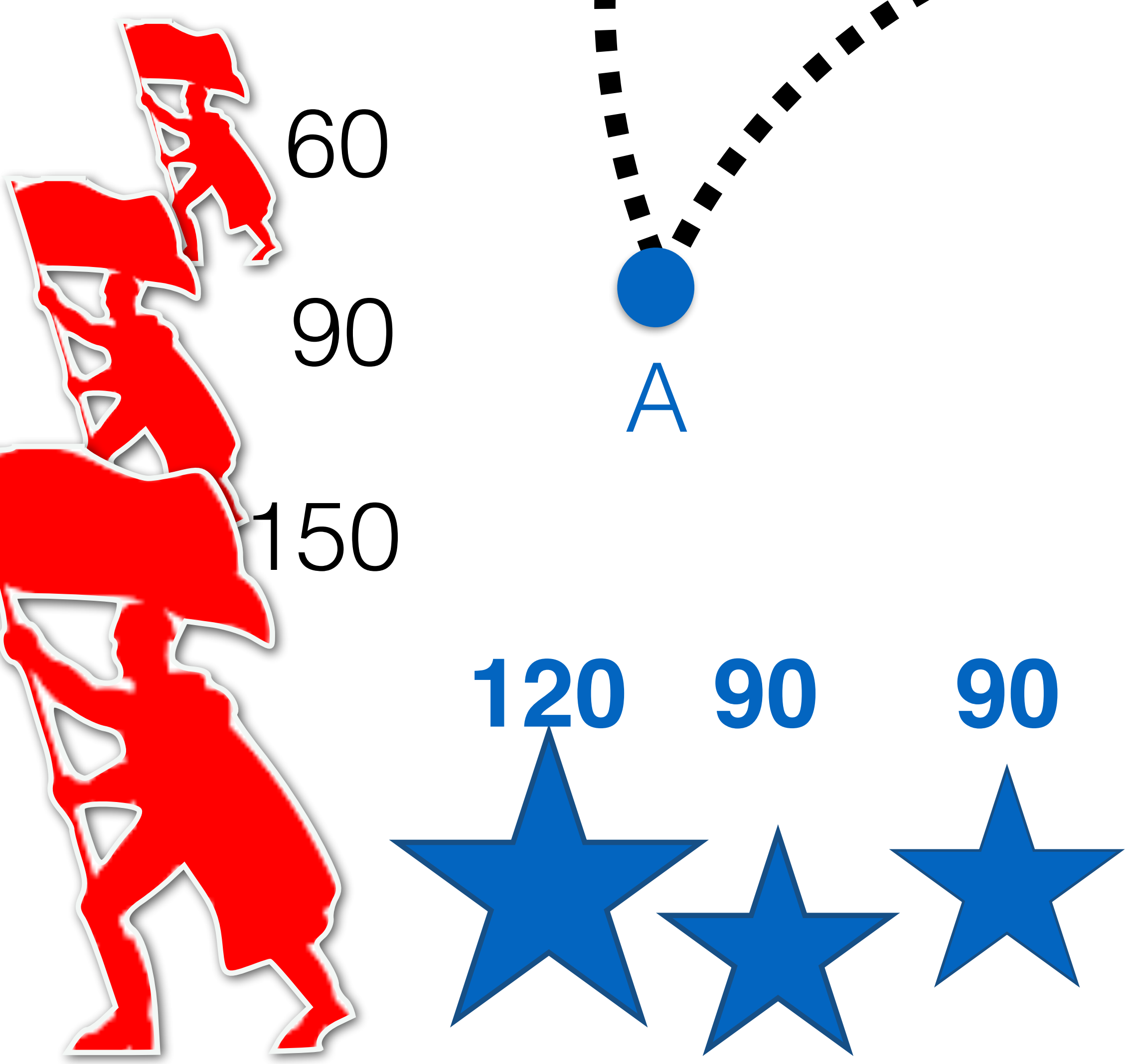


Kantorovich Problem



Kantorovich Problem





Kantorovich Problem

	120	90	90
60	?	?	?
90	?	?	?
150	?	?	?

	A	B	C
1	$d_{\mathbf{1}\mathrm{A}}$	$d_{\mathbf{1}\mathrm{B}}$	$d_{\mathbf{1}\mathrm{C}}$
2	$d_{\mathbf{2}\mathrm{A}}$	$d_{\mathbf{2}\mathrm{B}}$	$d_{\mathbf{2}\mathrm{C}}$
3	$d_{\mathbf{3}\mathrm{A}}$	$d_{\mathbf{3}\mathrm{B}}$	$d_{\mathbf{3}\mathrm{C}}$

Kantorovich Problem

Transportation matrix

60	?	?	?
90	?	?	?
150	?	?	?
	120	**90**	**90**

Distance matrix

1	$d_{\mathbf{1}\mathrm{A}}$	$d_{\mathbf{1}\mathrm{B}}$	$d_{\mathbf{1}\mathrm{C}}$
2	$d_{\mathbf{2}\mathrm{A}}$	$d_{\mathbf{2}\mathrm{B}}$	$d_{\mathbf{2}\mathrm{C}}$
3	$d_{\mathbf{3}\mathrm{A}}$	$d_{\mathbf{3}\mathrm{B}}$	$d_{\mathbf{3}\mathrm{C}}$
	A	B	C

Kantorovich Problem

The problem is entirely described by **counts** *and a* **cost/distance matrix**

Transportation matrix

60	?	?	?
90	?	?	?
150	?	?	?
	120	90	90

Distance matrix

	A	B	C
1	d_{1A}	d_{1B}	d_{1C}
2	d_{2A}	d_{2B}	d_{2C}
3	d_{3A}	d_{3B}	d_{3C}

Kantorovich Problem

Transportation matrix

60	?	?	?
90	?	?	?
150	?	?	?
	120	90	90

Distance matrix

1	d_{1A}	d_{1B}	d_{1C}
2	d_{2A}	d_{2B}	d_{2C}
3	d_{3A}	d_{3B}	d_{3C}
	A	B	C

Kantorovich Problem

Transportation matrix

60	p_{1A}	p_{1B}	p_{1C}
90	p_{2A}	p_{2B}	p_{2C}
150	p_{3A}	p_{3B}	p_{3C}
	120	**90**	**90**

Distance matrix

1	d_{1A}	d_{1B}	d_{1C}
2	d_{2A}	d_{2B}	d_{2C}
3	d_{3A}	d_{3B}	d_{3C}
	A	B	C

Kantorovich Problem

Transportation matrix

a_1	$p_{1\mathrm{A}}$	$p_{1\mathrm{B}}$	$p_{1\mathrm{C}}$
a_2	$p_{2\mathrm{A}}$	$p_{2\mathrm{B}}$	$p_{2\mathrm{C}}$
a_3	$p_{3\mathrm{A}}$	$p_{3\mathrm{B}}$	$p_{3\mathrm{C}}$
	b_{A}	b_{B}	b_{C}

Distance matrix

1	$d_{1\mathrm{A}}$	$d_{1\mathrm{B}}$	$d_{1\mathrm{C}}$
2	$d_{2\mathrm{A}}$	$d_{2\mathrm{B}}$	$d_{2\mathrm{C}}$
3	$d_{3\mathrm{A}}$	$d_{3\mathrm{B}}$	$d_{3\mathrm{C}}$
	A	B	C

Kantorovich Problem

Transportation matrix

$$\begin{array}{c|c|c|c|}
\cline{2-4}
a_1 & p_{1\mathrm{A}} & p_{1\mathrm{B}} & p_{1\mathrm{C}} \\
\cline{2-4}
a_2 & p_{2\mathrm{A}} & p_{2\mathrm{B}} & p_{2\mathrm{C}} \\
\cline{2-4}
a_3 & p_{3\mathrm{A}} & p_{3\mathrm{B}} & p_{3\mathrm{C}} \\
\cline{2-4}
\multicolumn{1}{c}{} & \multicolumn{1}{c}{b_\mathrm{A}} & \multicolumn{1}{c}{b_\mathrm{B}} & \multicolumn{1}{c}{b_\mathrm{C}}
\end{array}$$

Distance matrix

$$\begin{array}{c|c|c|c|}
\cline{2-4}
1 & d_{1\mathrm{A}} & d_{1\mathrm{B}} & d_{1\mathrm{C}} \\
\cline{2-4}
2 & d_{2\mathrm{A}} & d_{2\mathrm{B}} & d_{2\mathrm{C}} \\
\cline{2-4}
3 & d_{3\mathrm{A}} & d_{3\mathrm{B}} & d_{3\mathrm{C}} \\
\cline{2-4}
\multicolumn{1}{c}{} & \multicolumn{1}{c}{\mathrm{A}} & \multicolumn{1}{c}{\mathrm{B}} & \multicolumn{1}{c}{\mathrm{C}}
\end{array}$$

Constraints

$$\forall i \in \{1,2,3\}, \quad \sum_{j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}} p_{ij} = a_i$$

$$\forall j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}, \quad \sum_{i \in \{1,2,3\}} p_{ij} = b_j$$

$$p_{ij} \geq 0$$

Kantorovich Problem

Transportation matrix

a_1	$p_{1\mathrm{A}}$	$p_{1\mathrm{B}}$	$p_{1\mathrm{C}}$
a_2	$p_{2\mathrm{A}}$	$p_{2\mathrm{B}}$	$p_{2\mathrm{C}}$
a_3	$p_{3\mathrm{A}}$	$p_{3\mathrm{B}}$	$p_{3\mathrm{C}}$
	b_A	b_B	b_C

Distance matrix

1	$d_{1\mathrm{A}}$	$d_{1\mathrm{B}}$	$d_{1\mathrm{C}}$
2	$d_{2\mathrm{A}}$	$d_{2\mathrm{B}}$	$d_{2\mathrm{C}}$
3	$d_{3\mathrm{A}}$	$d_{3\mathrm{B}}$	$d_{3\mathrm{C}}$
	A	B	C

Constraints

$$\forall i \in \{1,2,3\}, \sum_{j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}} p_{ij} = a_i$$

$$\forall j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}, \sum_{i \in \{1,2,3\}} p_{ij} = b_j$$

$$p_{ij} \geq 0$$

Cost function

$$C(P) = \sum_{j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}} \sum_{i \in \{1,2,3\}} p_{ij} d_{ij}$$

Kantorovich Problem

Transportation matrix

a_1	$p_{1\mathrm{A}}$	$p_{1\mathrm{B}}$	$p_{1\mathrm{C}}$
a_2	$p_{2\mathrm{A}}$	$p_{2\mathrm{B}}$	$p_{2\mathrm{C}}$
a_3	$p_{3\mathrm{A}}$	$p_{3\mathrm{B}}$	$p_{3\mathrm{C}}$
	b_A	b_B	b_C

Distance matrix

1	$d_{1\mathrm{A}}$	$d_{1\mathrm{B}}$	$d_{1\mathrm{C}}$
2	$d_{2\mathrm{A}}$	$d_{2\mathrm{B}}$	$d_{2\mathrm{C}}$
3	$d_{3\mathrm{A}}$	$d_{3\mathrm{B}}$	$d_{3\mathrm{C}}$
	A	B	C

Constraints

$$\forall i \in \{1,2,3\}, \quad \sum_{j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}} p_{ij} = a_i$$

$$\forall j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}, \quad \sum_{i \in \{1,2,3\}} p_{ij} = b_j$$

$$p_{ij} \geq 0$$

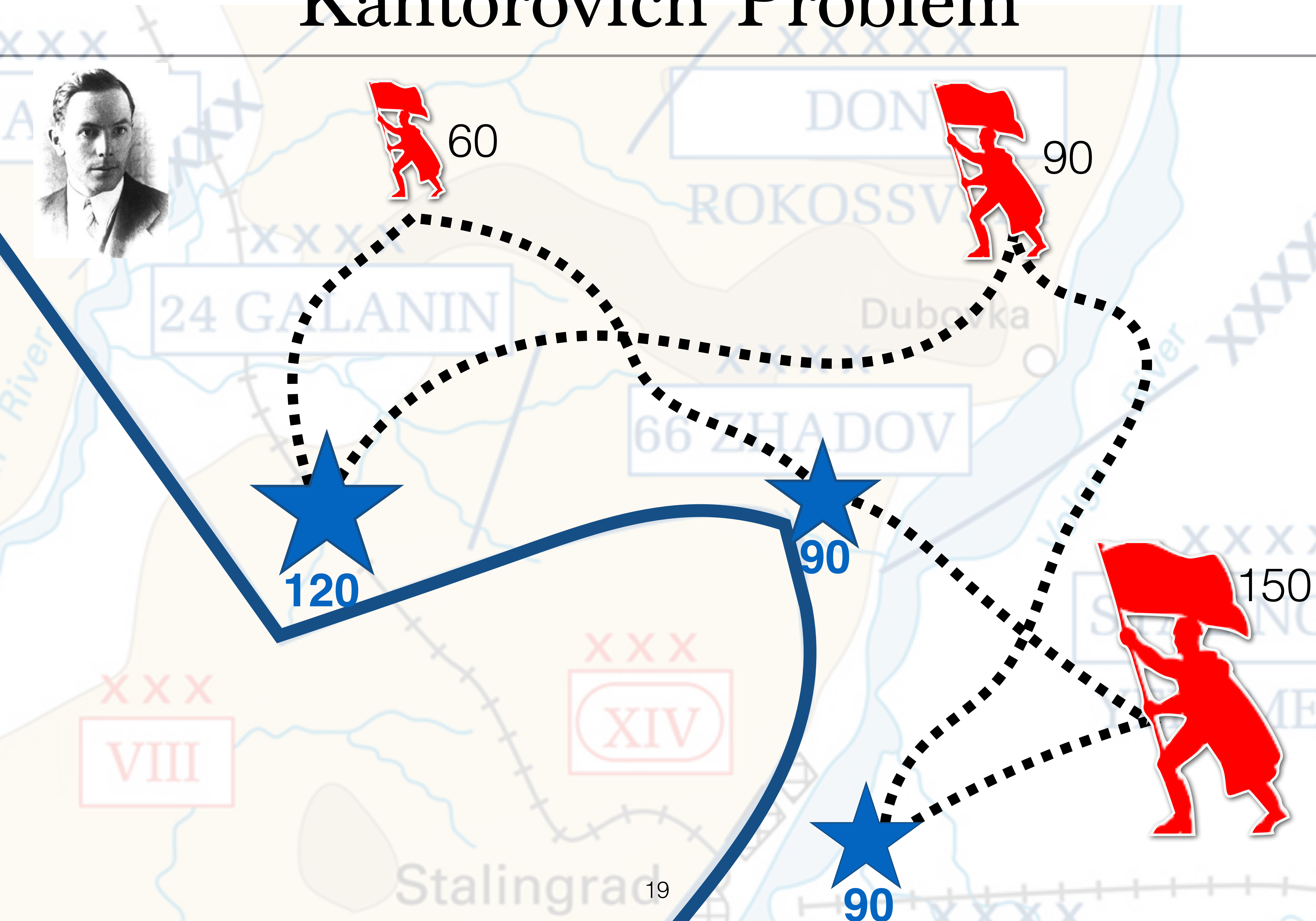
Cost function

$$C(P) = \sum_{j \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}} \sum_{i \in \{1,2,3\}} p_{ij} d_{ij}$$

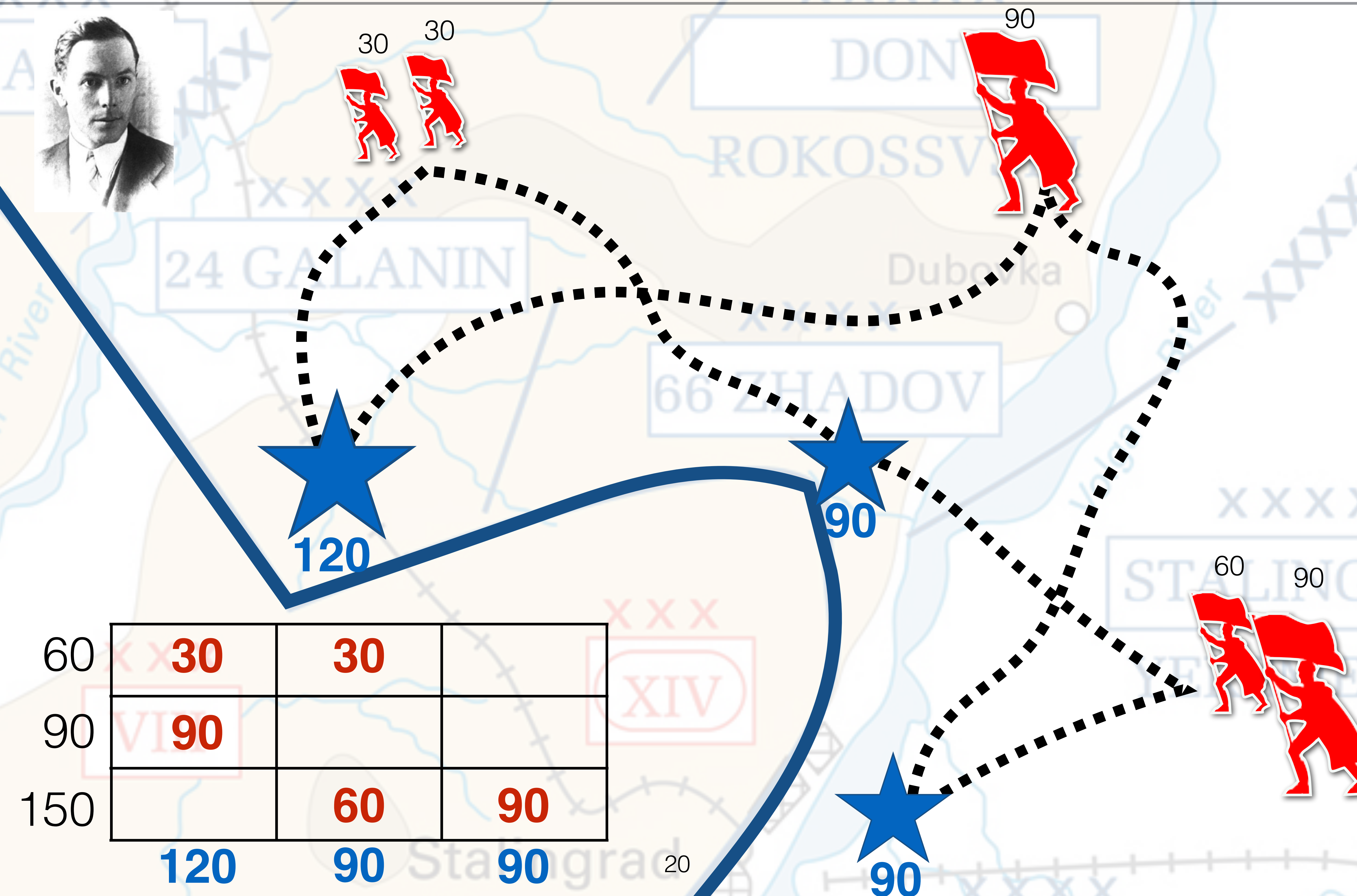
Problem

$$\min_{\text{all valid } P} C(P)$$

Kantorovich Problem

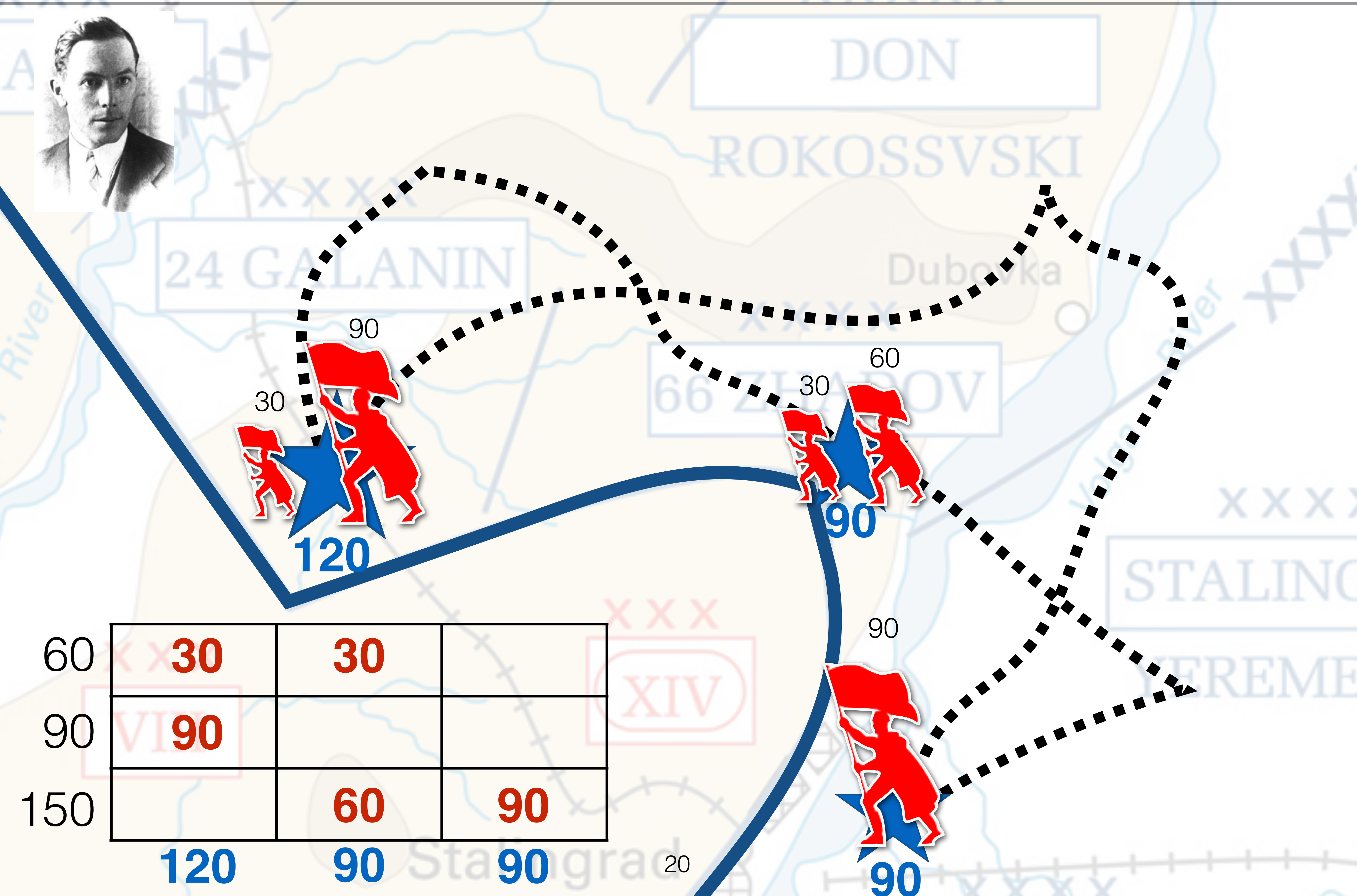


Kantorovich Problem



60	30	30	
90	90		
150		60	90
	120	90	90

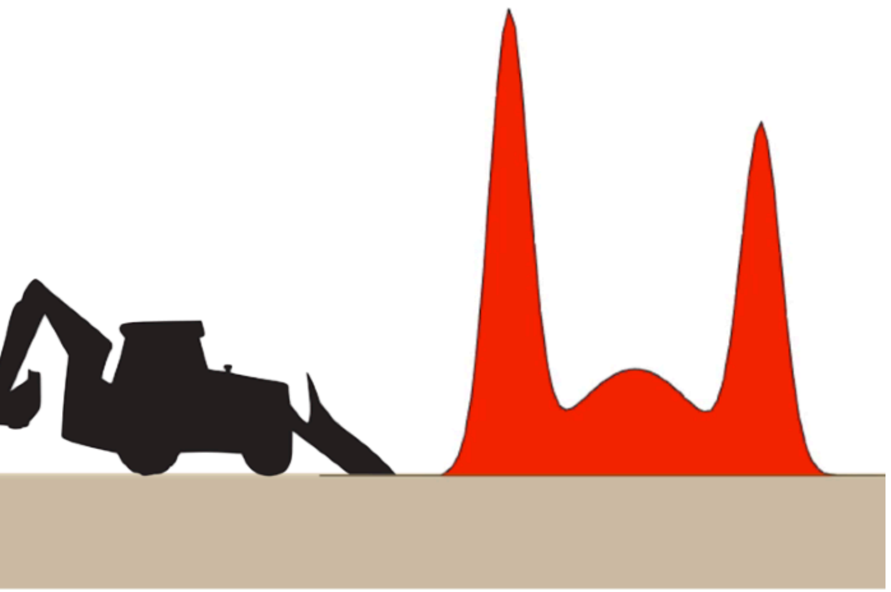
Kantorovich Problem



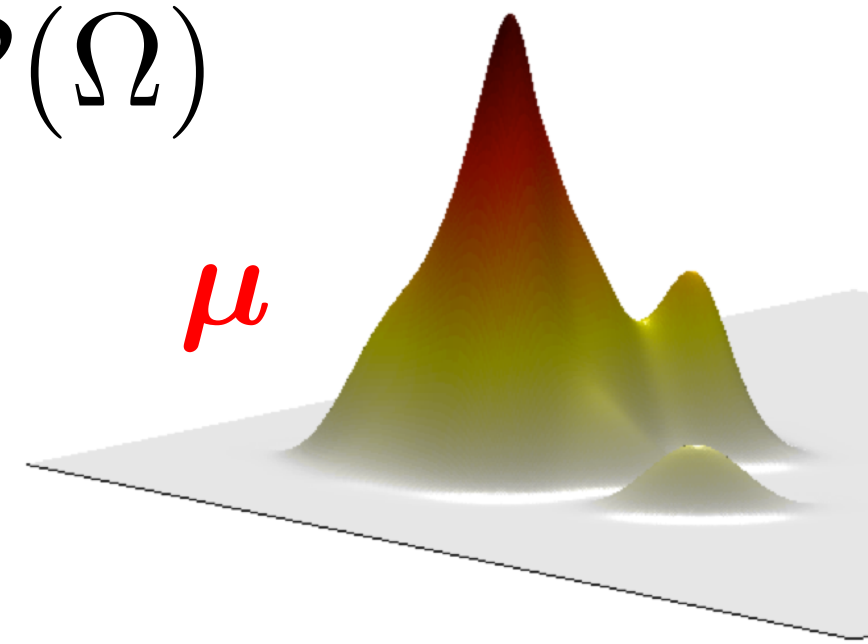
60	30	30	
90	90		
150		60	90
	120	90	90

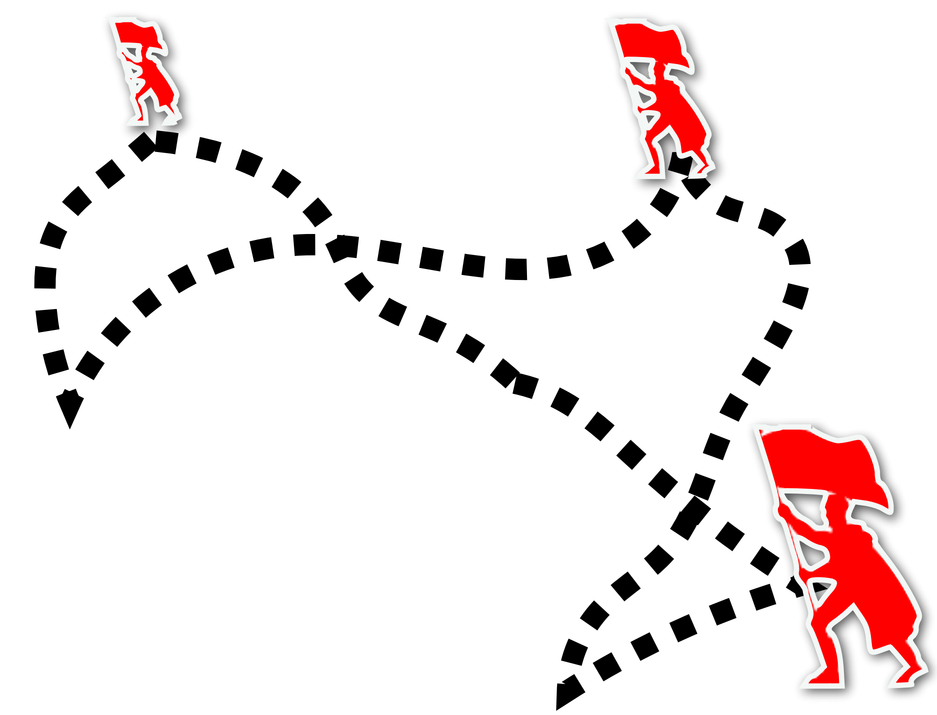
Mathematical Formalism

These problems involve discrete and continuous **probability measures** on a geometric space



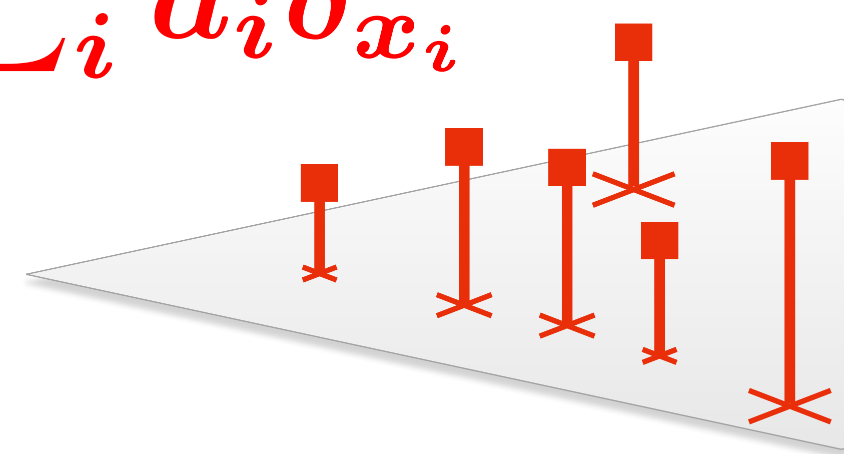
$\mathcal{P}(\Omega)$





$\longmapsto$

$\sum_i a_i \delta_{x_i}$

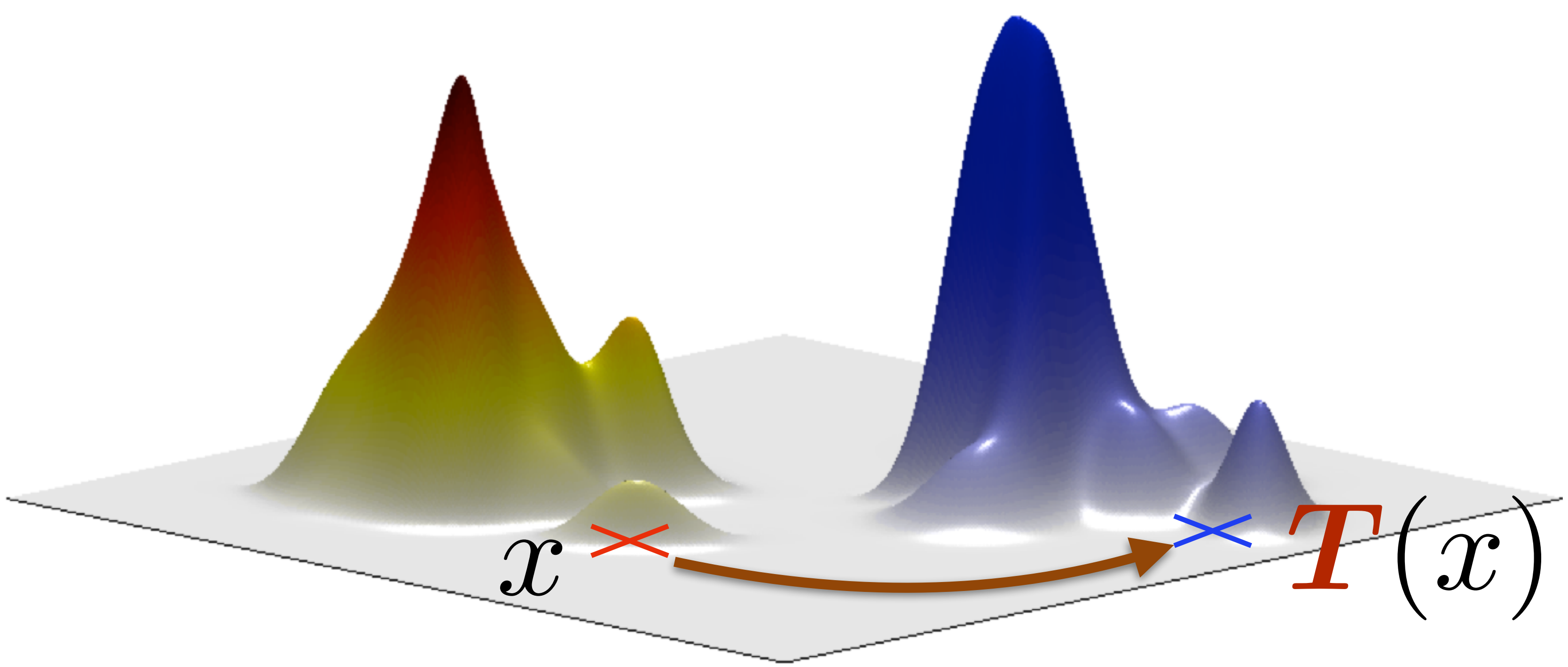


Monge Problem

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 $\boldsymbol{T} : \Omega \to \Omega$

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

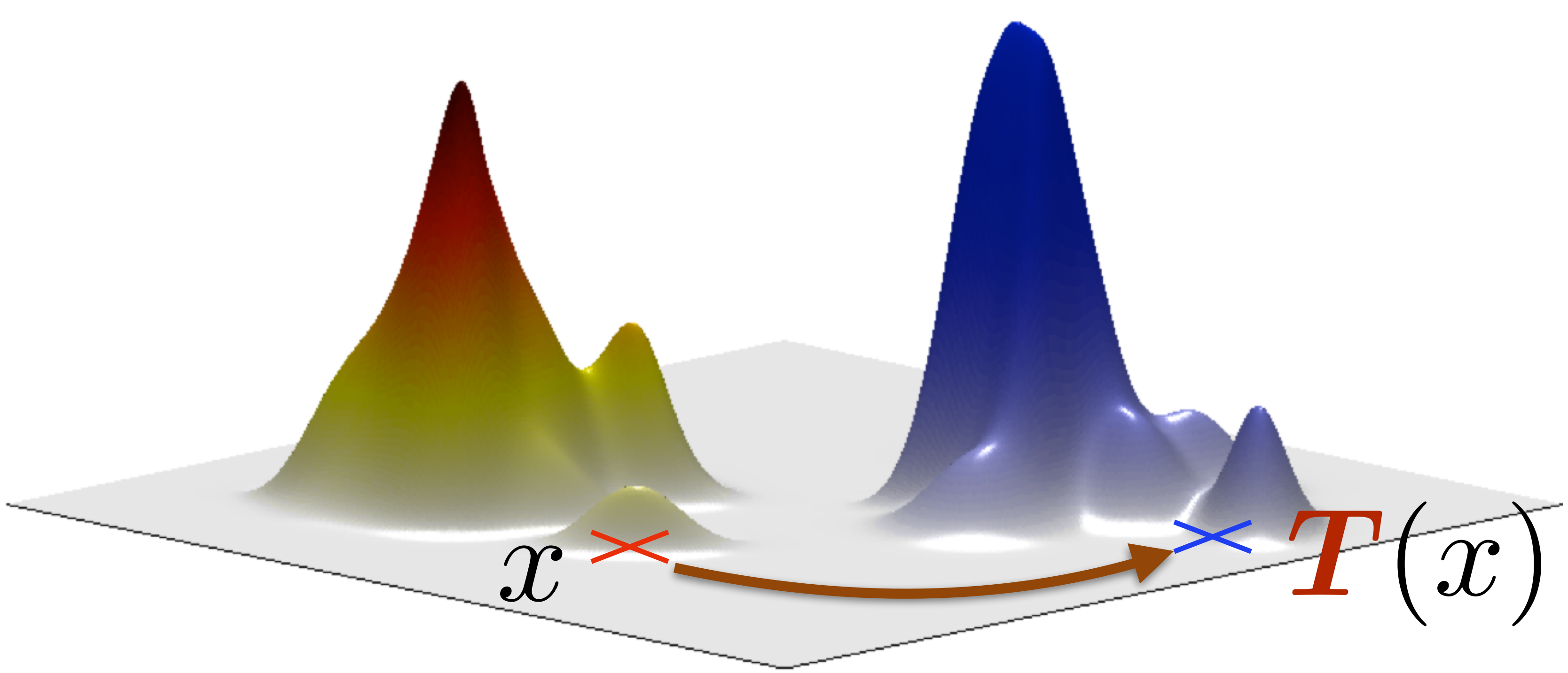


Monge Problem

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

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

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

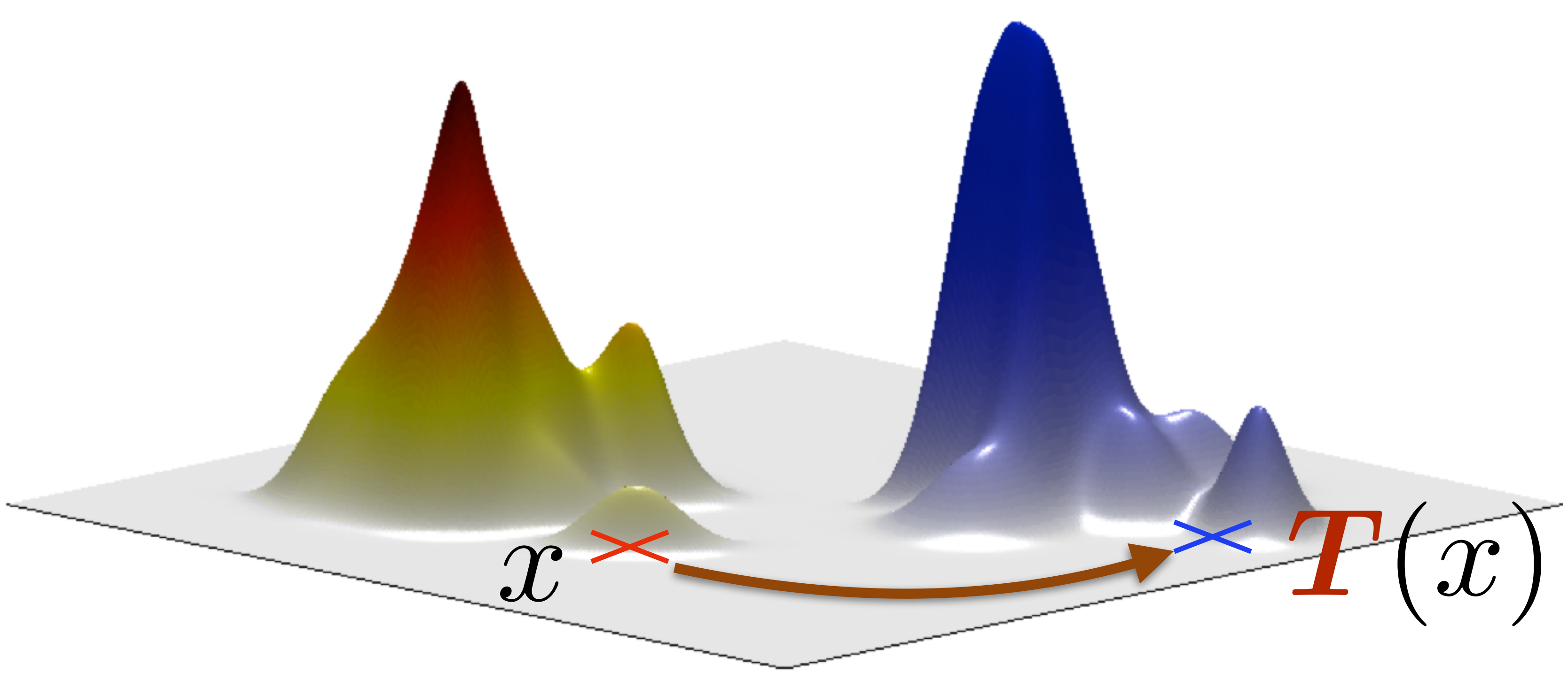


Monge Problem

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 $\boldsymbol{T} : \Omega \to \Omega$

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

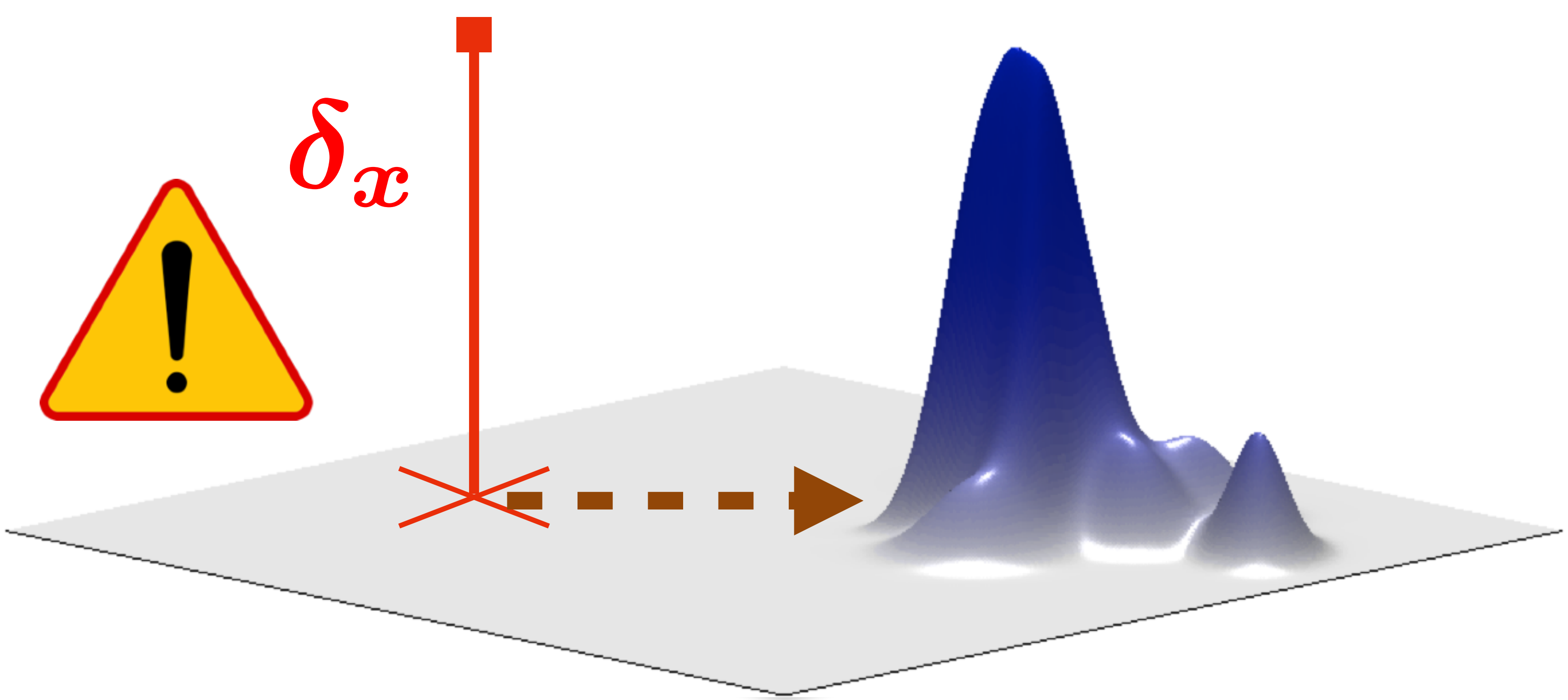


Monge Problem

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 $\boldsymbol{T} : \Omega \to \Omega$

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



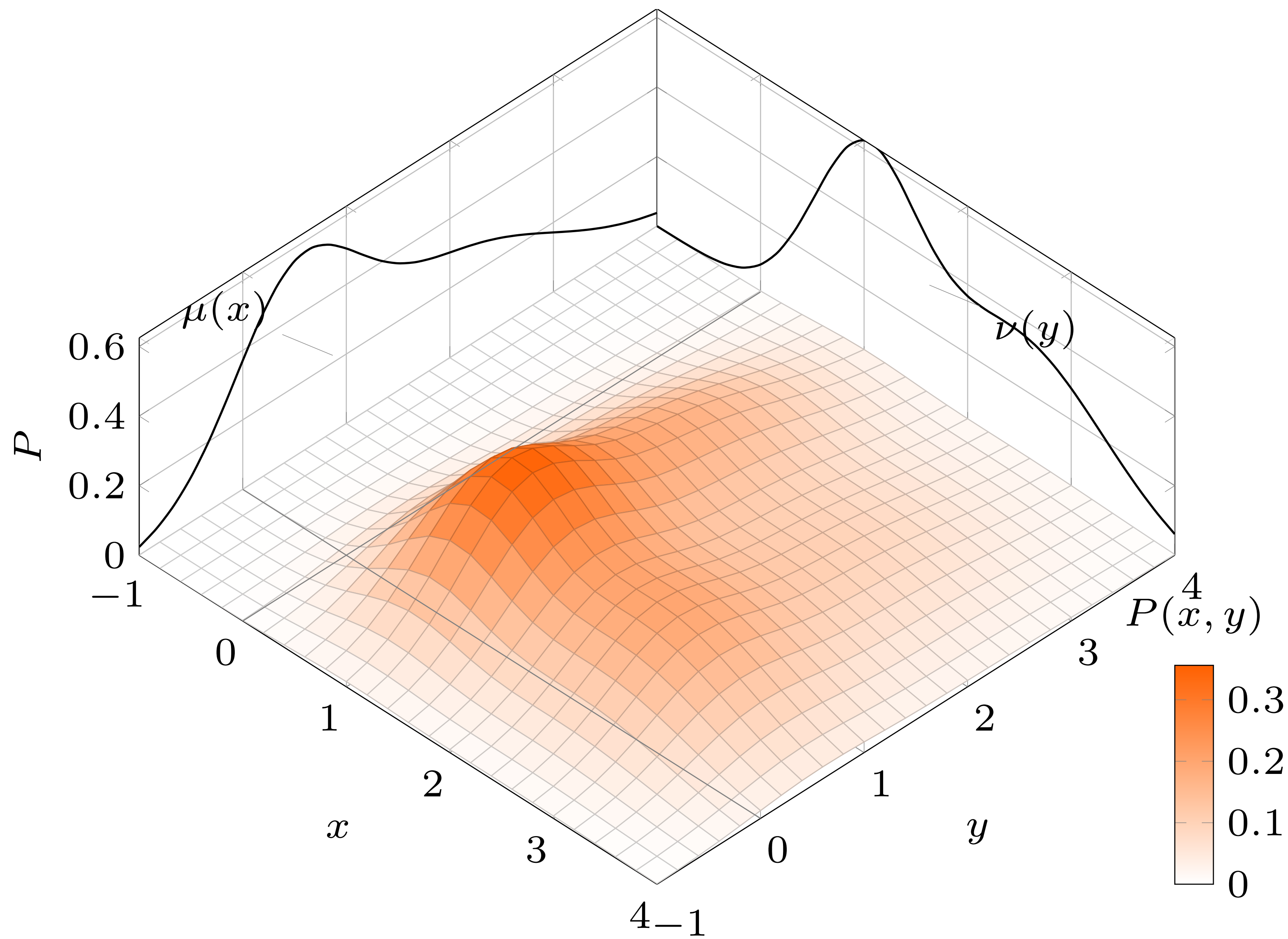
Kantorovich Relaxation

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

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

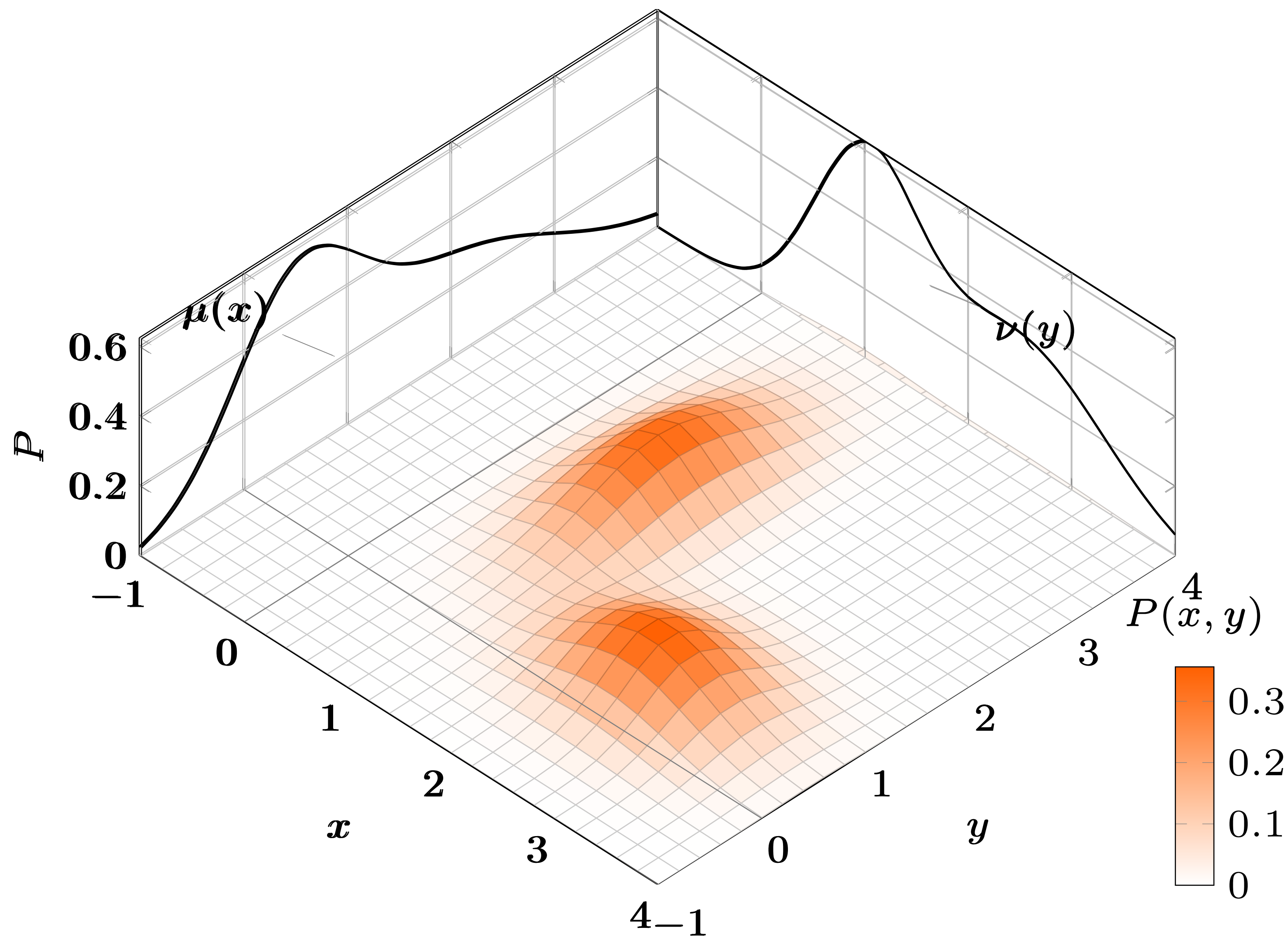
Kantorovich Relaxation

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



Kantorovich 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 Problem

Def. Given $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$; a cost function $\boldsymbol{c}$ on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{c}(x, y) \boldsymbol{P}(dx, dy).$$

PRIMAL

Kantorovich Problem

Def. Given μ, ν in $\mathcal{P}(\Omega)$; a cost function c on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{P \in \Pi(\mu,\nu)} \iint c(x,y) P(dx, dy).$$

PRIMAL

$$\inf_{P \in \Pi(\mu,\nu)} \mathbb{E}_P[c(X,Y)]$$

Kantorovich Problem

Def. Given $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$; a cost function $\boldsymbol{c}$ on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{c}(x, y) \boldsymbol{P}(dx, dy).$$

PRIMAL

$$\sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{c}(x,y)}} \int \boldsymbol{\varphi} \, d\boldsymbol{\mu} + \int \boldsymbol{\psi} \, d\boldsymbol{\nu}.$$

DUAL

Links between Monge & Kantorovich

Prop. For "well behaved" costs $\boldsymbol{c}$, if $\boldsymbol{\mu}$ has a density then an *optimal* Monge map $T^\star$ between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ must exist.

Prop. In that case

$$\boldsymbol{P}^\star := (\mathrm{Id}, T^\star)_\sharp \boldsymbol{\mu} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$$

is also *optimal* for the Kantorovich problem.

[Brenier'91] [Smith&Knott'87] [McCann'01]

(Kantorovich) Wasserstein Distances

Let $p \geq 1$. Let $\boldsymbol{c} := \boldsymbol{D}$, a metric.

Def. The p-**Wasserstein distance** between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$ is

$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \left(\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy) \right)^{1/p}.$$

(Kantorovich) Wasserstein Distances

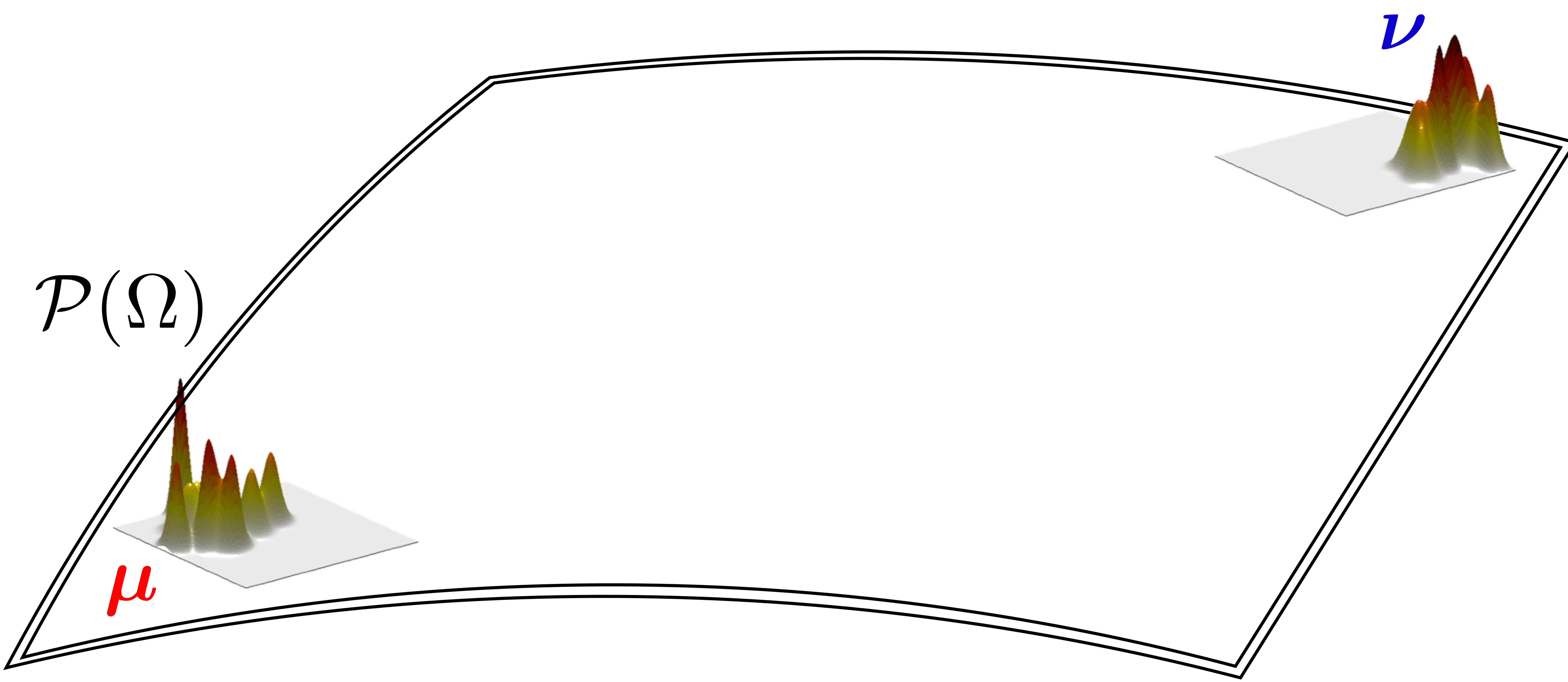
Let $p \geq 1$. Let $\boldsymbol{c} := \boldsymbol{D}$, a metric.

Def. The p-**Wasserstein distance** between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$ is

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \left(\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy) \right)^{\cancel{1/p}}.$$

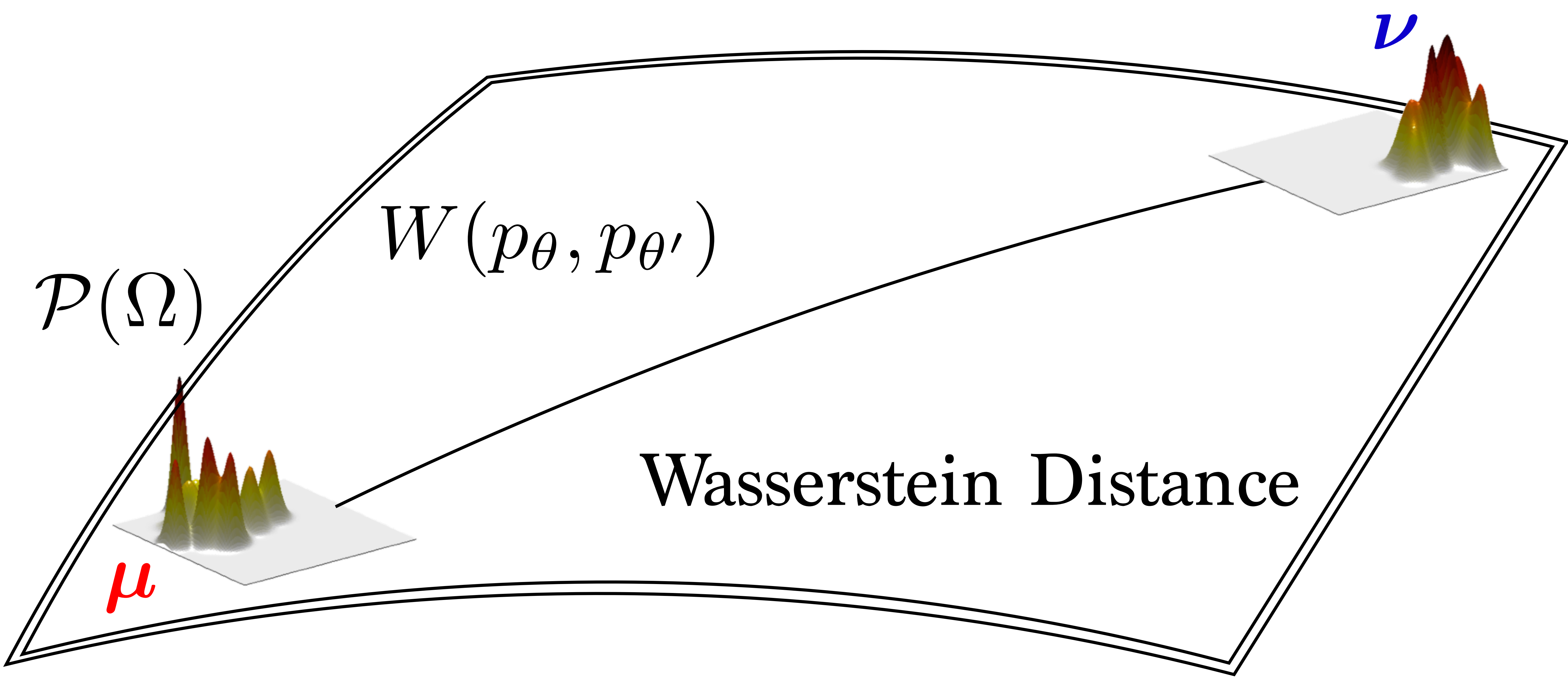
Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)



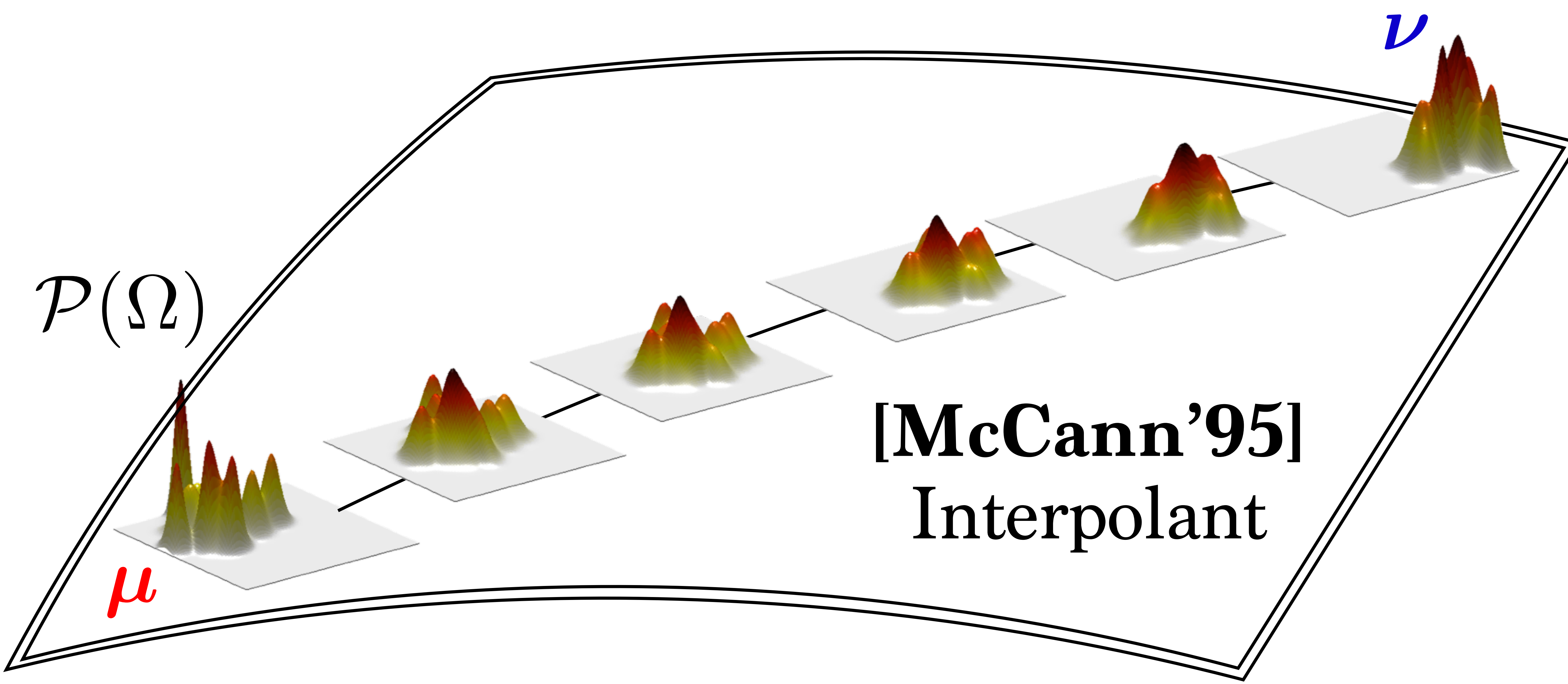
Optimal Transport Geometry

Very different geometry than standard information divergences (KL, Euclidean)



Optimal Transport Geometry

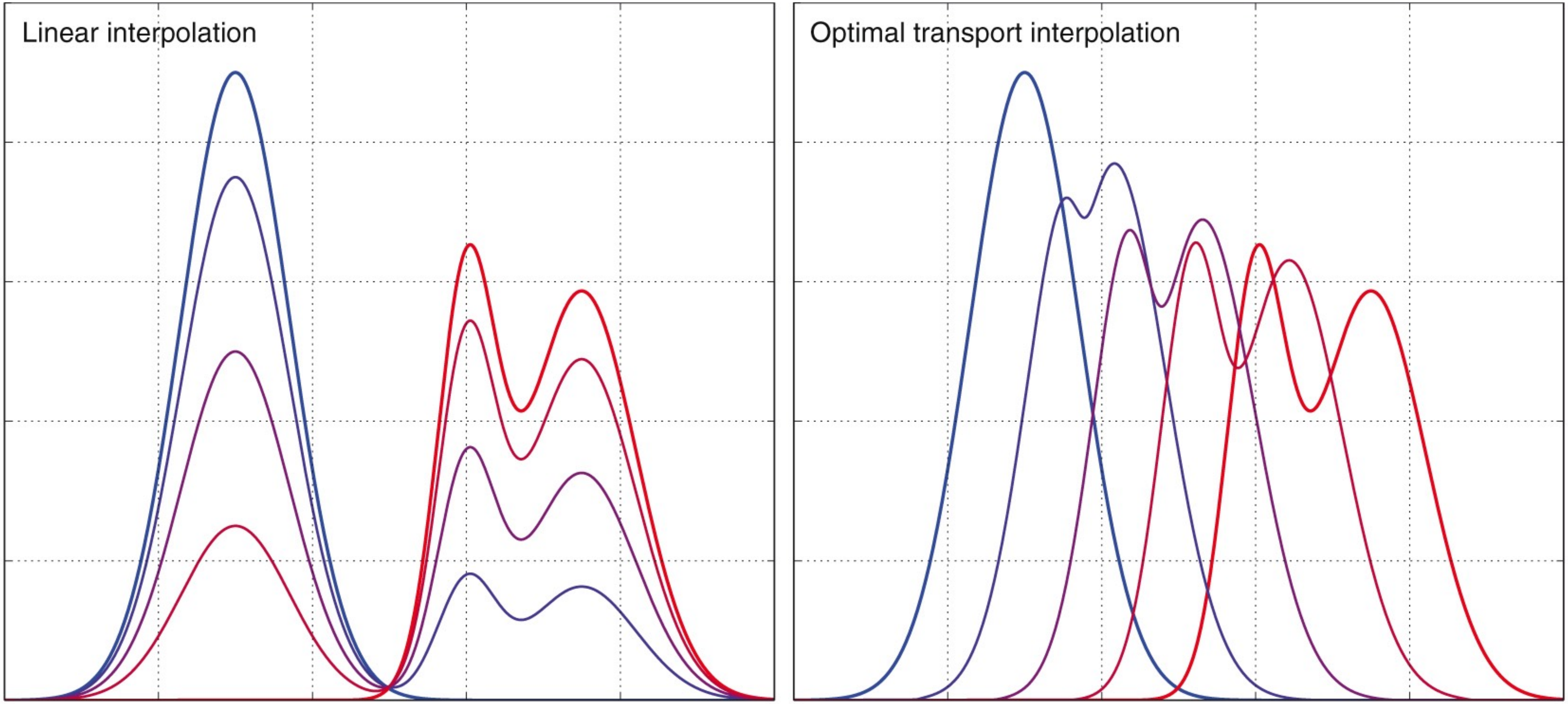
Very different geometry than standard information divergences (KL, Euclidean)



[McCann'95]
Interpolant

Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)



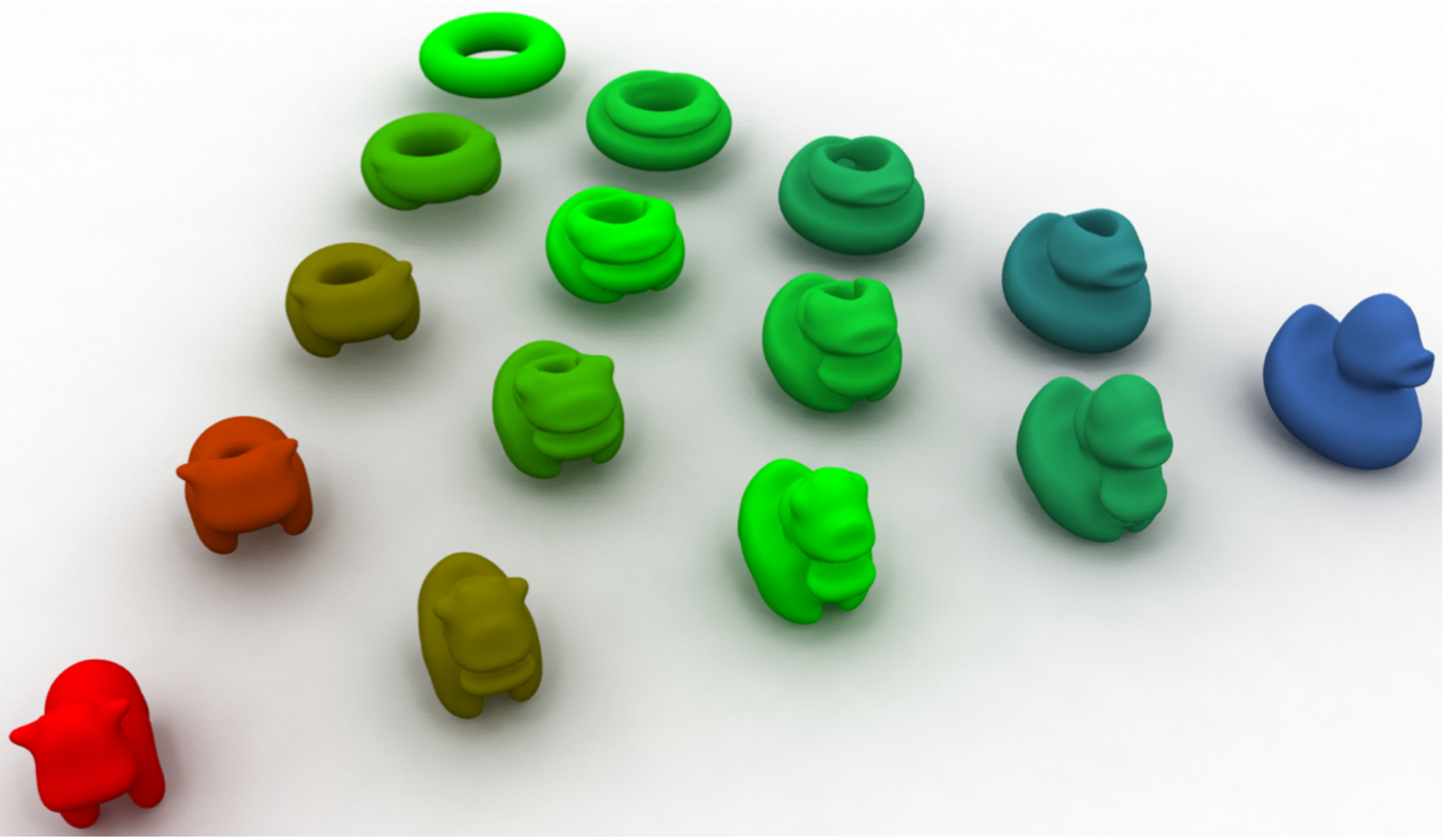
Optimal Transport Geometry

Very different geometry than standard information divergences (KL, Euclidean)

[Solomon'15]

Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)



[Solomon'15]

Computational OT

Up to 2010: OT solvers $W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = ?$

Goal now: use OT as a **loss or fidelity** term

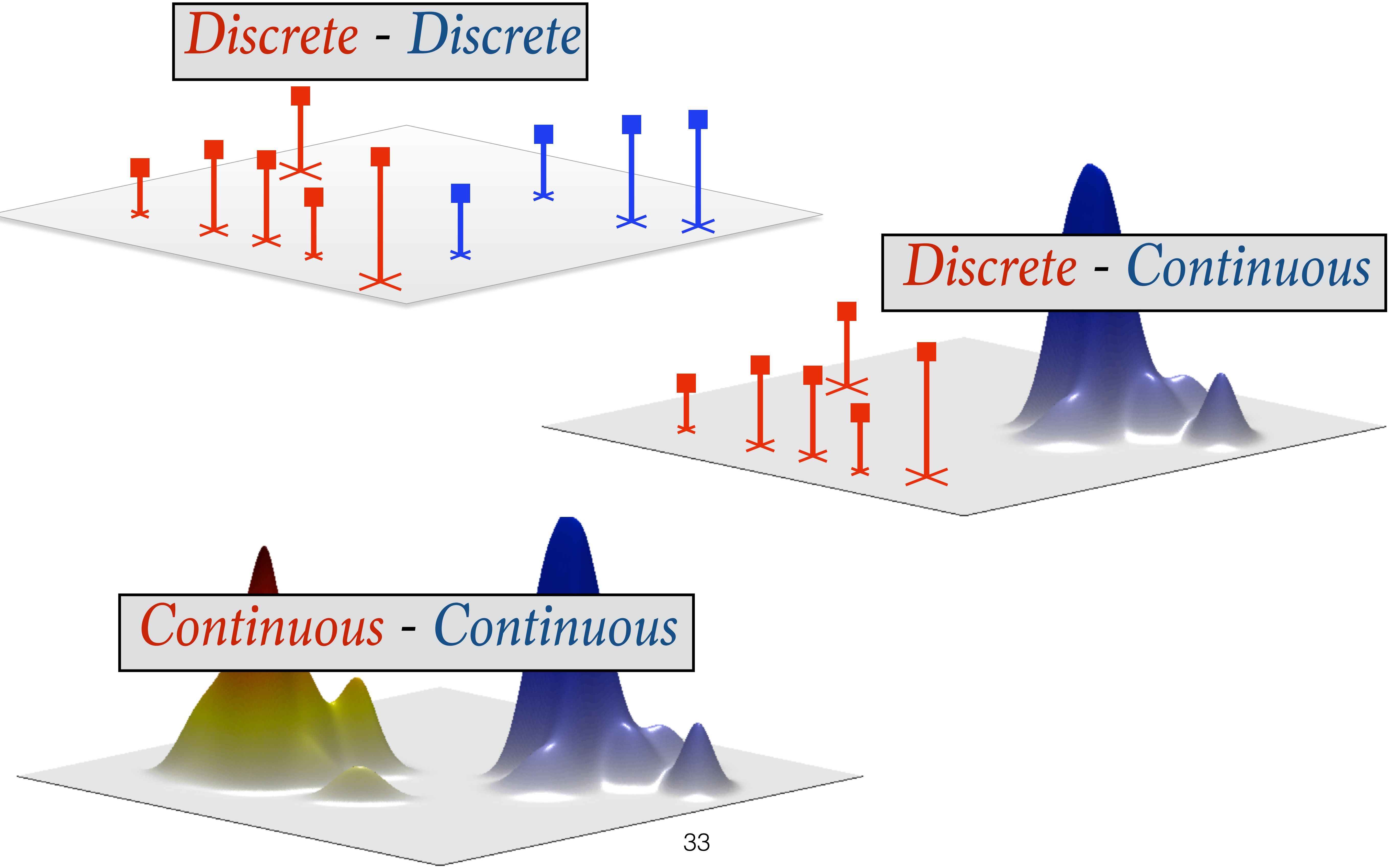
$$\underset{\boldsymbol{\mu} \in \mathcal{P}(\Omega)}{\operatorname{argmin}} \; F(W_p(\boldsymbol{\mu}, \boldsymbol{\nu_1}), W_p(\boldsymbol{\mu}, \boldsymbol{\nu_2}), \ldots, \boldsymbol{\mu}) = ?$$

$$\nabla_{\boldsymbol{\mu}} W_p(\boldsymbol{\mu}, \boldsymbol{\nu_1}) = ?$$

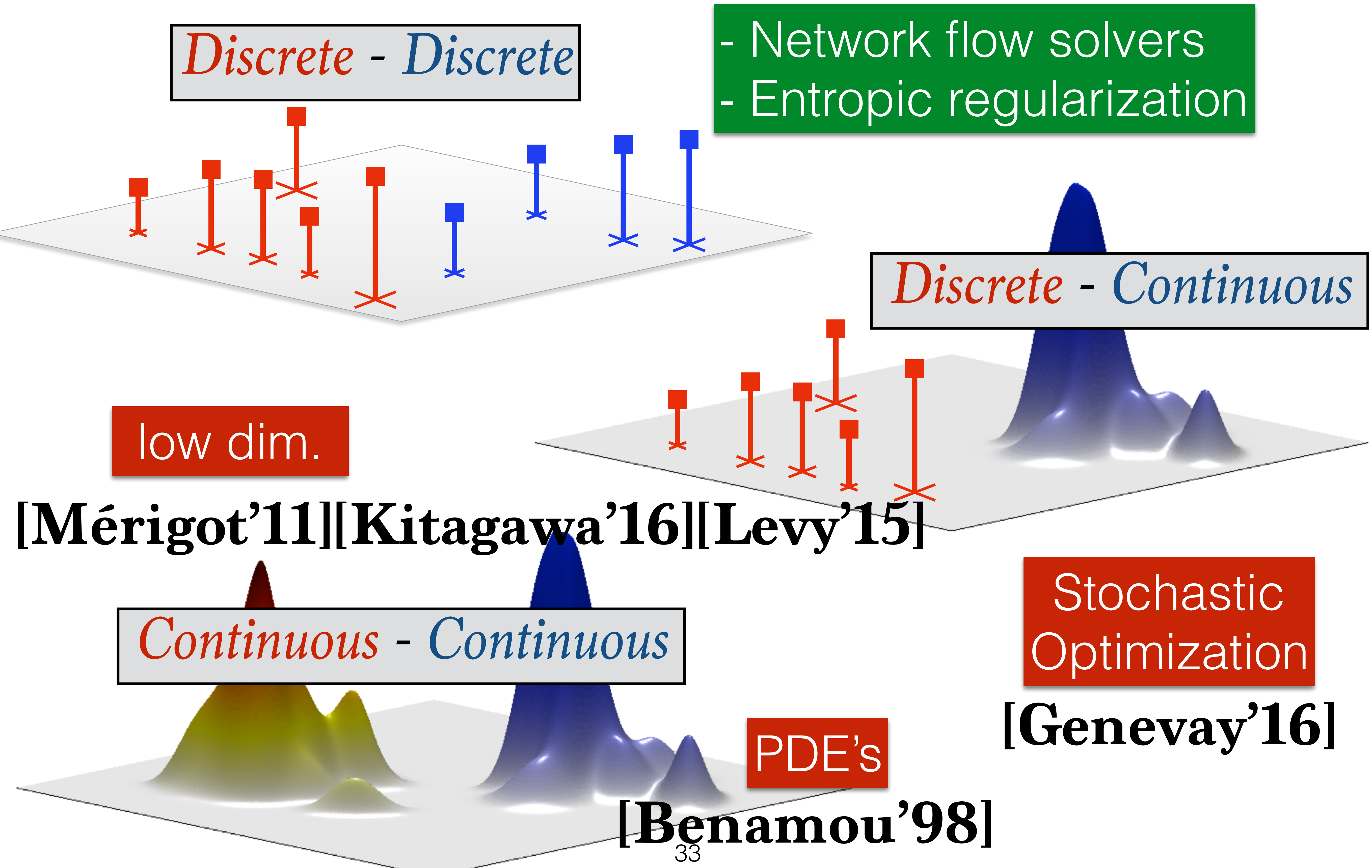
2. How to compute OT

- Typology: discrete/continuous problems
- Easy cases, zoo of solvers
- Entropic regularization
- Differentiability of the W distance

How can we compute OT?



How can we compute OT?



Easy (1): Univariate Measures

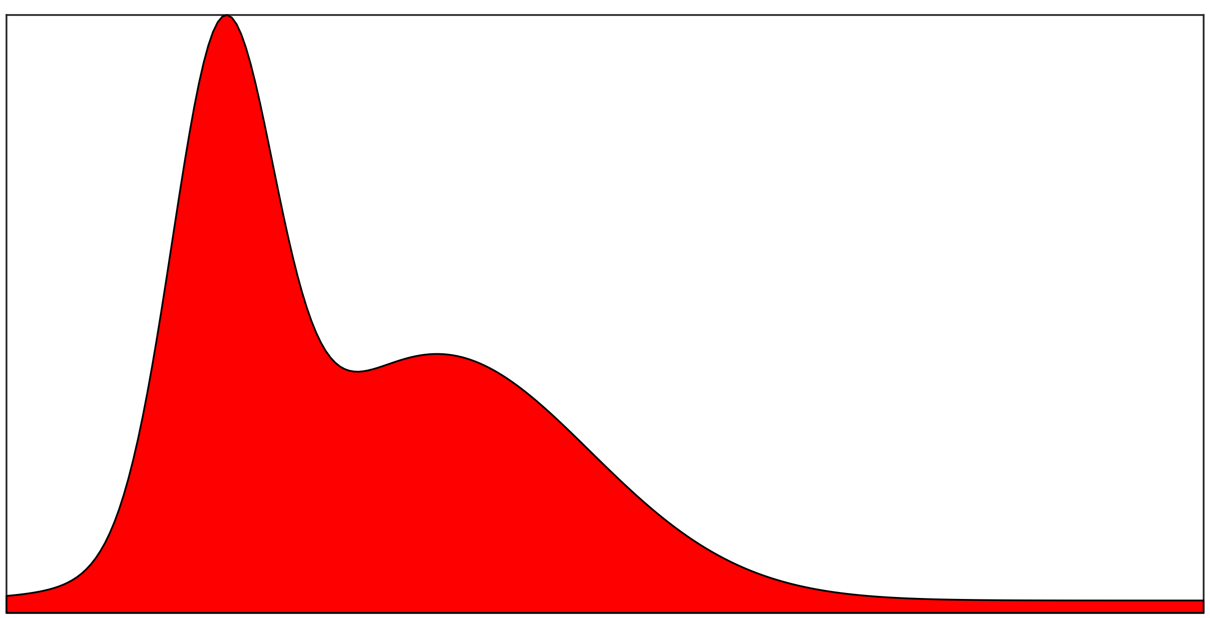
Remark. If $\Omega = \mathbb{R}$, $\boldsymbol{c}(x, y) = \boldsymbol{c}(|x - y|)$, $\boldsymbol{c}$ convex, $F_{\boldsymbol{\mu}}^{-1}, F_{\boldsymbol{\nu}}^{-1}$ quantile functions,

$$W(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int_0^1 \boldsymbol{c}(|F_{\boldsymbol{\mu}}^{-1}(x) - F_{\boldsymbol{\nu}}^{-1}(x)|)dx$$

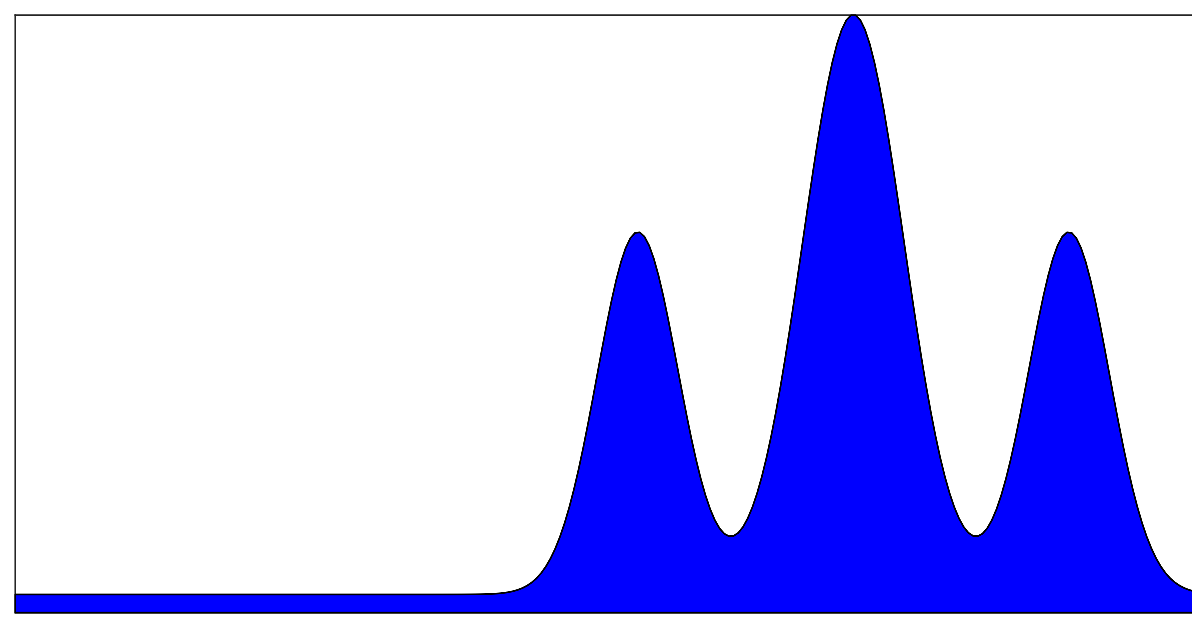
Easy (1): Univariate Measures

Remark. If $\Omega = \mathbb{R}$, $\boldsymbol{c}(x,y) = \boldsymbol{c}(|x-y|)$, $\boldsymbol{c}$ convex, $F_{\boldsymbol{\mu}}^{-1}, F_{\boldsymbol{\nu}}^{-1}$ quantile functions,

$$W(\boldsymbol{\mu},\boldsymbol{\nu}) = \int_0^1 \boldsymbol{c}(|F_{\boldsymbol{\mu}}^{-1}(x) - F_{\boldsymbol{\nu}}^{-1}(x)|)dx$$



$\boldsymbol{\mu}$

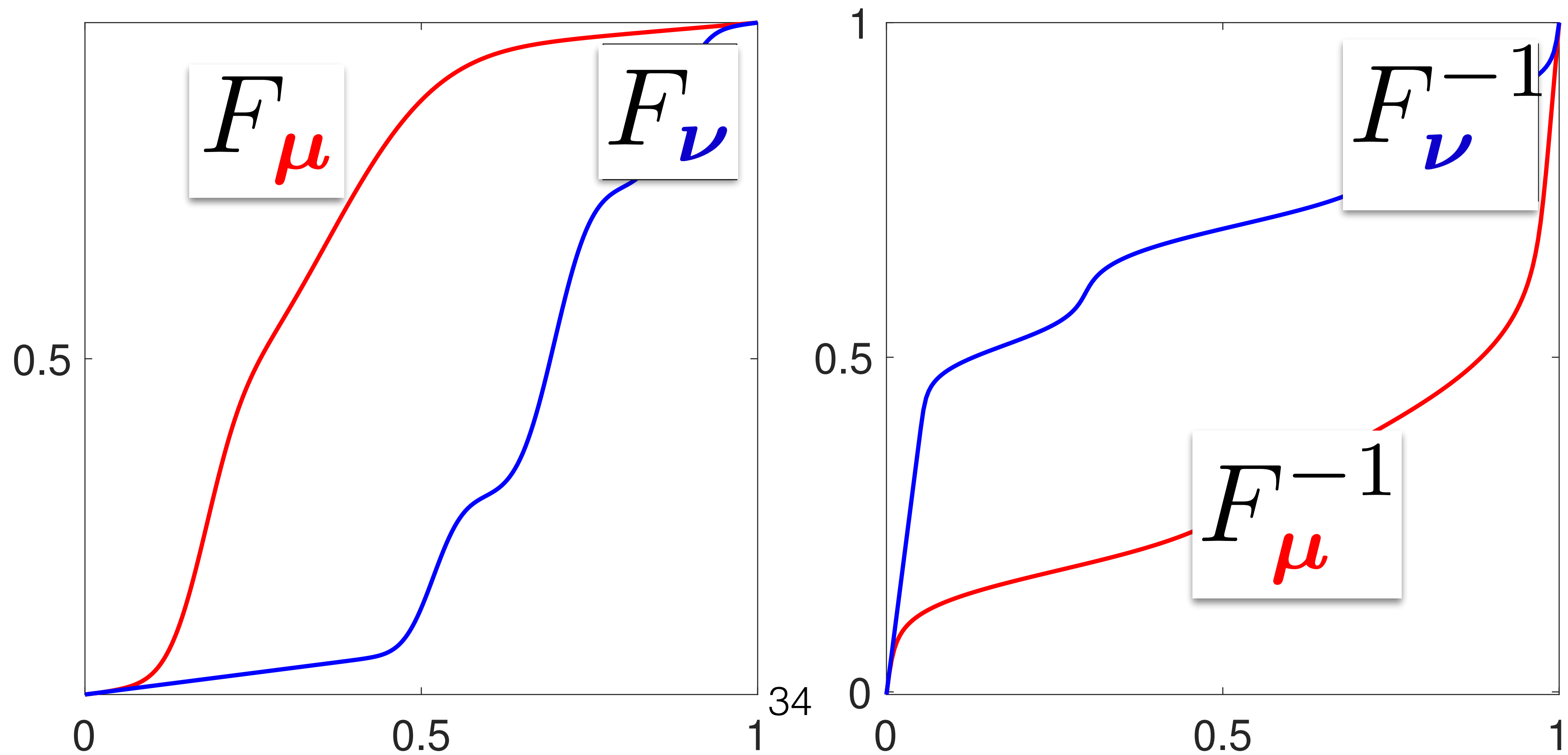


$\boldsymbol{\nu}$

Easy (1): Univariate Measures

Remark. If $\Omega = \mathbb{R}$, $c(x,y) = c(|x-y|)$, c convex, F_{μ}^{-1}, F_{ν}^{-1} quantile functions,

$$W(\mu,\nu) = \int_0^1 c(|F_{\mu}^{-1}(x) - F_{\nu}^{-1}(x)|)dx$$



Easy (2): Gaussian Measures

Remark. If $\Omega = \mathbb{R}^d$, $\boldsymbol{c}(x,y) = \|x-y\|^2$, and $\boldsymbol{\mu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}), \boldsymbol{\nu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\nu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})$ then

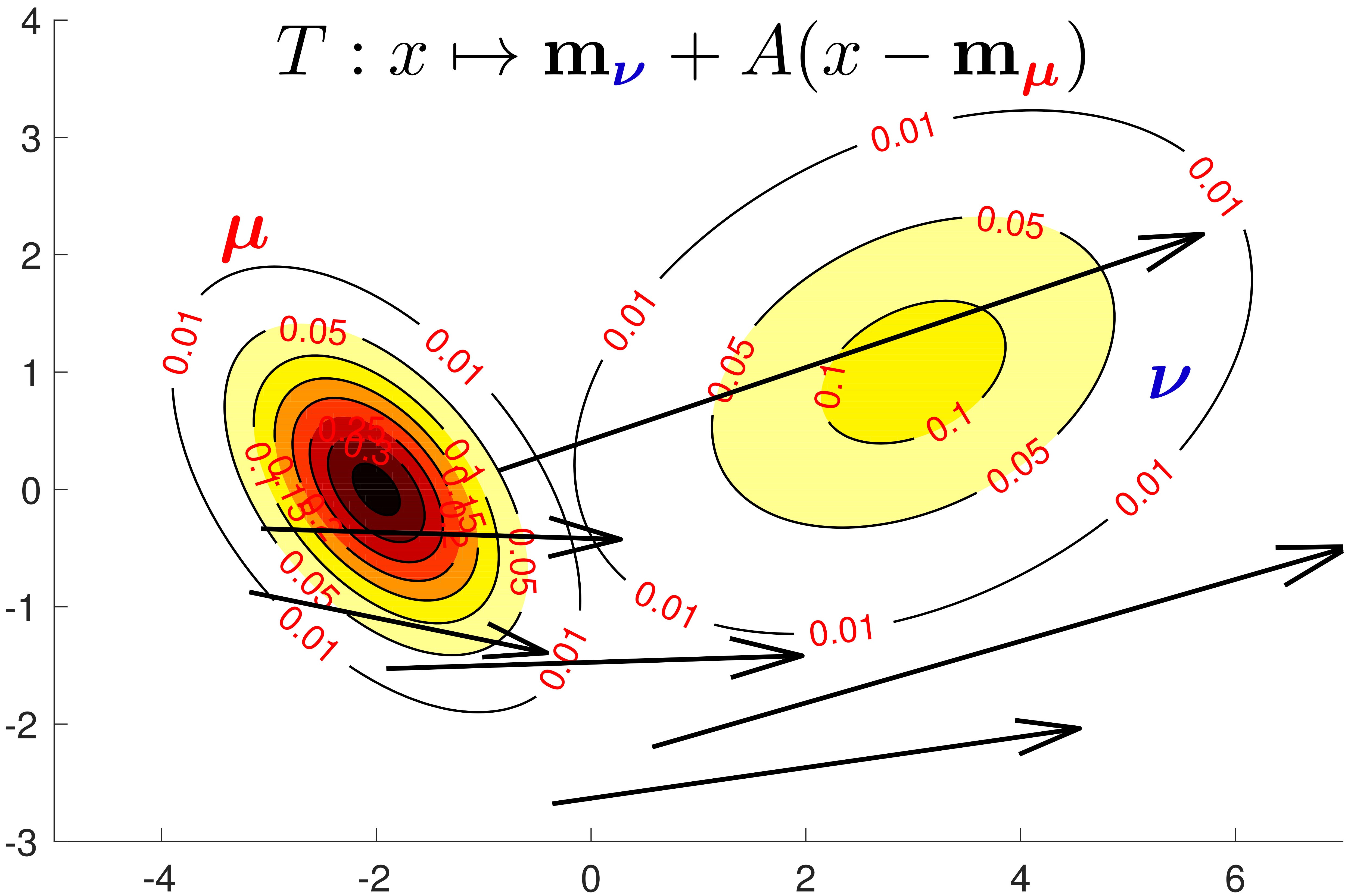
$$W_2^2(\boldsymbol{\mu},\boldsymbol{\nu}) = \|\mathbf{m}_{\boldsymbol{\mu}} - \mathbf{m}_{\boldsymbol{\nu}}\|^2 + B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2$$

where B is the Bures metric

$$B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2 = \mathrm{trace}(\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_{\boldsymbol{\nu}} - 2(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\nu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2})^{1/2}).$$

The map $T : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$ is **optimal**, where $A = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}}\right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}}$.

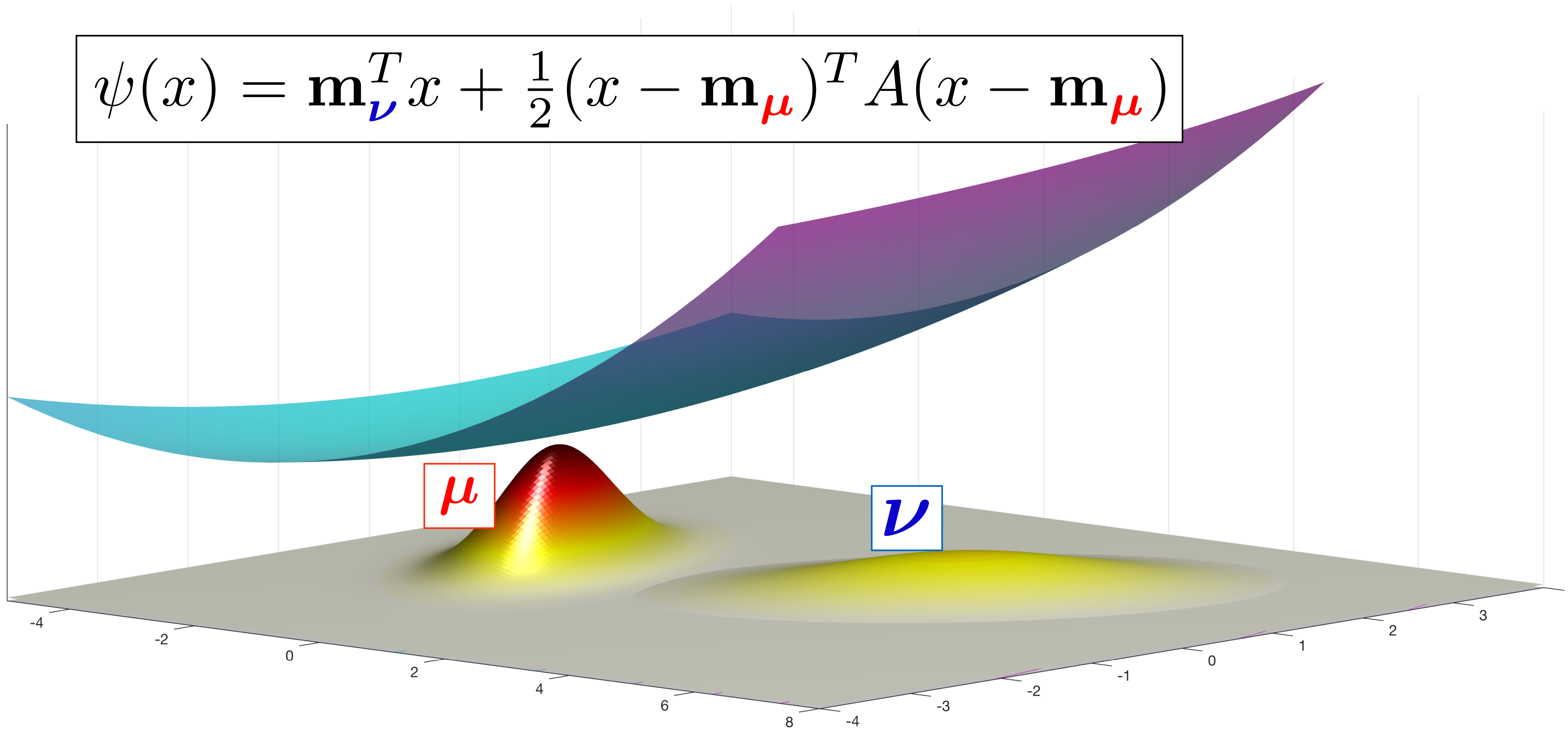
Easy (2): Gaussian Measures



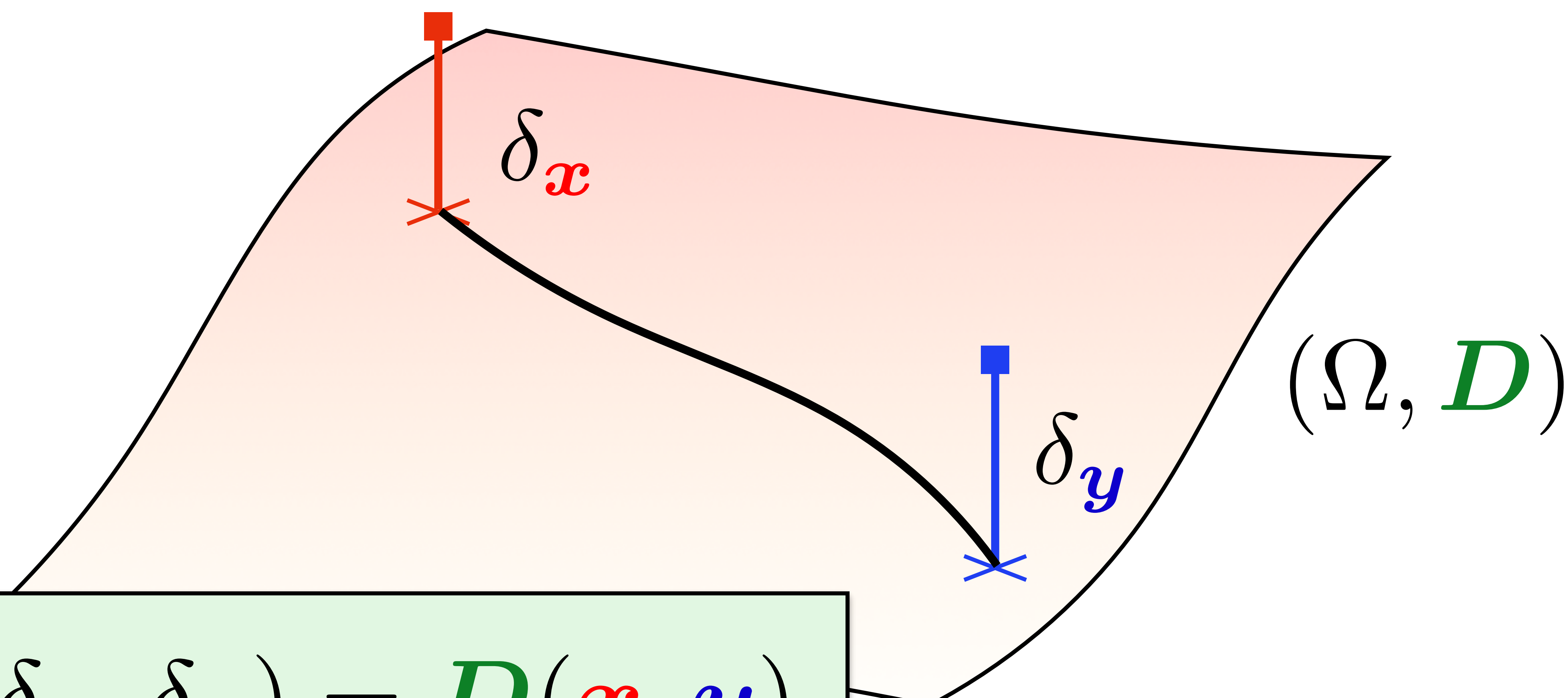
Easy (2): Gaussian Measures

$$T = \nabla\psi : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$$

$$\boxed{\psi(x) = \mathbf{m}_{\boldsymbol{\nu}}^T x + \frac{1}{2}(x - \mathbf{m}_{\boldsymbol{\mu}})^T A (x - \mathbf{m}_{\boldsymbol{\mu}})}$$

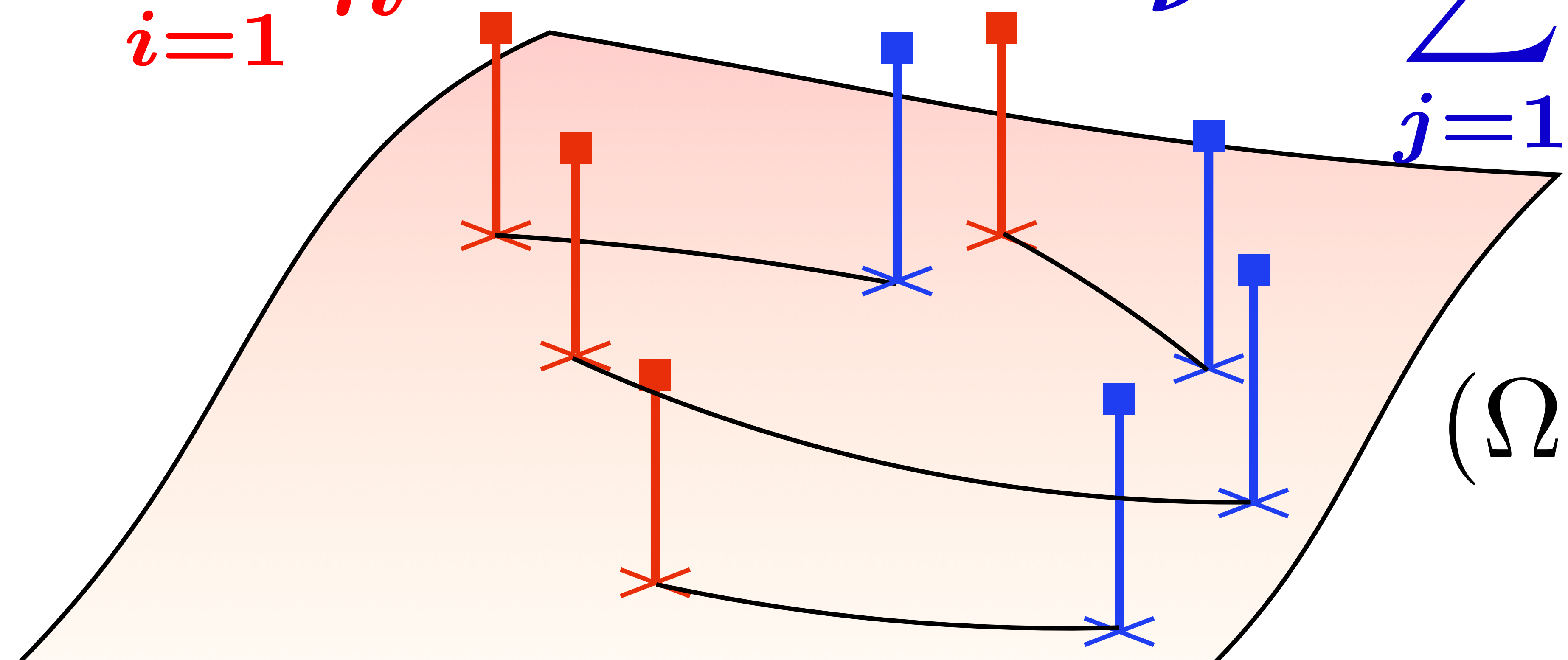


Wasserstein Between Two Diracs



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

Linear Assignment ⊂ Wasserstein



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

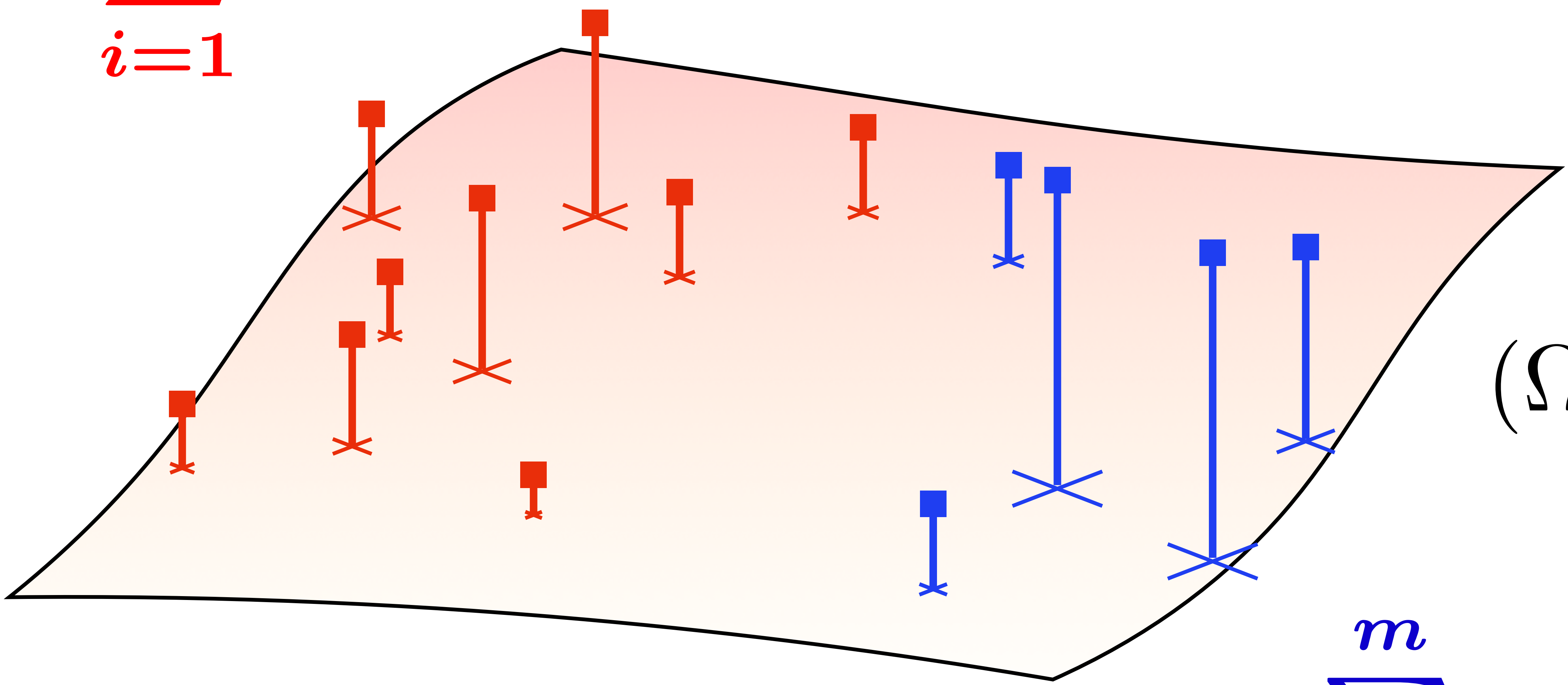
$$\nu = \sum_{j=1}^{n} \frac{1}{n} \delta_{y_j}$$

$$(\Omega, D)$$

$$W_p^p(\mu, \nu) = \min_{\sigma \in S_n} \frac{1}{n} \sum_{i=1}^{n} D(x_i, y_{\sigma_i})^p$$

OT on Two Empirical Measures

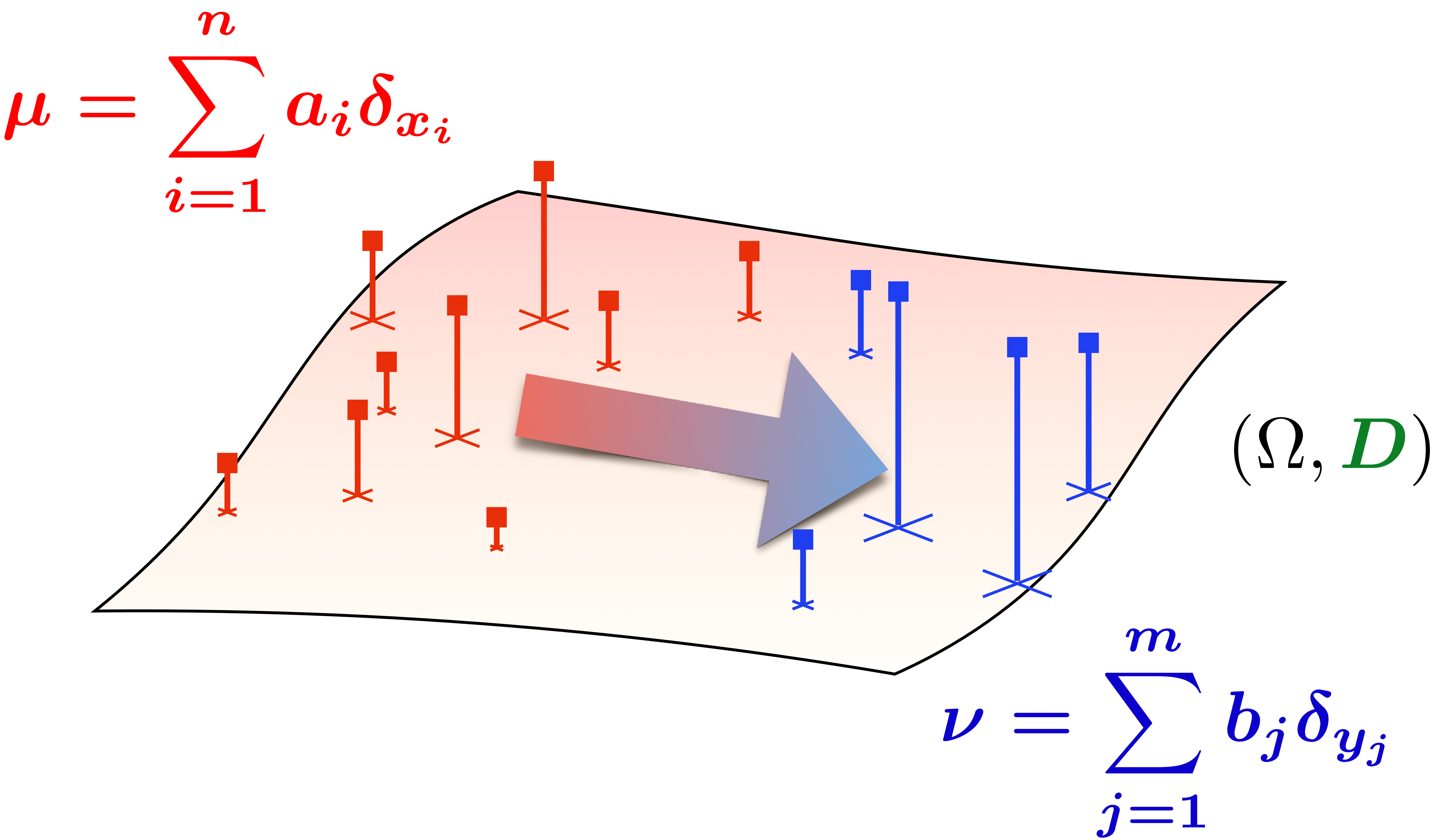
$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



$$(\Omega, D)$$

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

OT on Two Empirical Measures



Wasserstein on Empirical Measures

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

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

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

Def. Optimal Transport Problem

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

Dual Kantorovich Problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} \\ \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

Dual Kantorovich Problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} \\ \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}}} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle$$

Def. Dual OT problem

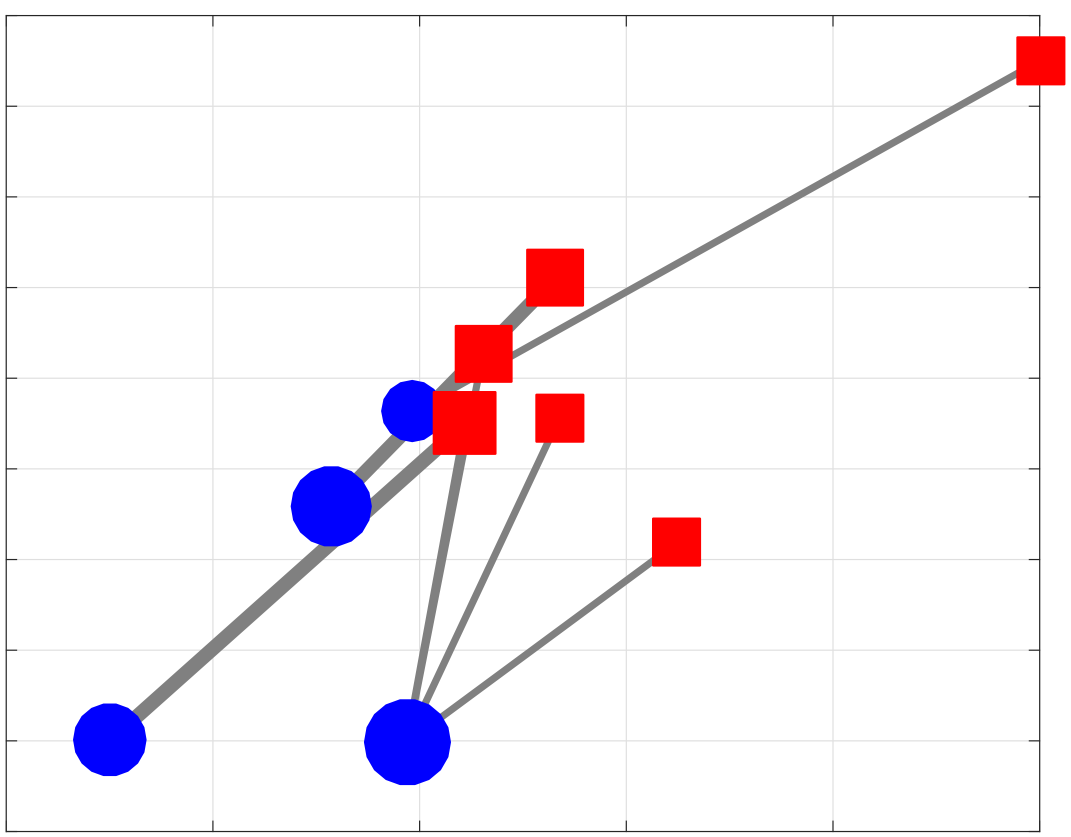
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^m \\ \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j \leq D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p}} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b}$$

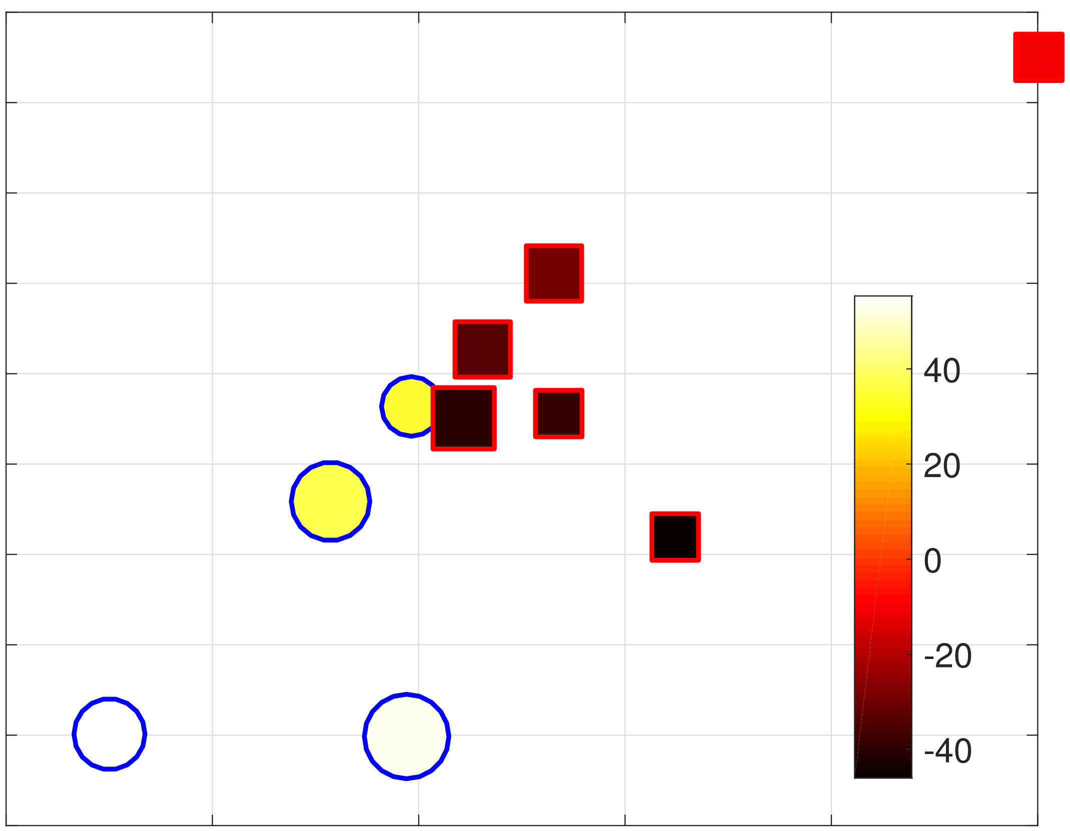
Dual Kantorovich Problem

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}\\ \boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

Def. Dual OT problem

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



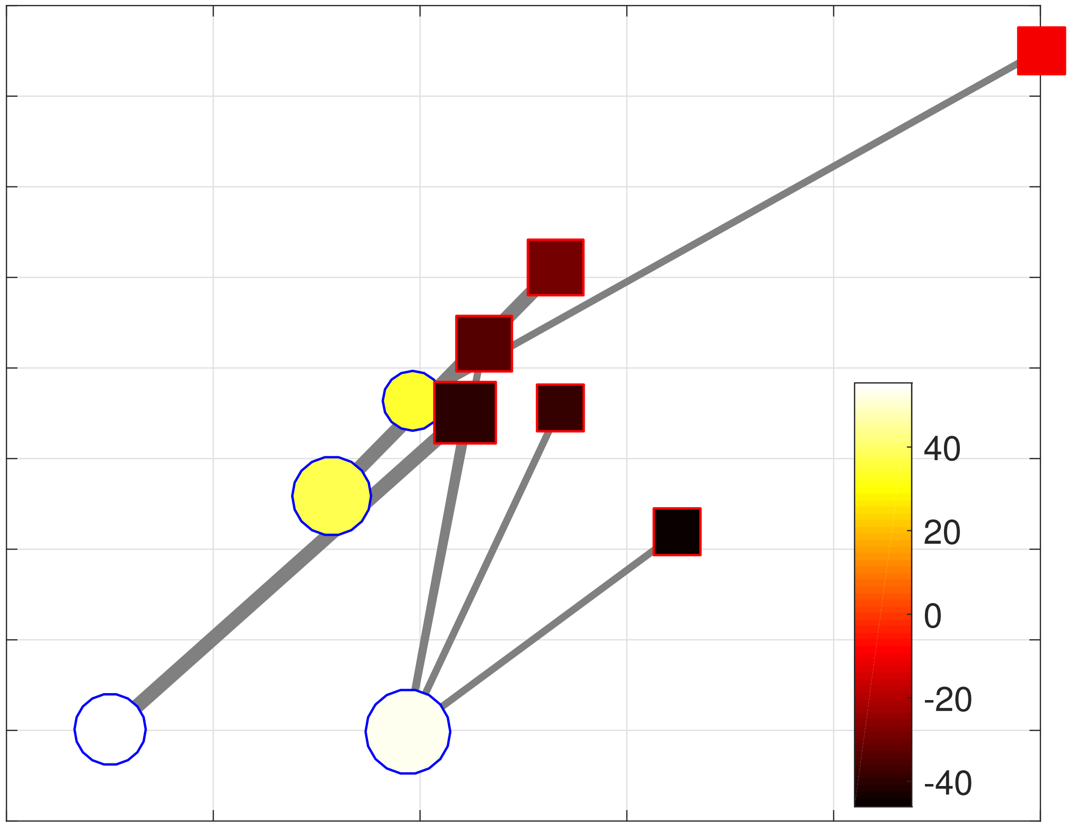


Dual Kantorovich Problem

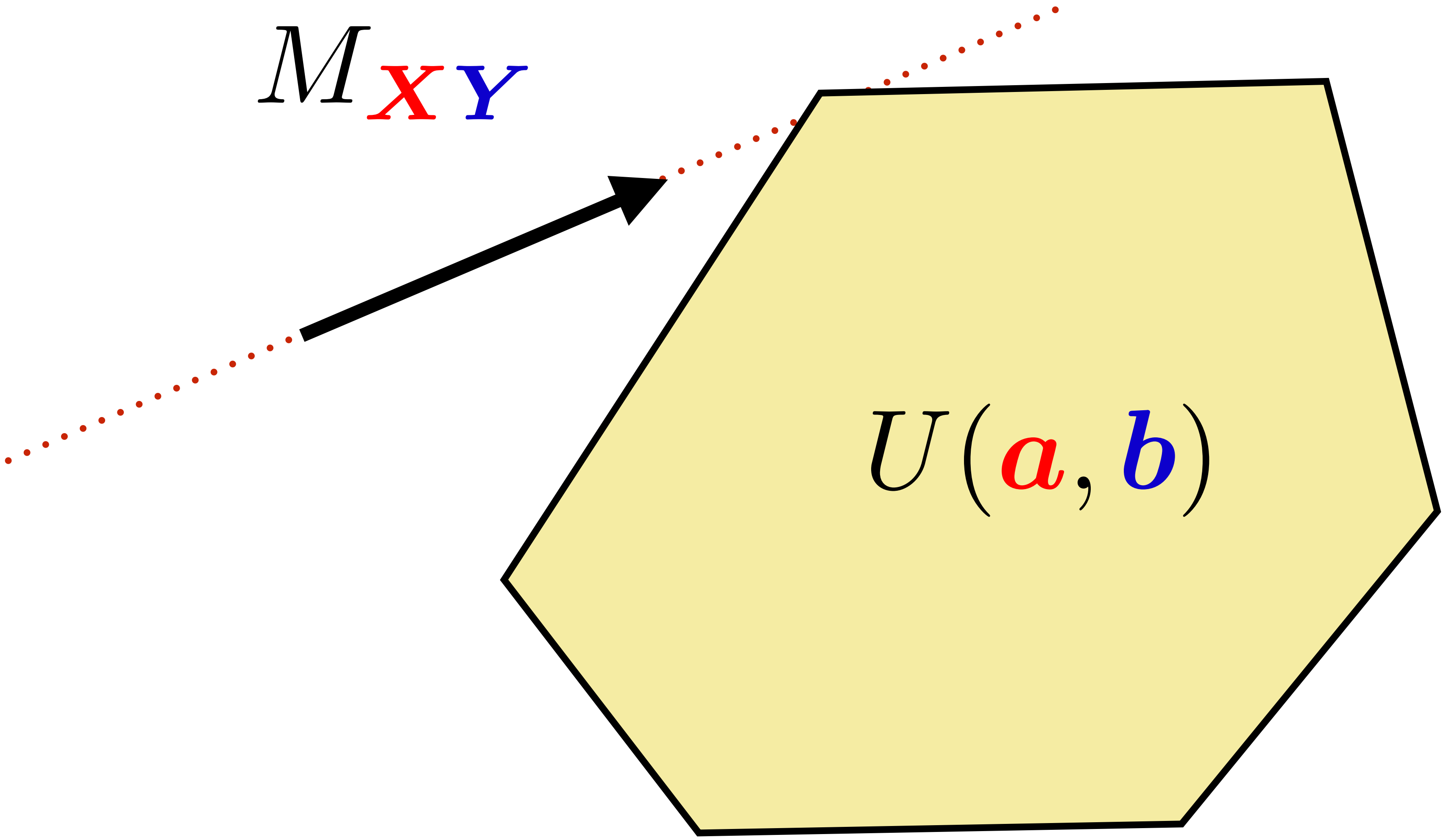
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} \\ \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

Def. Dual OT problem

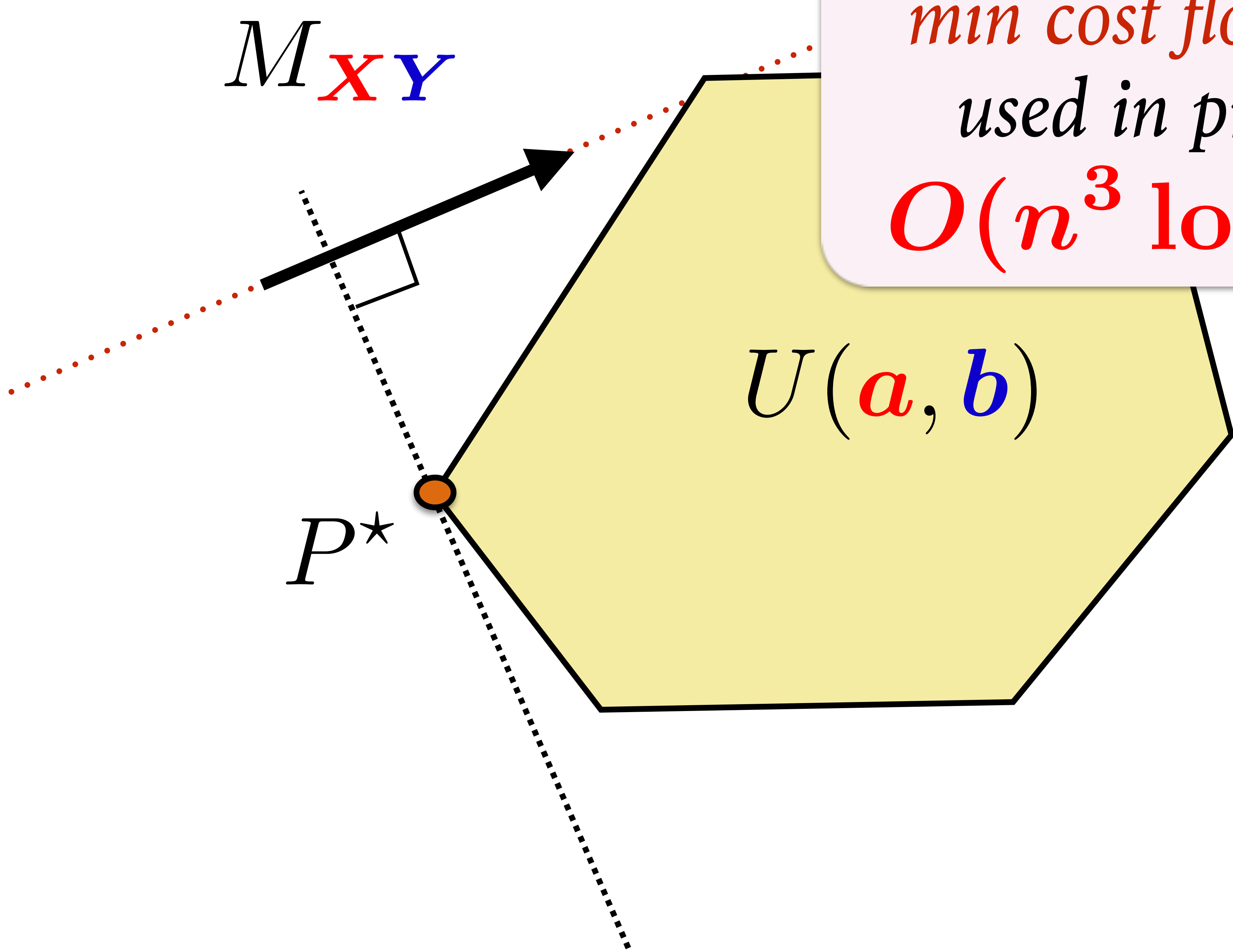
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^m \\ \boldsymbol{\alpha}_{\boldsymbol{i}} + \boldsymbol{\beta}_{\boldsymbol{j}} \leq D(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{j}})^p}} \alpha^T \boldsymbol{a}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



Solving the OT Problem



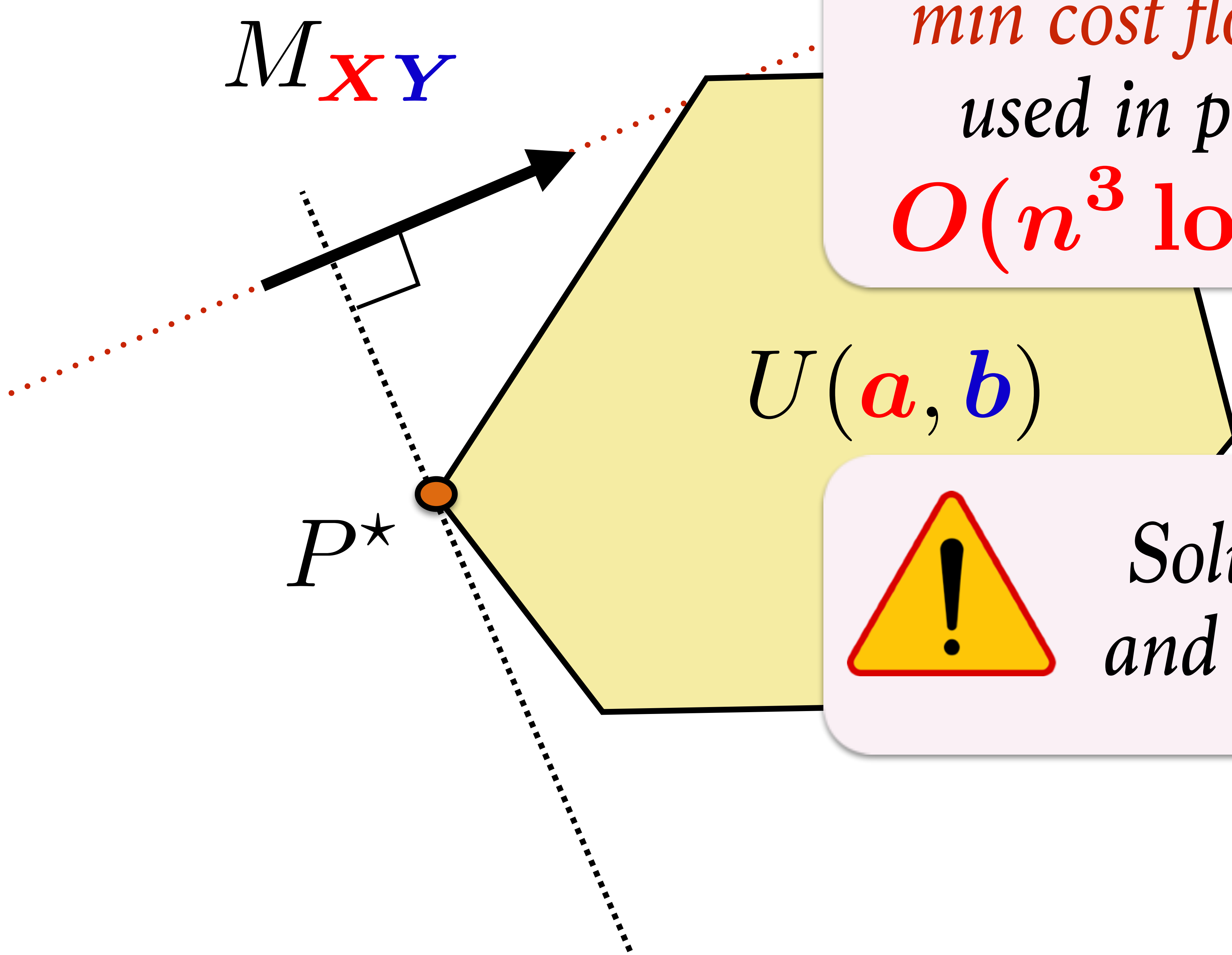
Solving the OT Problem



min cost flow solver used in practice.

$\boldsymbol{O(n^3 \log(n))}$

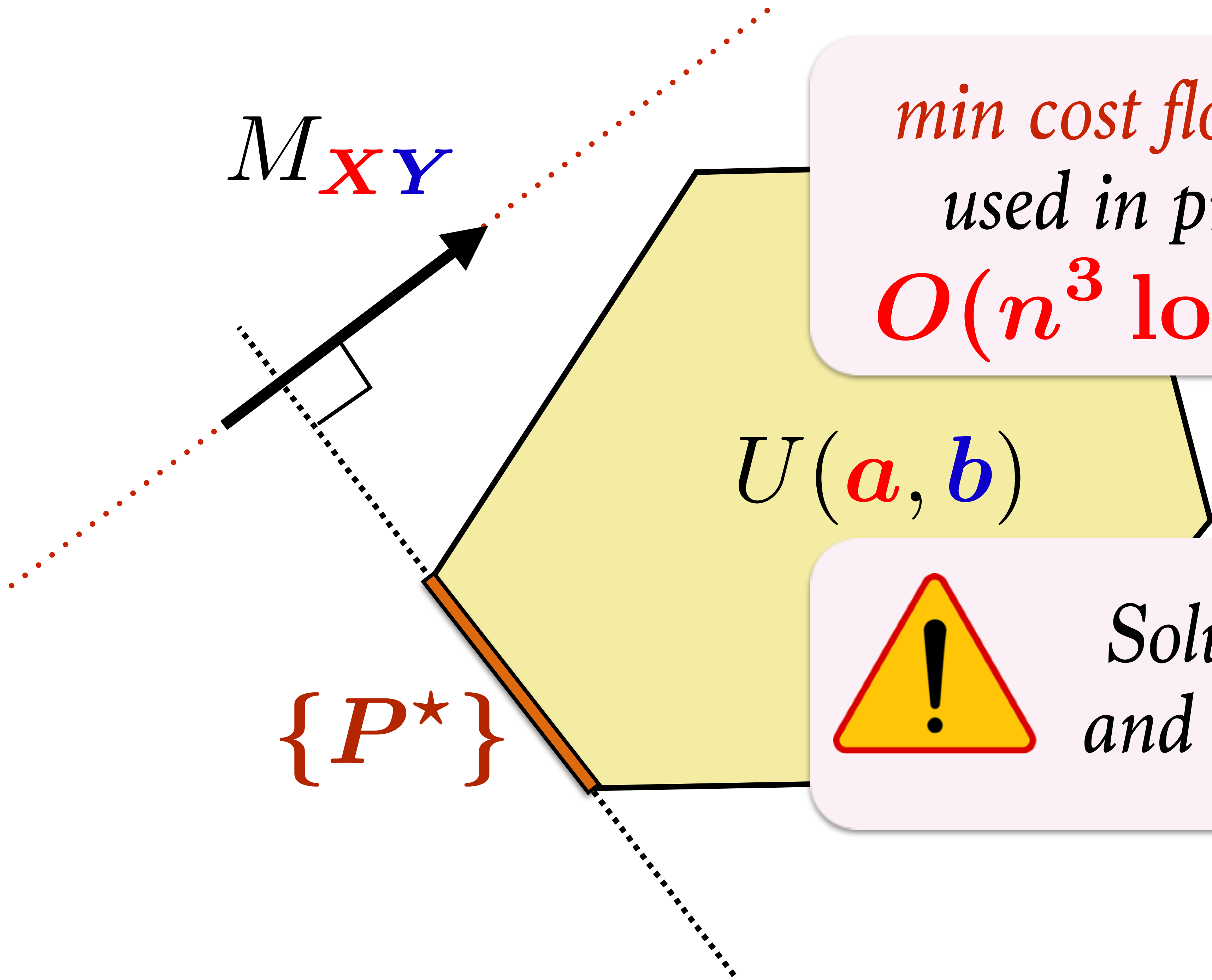
Solving the OT Problem



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

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

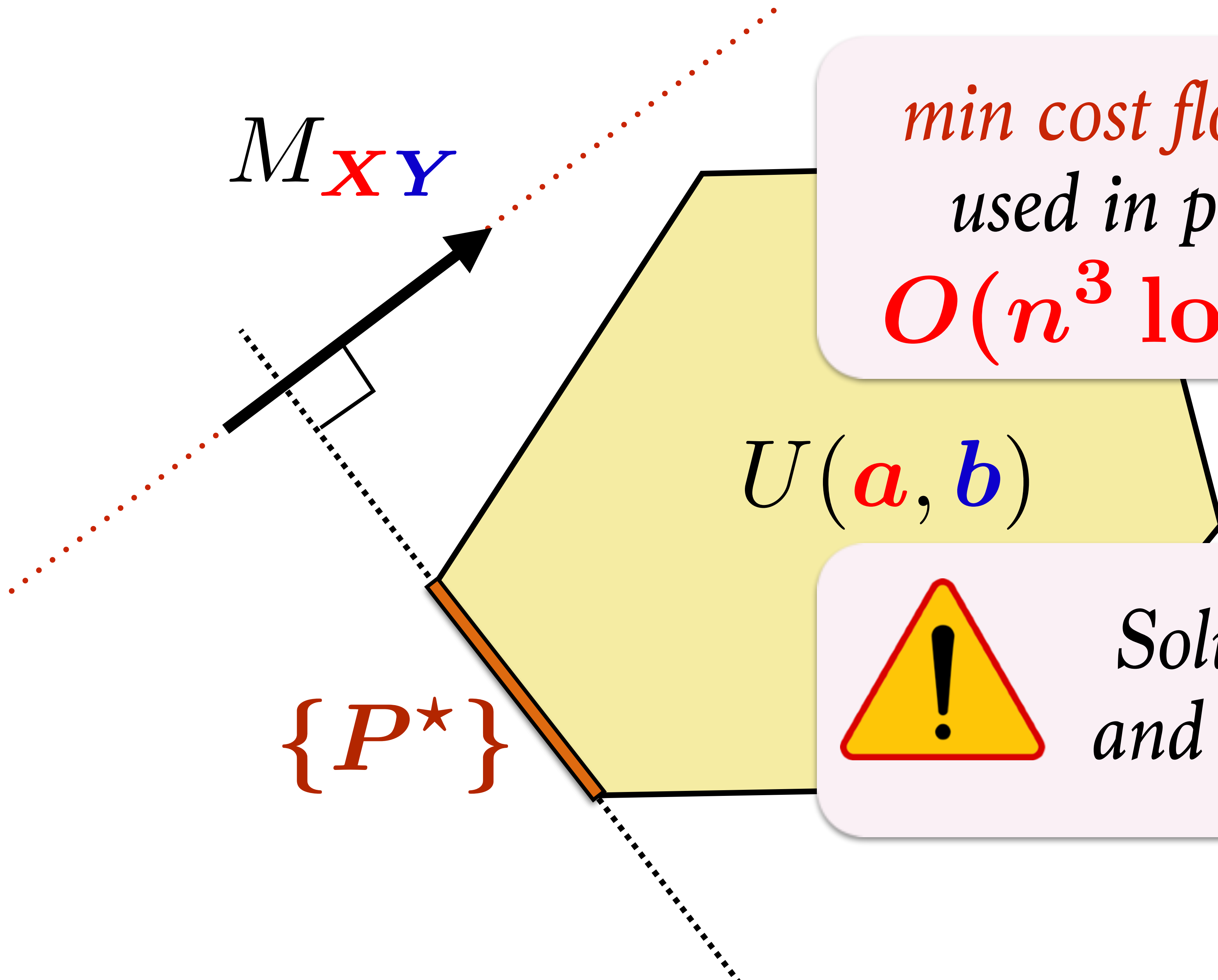
Solving the OT Problem



min cost flow solver *used in practice.*
$O(n^3 \log(n))$

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

Solving the OT Problem

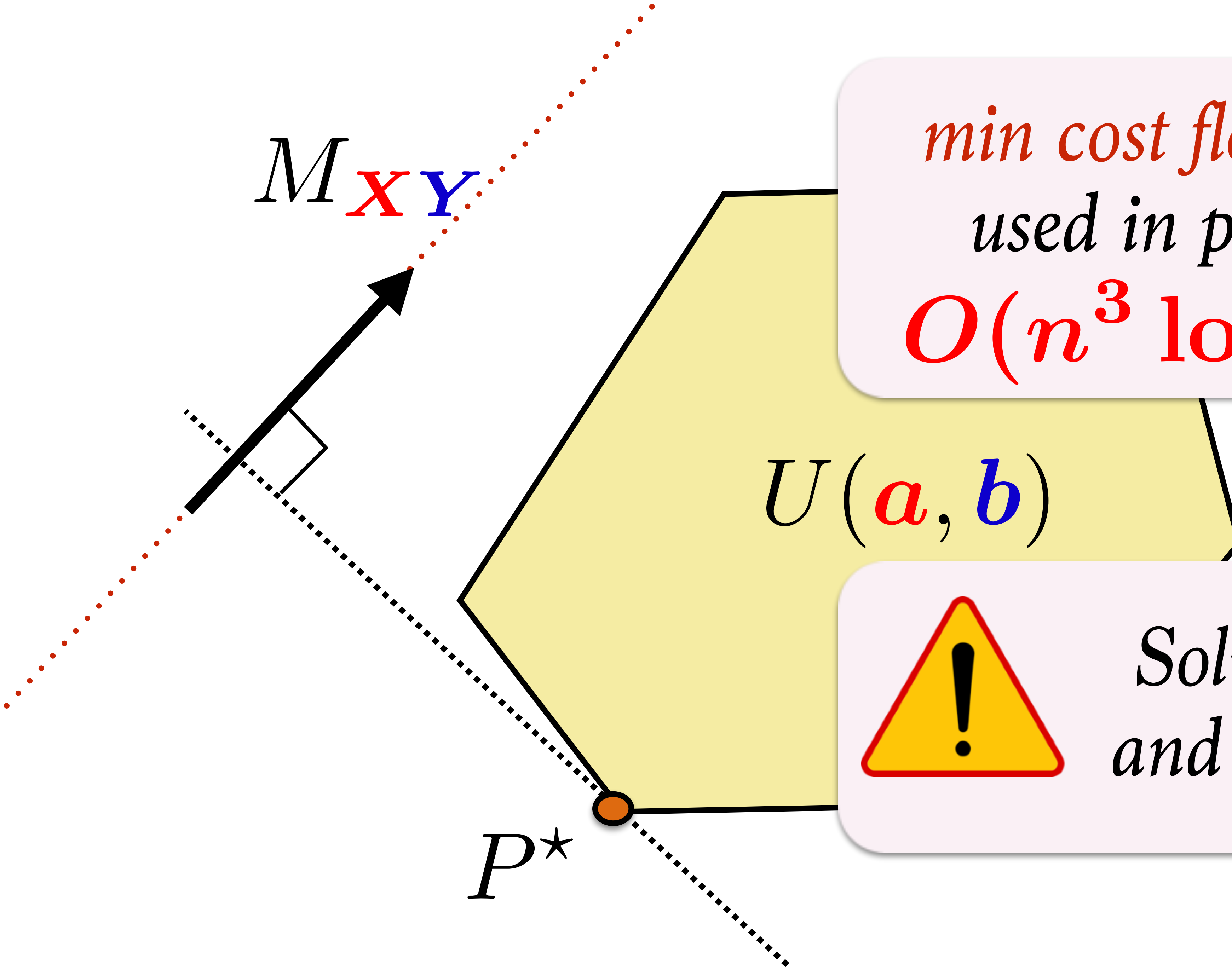


min cost flow solver used in practice.

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

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

Solving the OT Problem

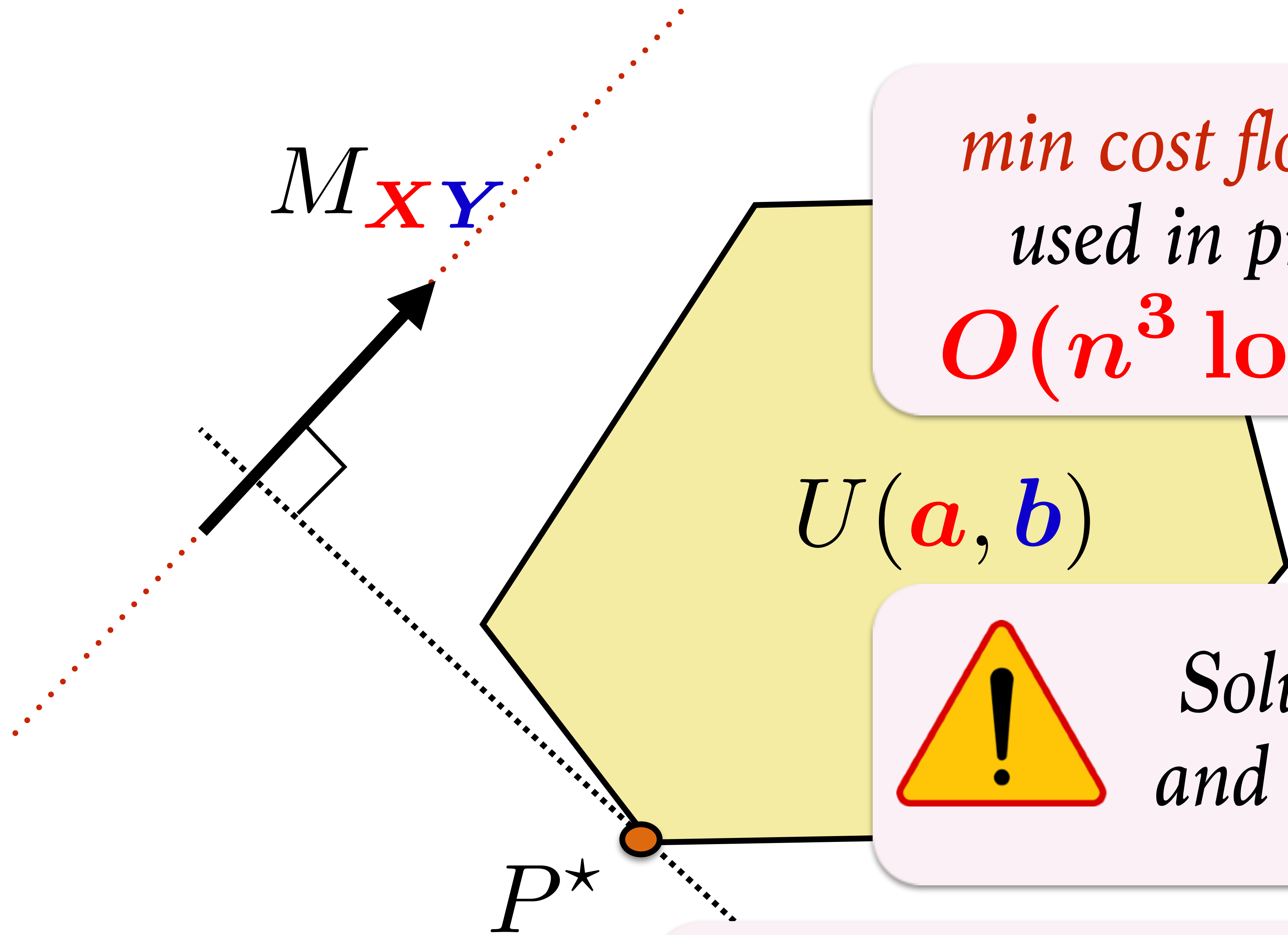


min cost flow solver *used in practice.*

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

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

Solving the OT Problem

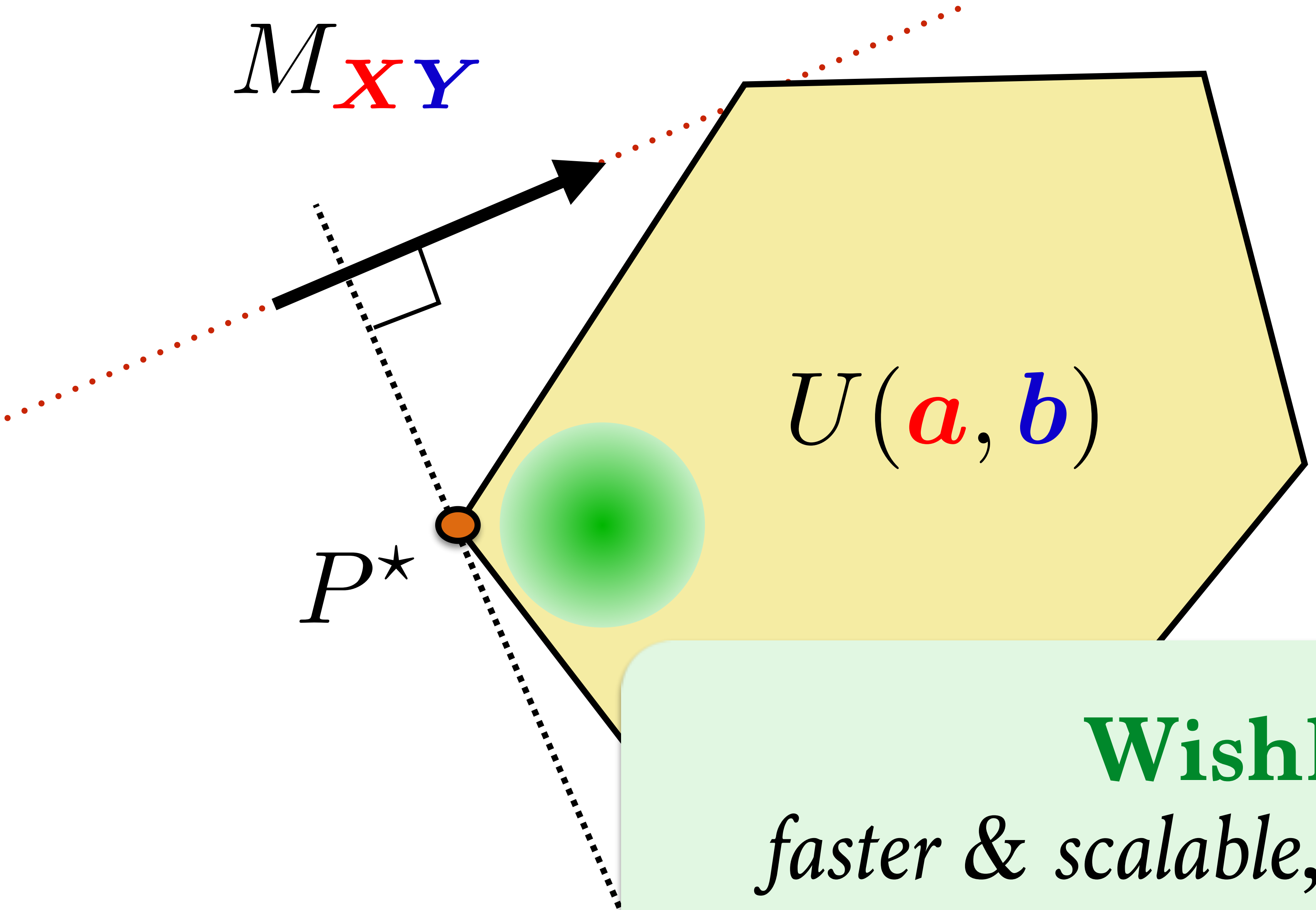


min cost flow solver *used in practice.* $O(n^3 \log(n))$

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

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

Solution: Regularization



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

Entropic Regularization [**Wilson'62**]

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

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

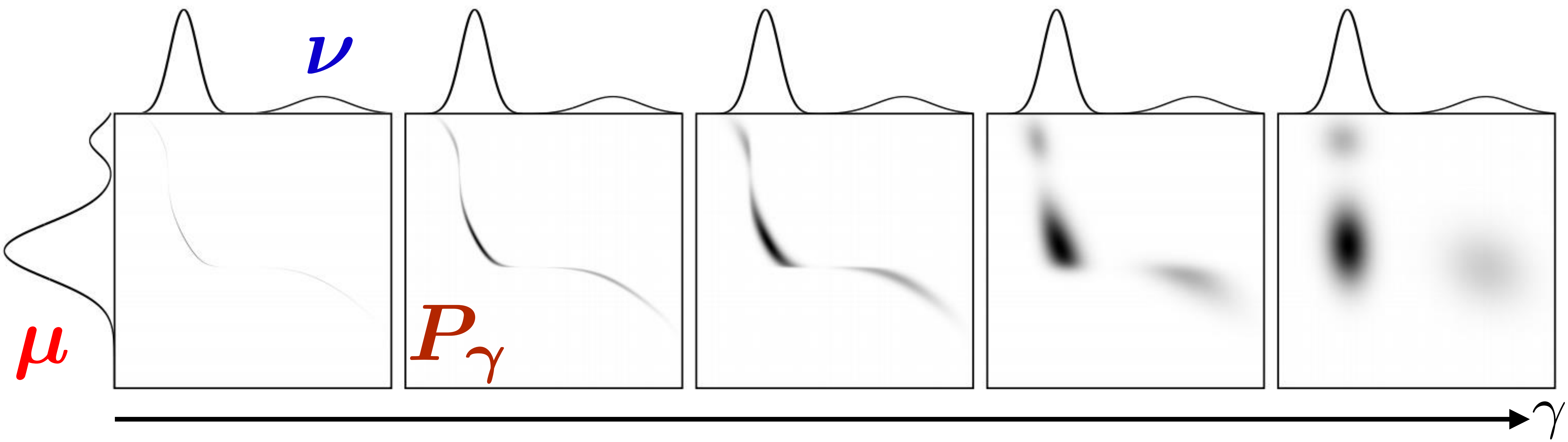
$$E(P) \stackrel{\text{def}}{=} -\sum_{i,j=1}^{nm} P_{ij}(\log P_{ij} - 1)$$

Note: Unique optimal solution because of strong concavity of entropy

Entropic Regularization [**Wilson'62**]

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

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



Note: Unique optimal solution because of strong concavity of entropy

Fast & Scalable Algorithm

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

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

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

Fast & Scalable Algorithm

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

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

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

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

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

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

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

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

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}) \mathbf{1}_m & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \mathbf{diag}(\boldsymbol{u}) \mathbf{1}_n & = \boldsymbol{b} \end{cases}$$

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

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

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}) \mathbf{1}_m & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \underbrace{\mathbf{diag}(\boldsymbol{u}) \mathbf{1}_n}_{\boldsymbol{u}} & = \boldsymbol{b} \end{cases}$$

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

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

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \overbrace{\mathbf{diag}(\boldsymbol{v}) \mathbf{1}_m}^{\boldsymbol{v}} & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \underbrace{\mathbf{diag}(\boldsymbol{u}) \mathbf{1}_n}_{\boldsymbol{u}} & = \boldsymbol{b} \end{cases}$$

Fast & Scalable Algorithm

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

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

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

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \boldsymbol{v} & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \boldsymbol{u} & = \boldsymbol{b} \end{cases}$$

Fast & Scalable Algorithm

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle - \gamma E(\boldsymbol{P})$
then $\exists! \boldsymbol{u} \in \mathbb{R}^n_+, \boldsymbol{v} \in \mathbb{R}^m_+$, such that

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

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \boldsymbol{u} \odot K\boldsymbol{v} & = \boldsymbol{a} \\ \boldsymbol{v} \odot K^T \boldsymbol{u} & = \boldsymbol{b} \end{cases}$$

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

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

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \boldsymbol{u} = \boldsymbol{a}/K\boldsymbol{v} \\ \boldsymbol{v} = \boldsymbol{b}/K^T\boldsymbol{u} \end{cases}$$

Fast & Scalable Algorithm

Sinkhorn's Algorithm : Repeat

1. $\boldsymbol{u} = \boldsymbol{a} / K\boldsymbol{v}$
2. $\boldsymbol{v} = \boldsymbol{b} / K^T\boldsymbol{u}$

Fast & Scalable Algorithm

Sinkhorn's Algorithm : Repeat

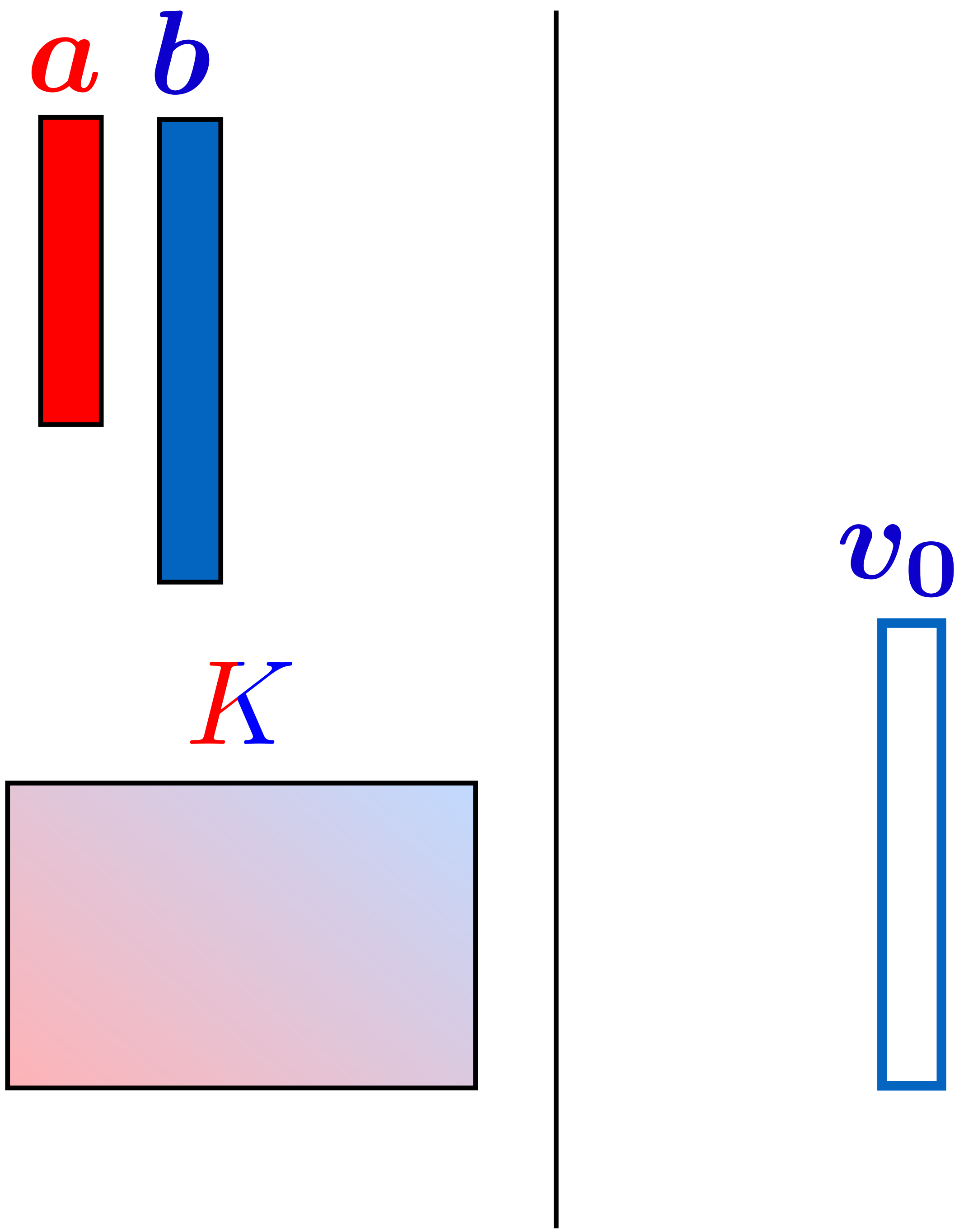
1. $\boldsymbol{u} = \boldsymbol{a} / K\boldsymbol{v}$
2. $\boldsymbol{v} = \boldsymbol{b} / K^T\boldsymbol{u}$

- **[Sinkhorn'64]** proved convergence for the first time.
- **[Lorenz'89]** linear convergence, see **[Altschuler'17]**
- $O(nm)$ complexity, GPGPU parallel **[Cuturi'13]** .
- $O(n \log n)$ on gridded spaces using convolutions. **[Solomon'15]**

Fast & Scalable Algorithm

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

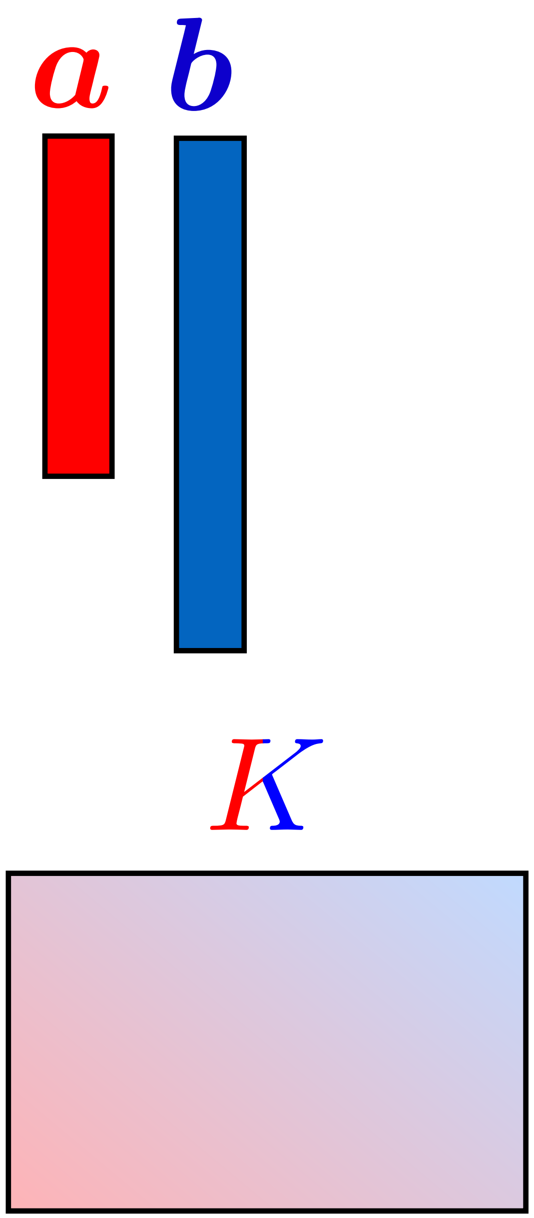
$$\boldsymbol{u} \leftarrow \boldsymbol{a} / K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b} / K^T\boldsymbol{u}$$

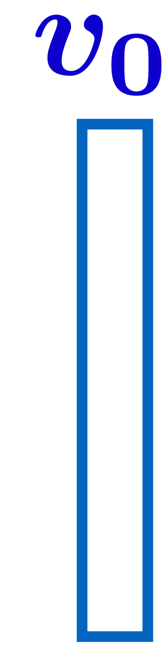


Fast & Scalable Algorithm

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

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

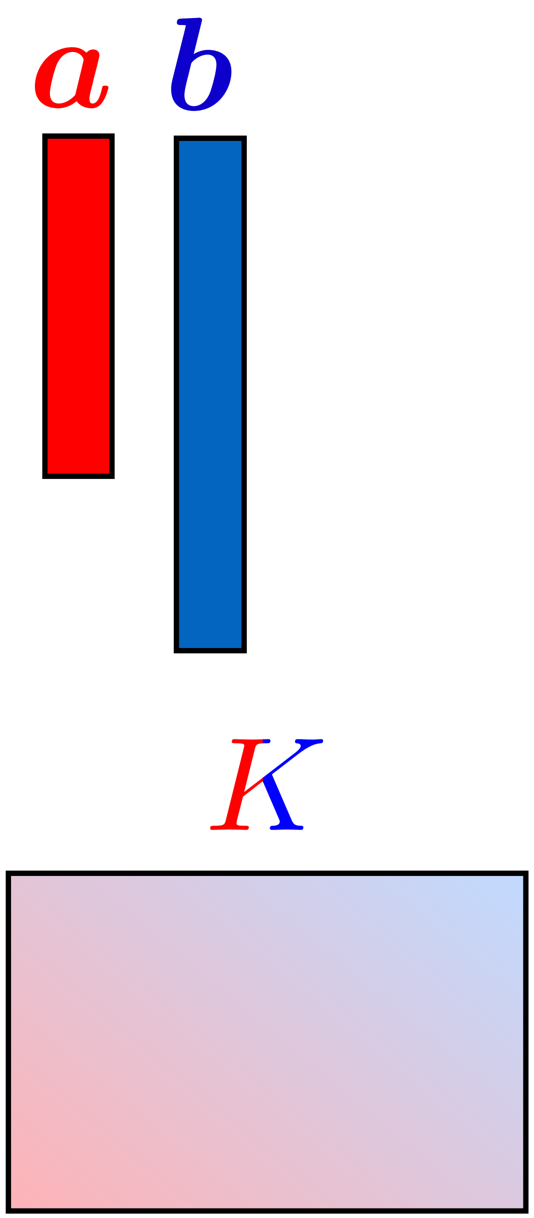


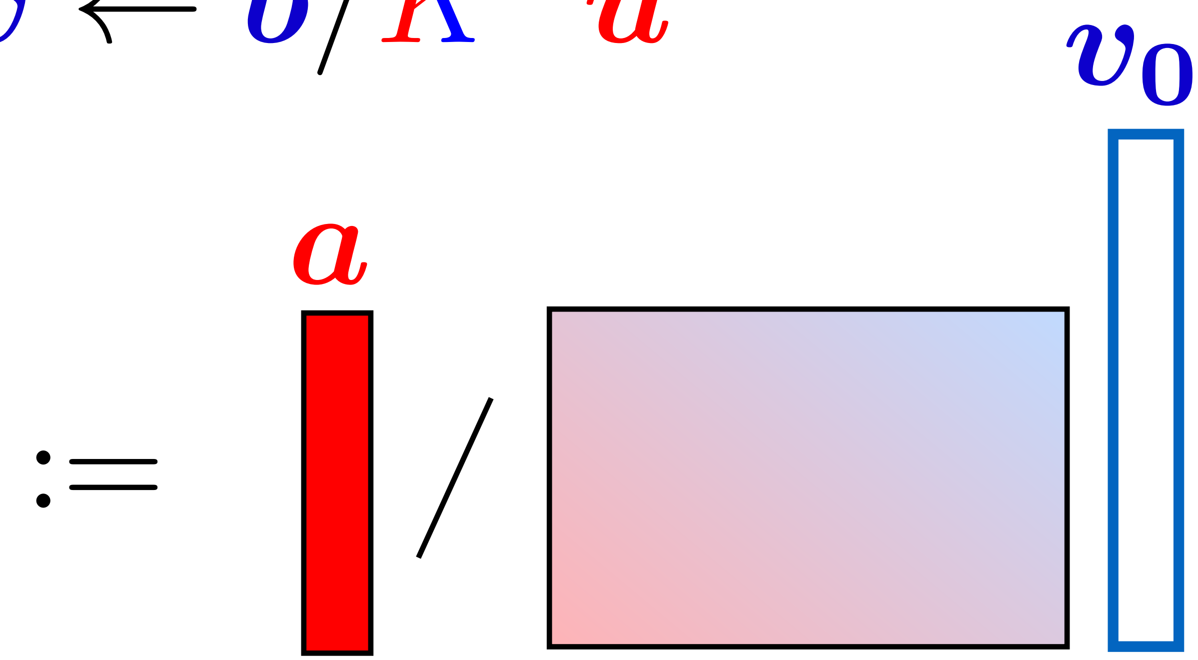


Fast & Scalable Algorithm

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

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

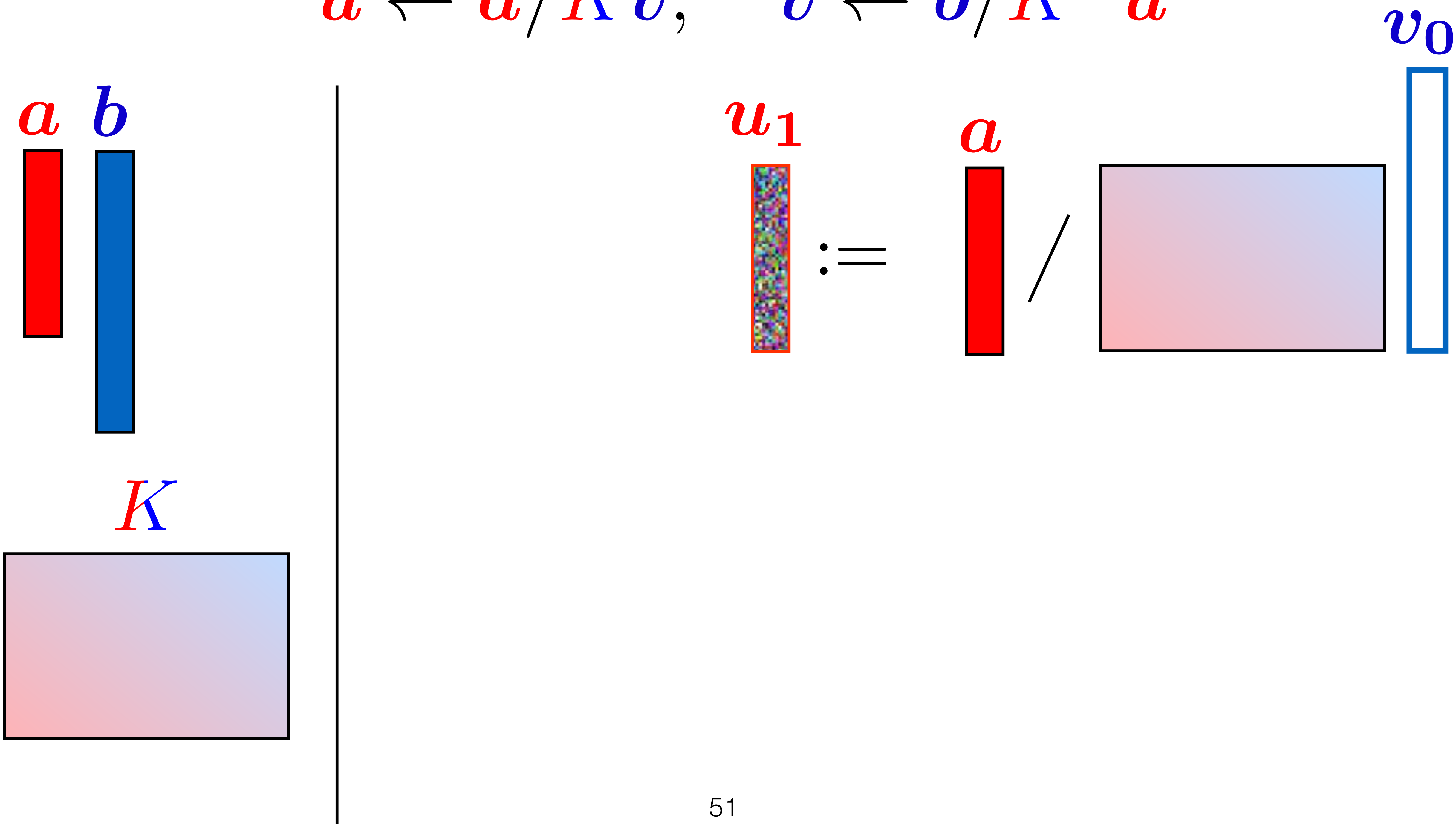




Fast & Scalable Algorithm

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

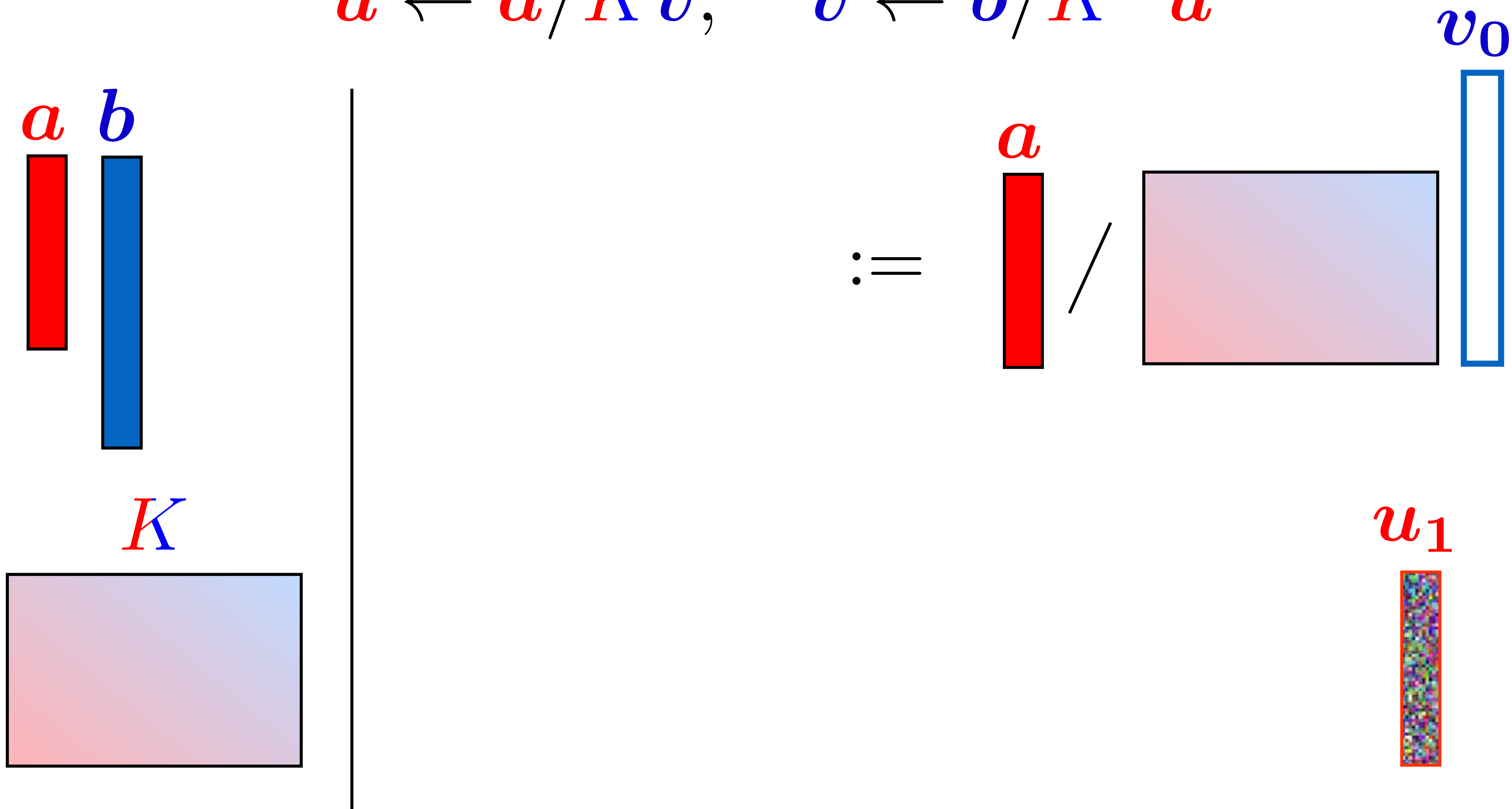
$$\boldsymbol{u} \leftarrow \boldsymbol{a} / K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b} / K^T \boldsymbol{u}$$



Fast & Scalable Algorithm

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

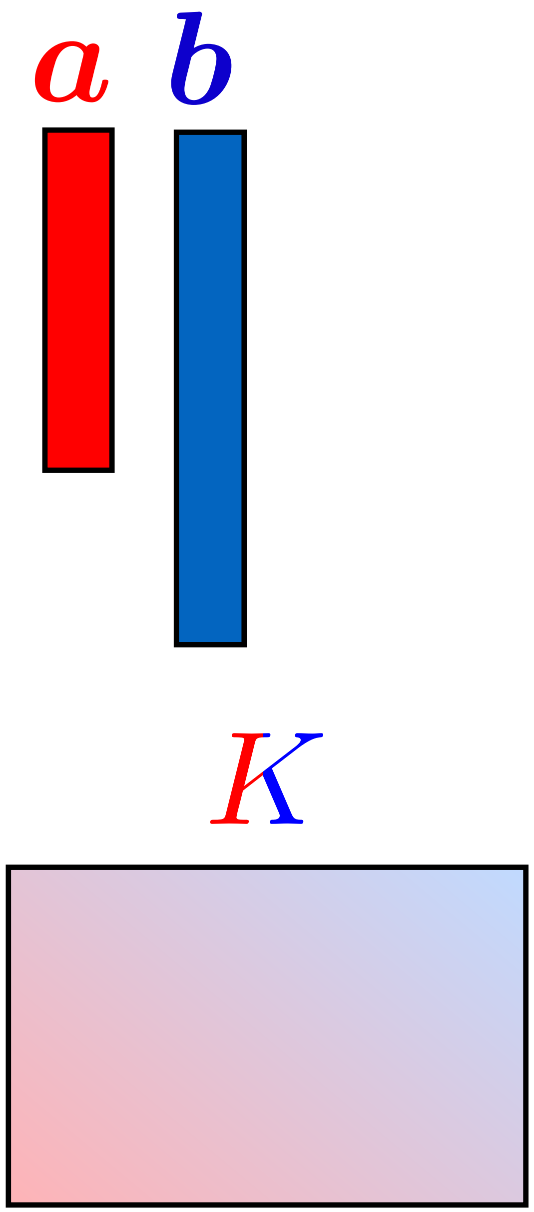
$$\mathbf{u} \leftarrow \mathbf{a}/K\mathbf{v}, \quad \mathbf{v} \leftarrow \mathbf{b}/K^T\mathbf{u}$$

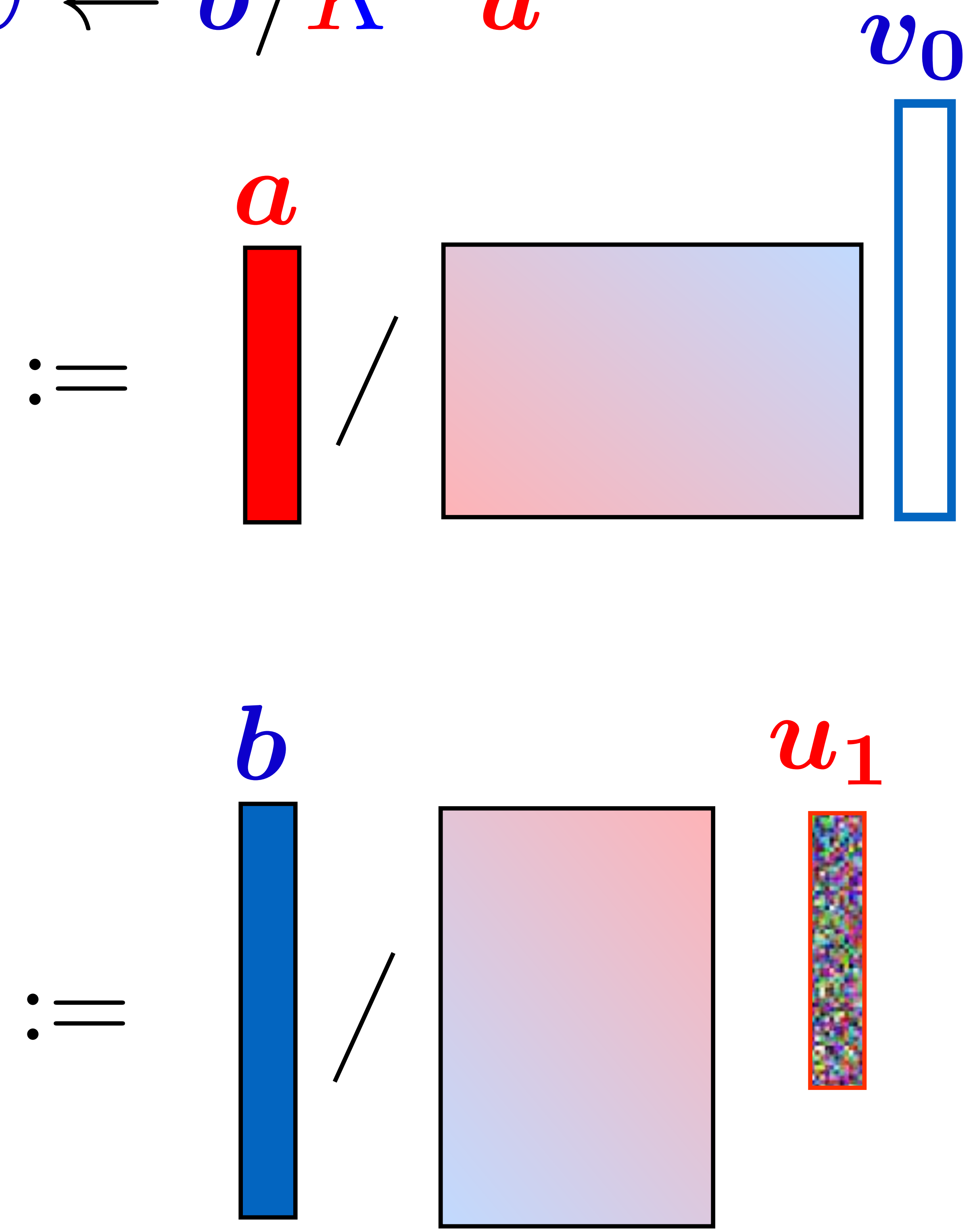


Fast & Scalable Algorithm

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

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

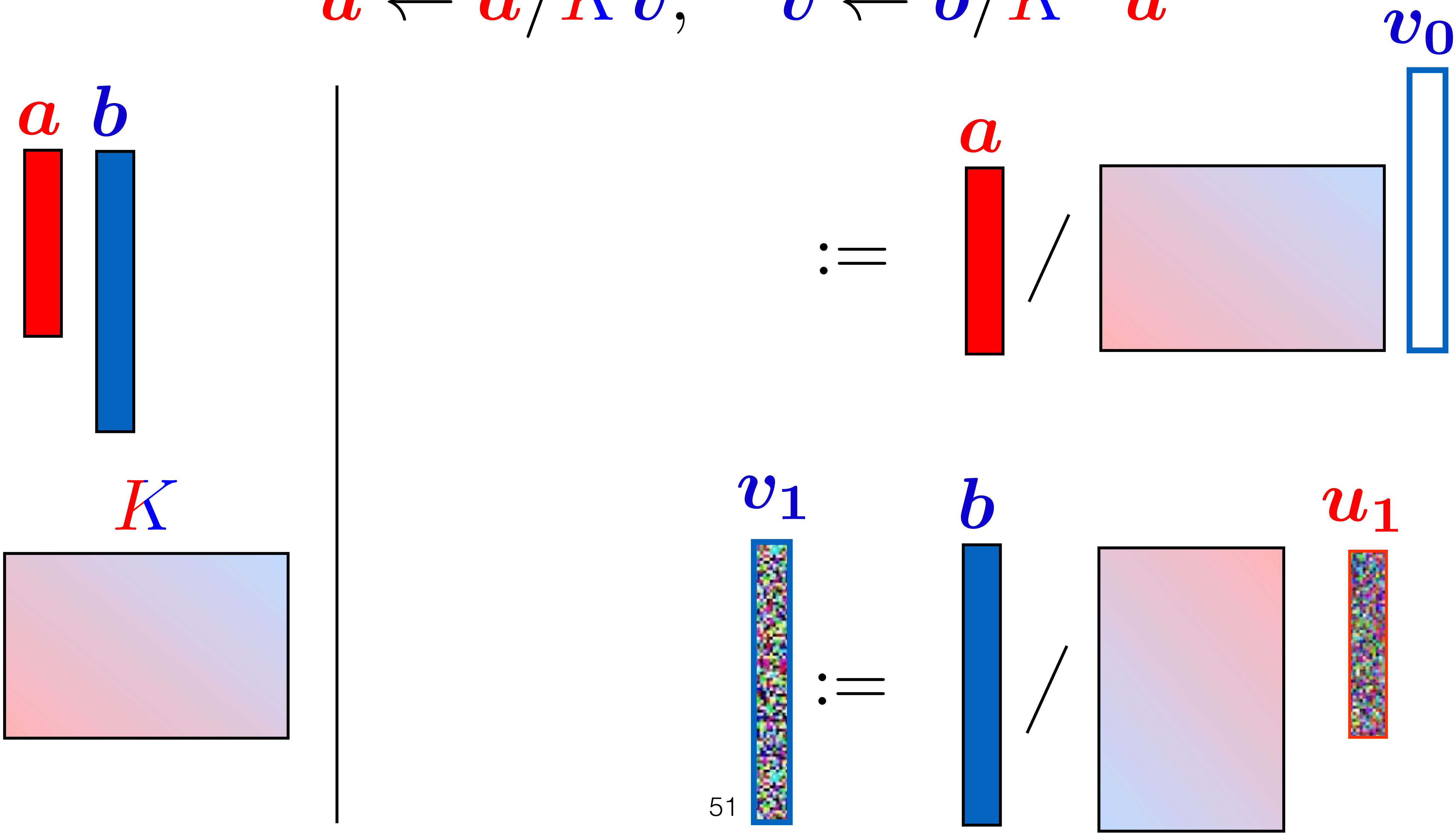




Fast & Scalable Algorithm

- **[Sinkhorn'64]** fixed-point iterations for (u, v)

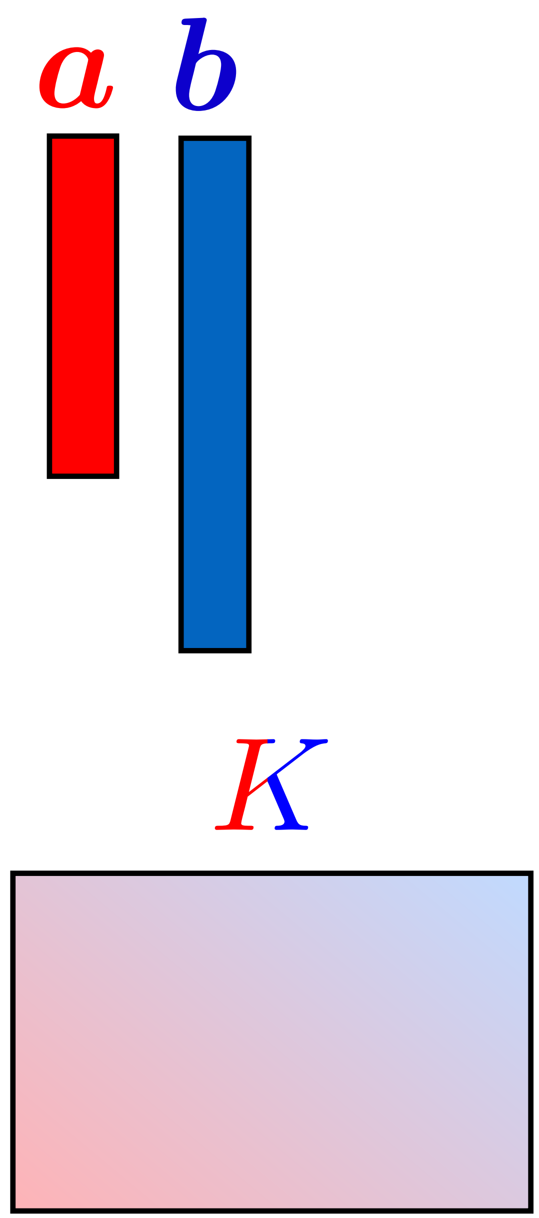
$$u \leftarrow a/Kv, \quad v \leftarrow b/K^T u$$

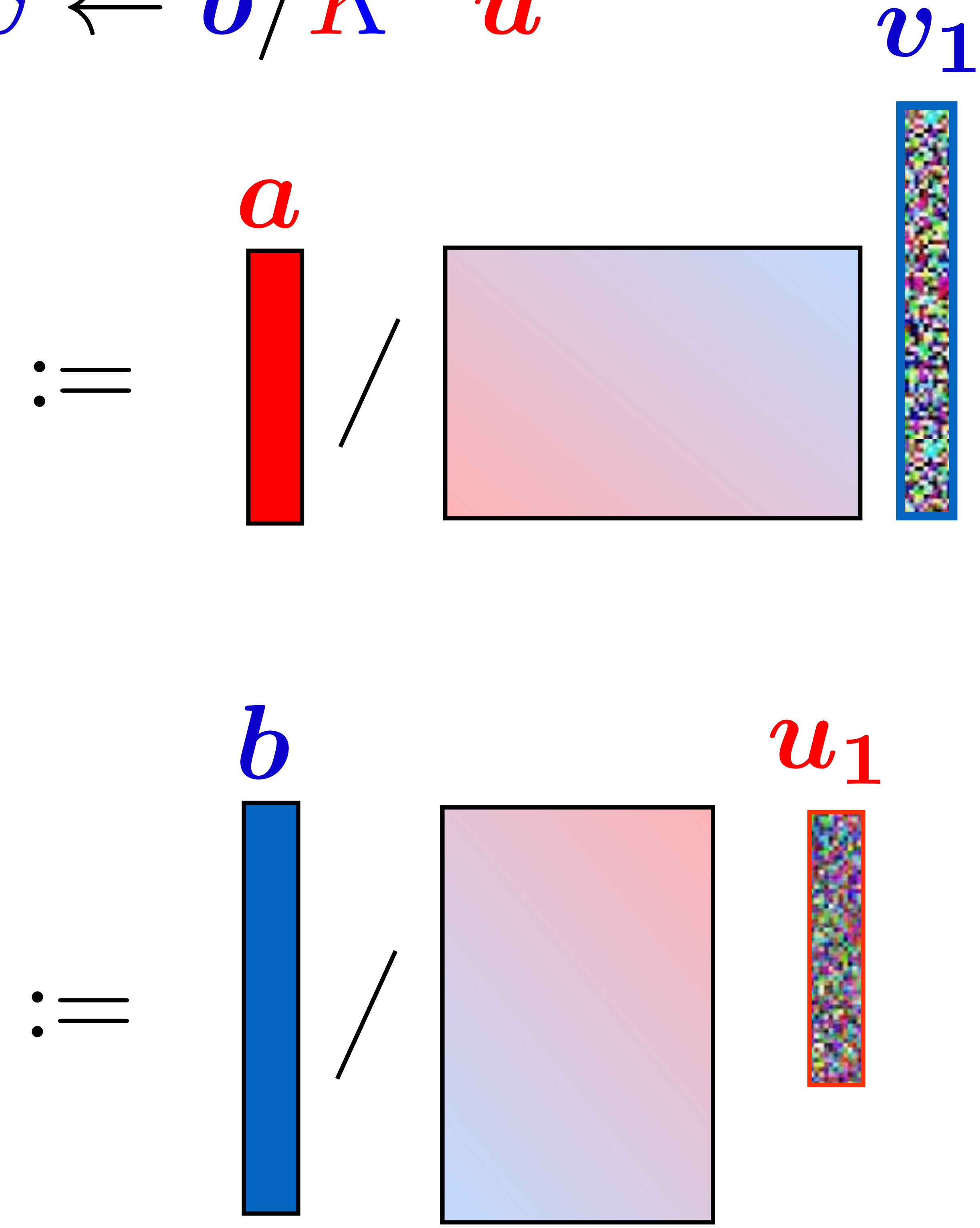


Fast & Scalable Algorithm

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

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

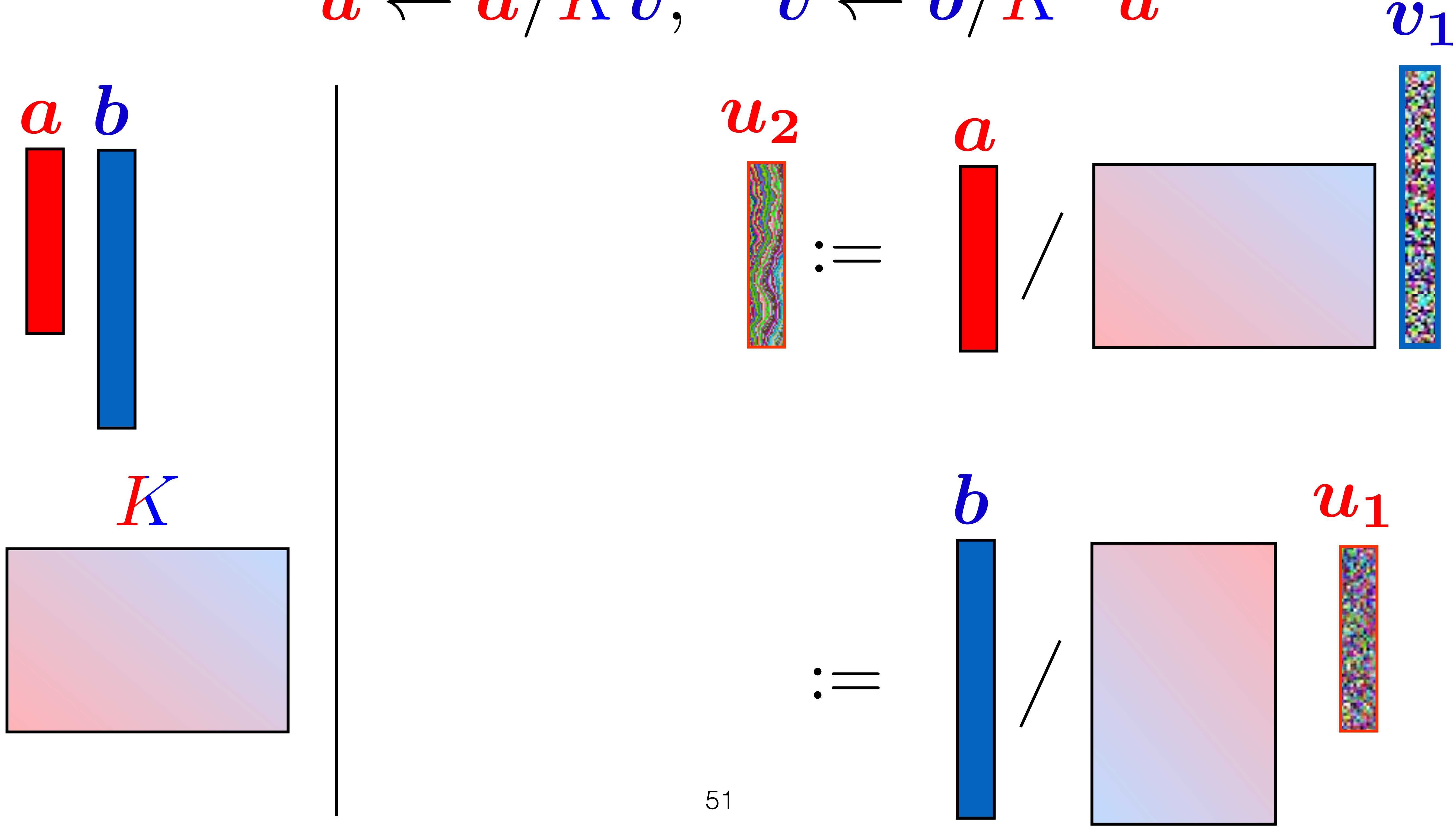




Fast & Scalable Algorithm

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

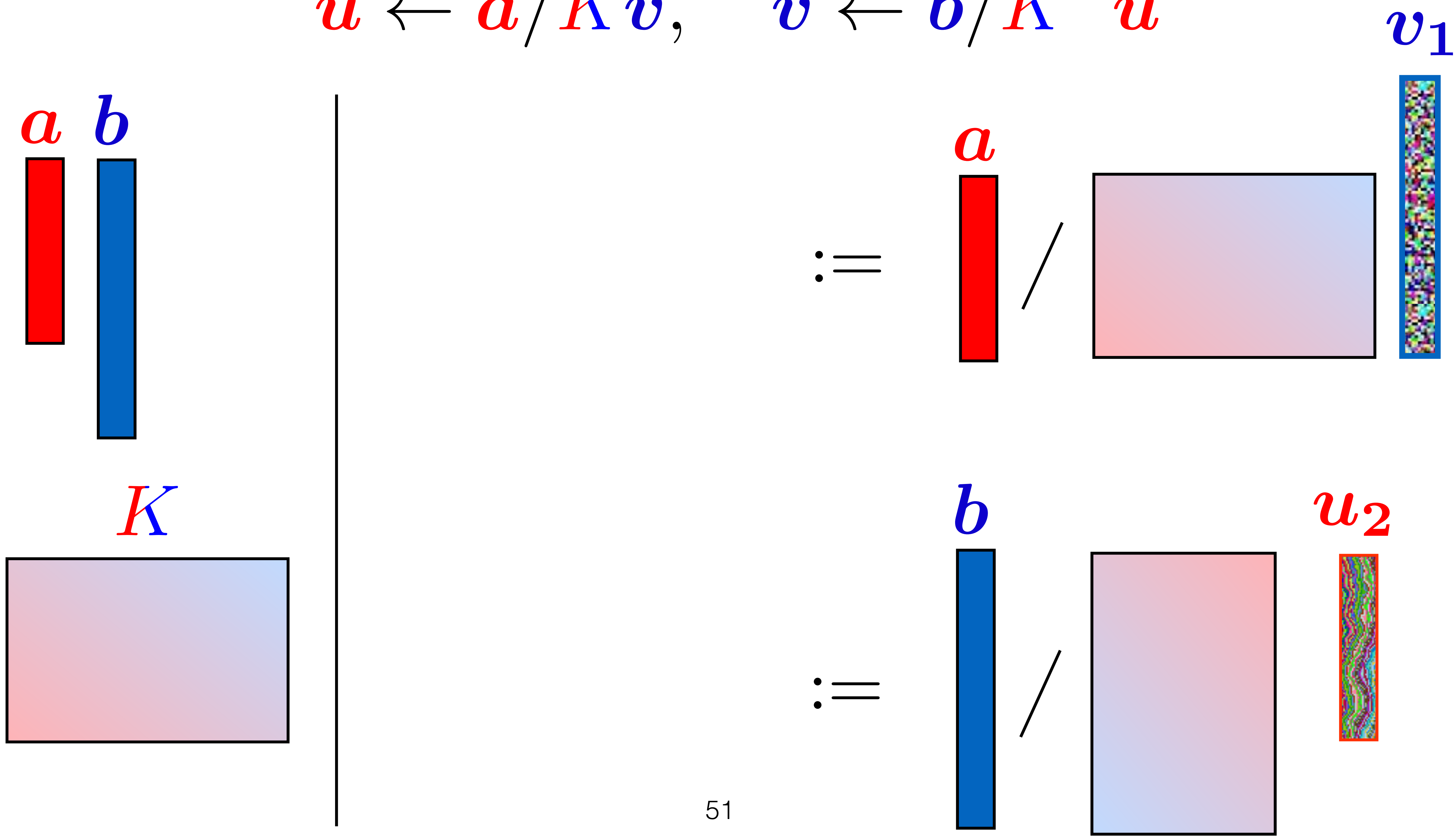
$$\boldsymbol{u} \leftarrow \boldsymbol{a}/K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b}/K^T\boldsymbol{u}$$



Fast & Scalable Algorithm

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

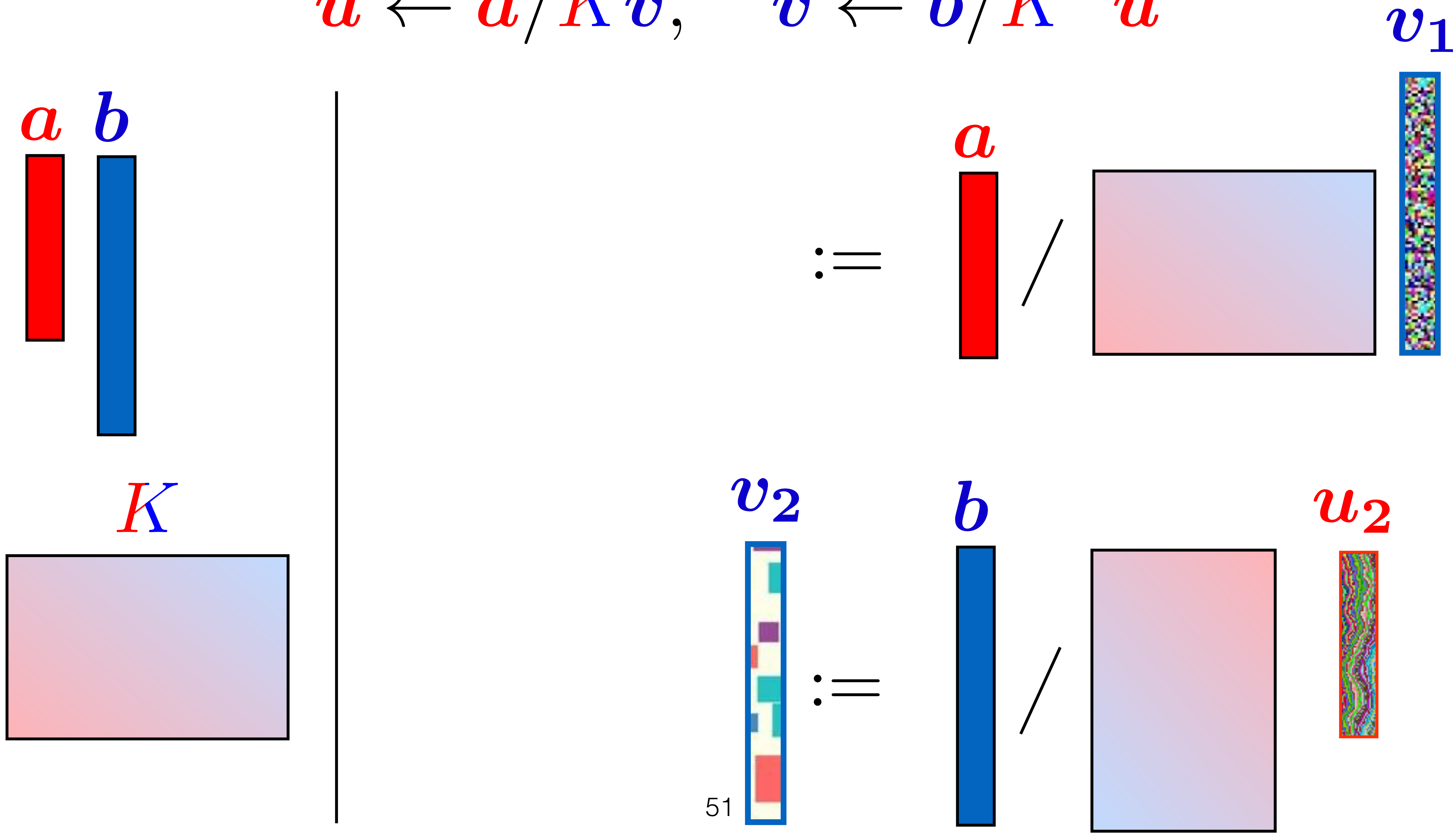
$$\boldsymbol{u} \leftarrow \boldsymbol{a}/K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b}/K^T\boldsymbol{u}$$



Fast & Scalable Algorithm

- **[Sinkhorn'64]** fixed-point iterations for (u, v)

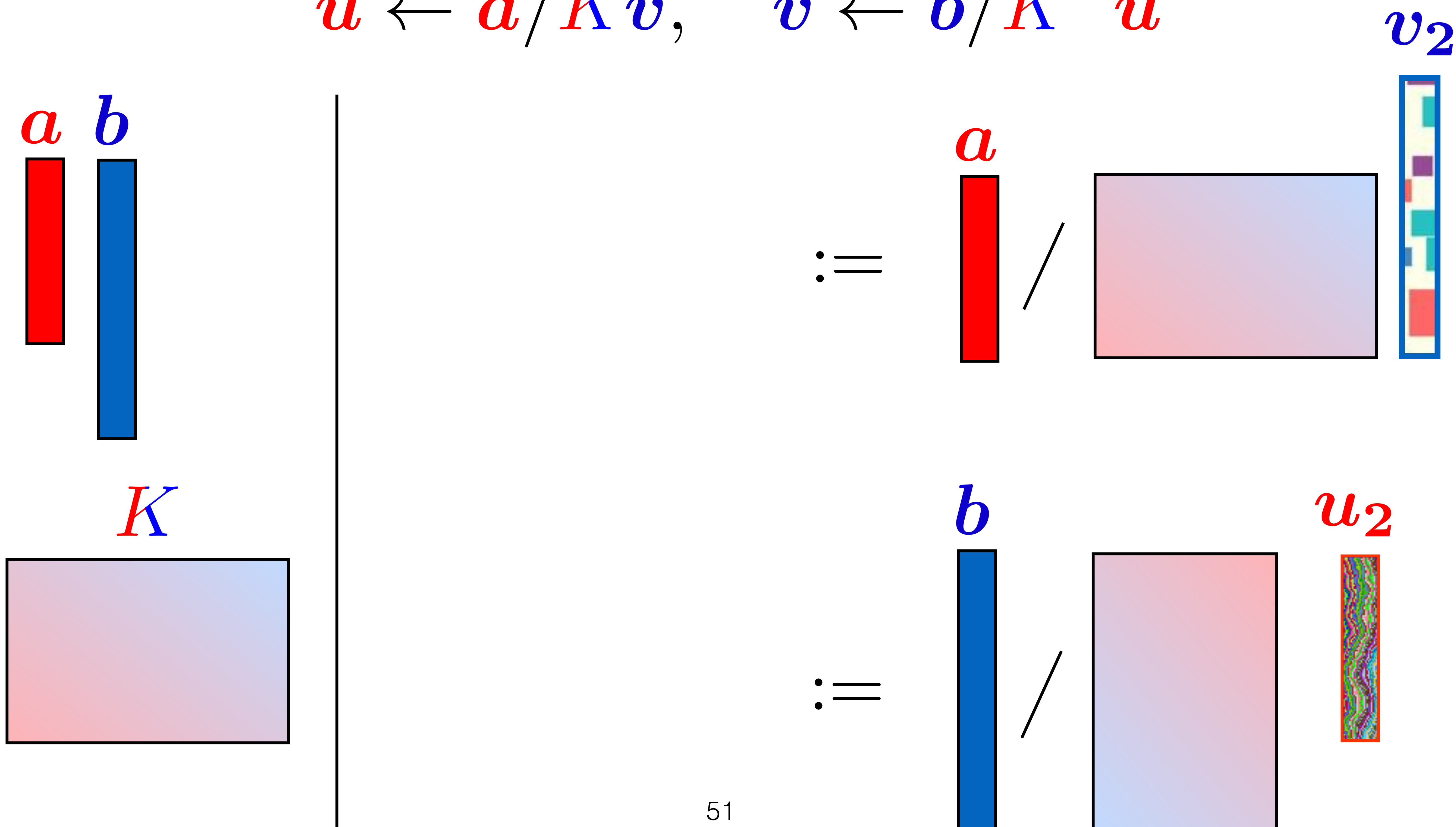
$$u \leftarrow a/Kv, \quad v \leftarrow b/K^T u$$



Fast & Scalable Algorithm

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

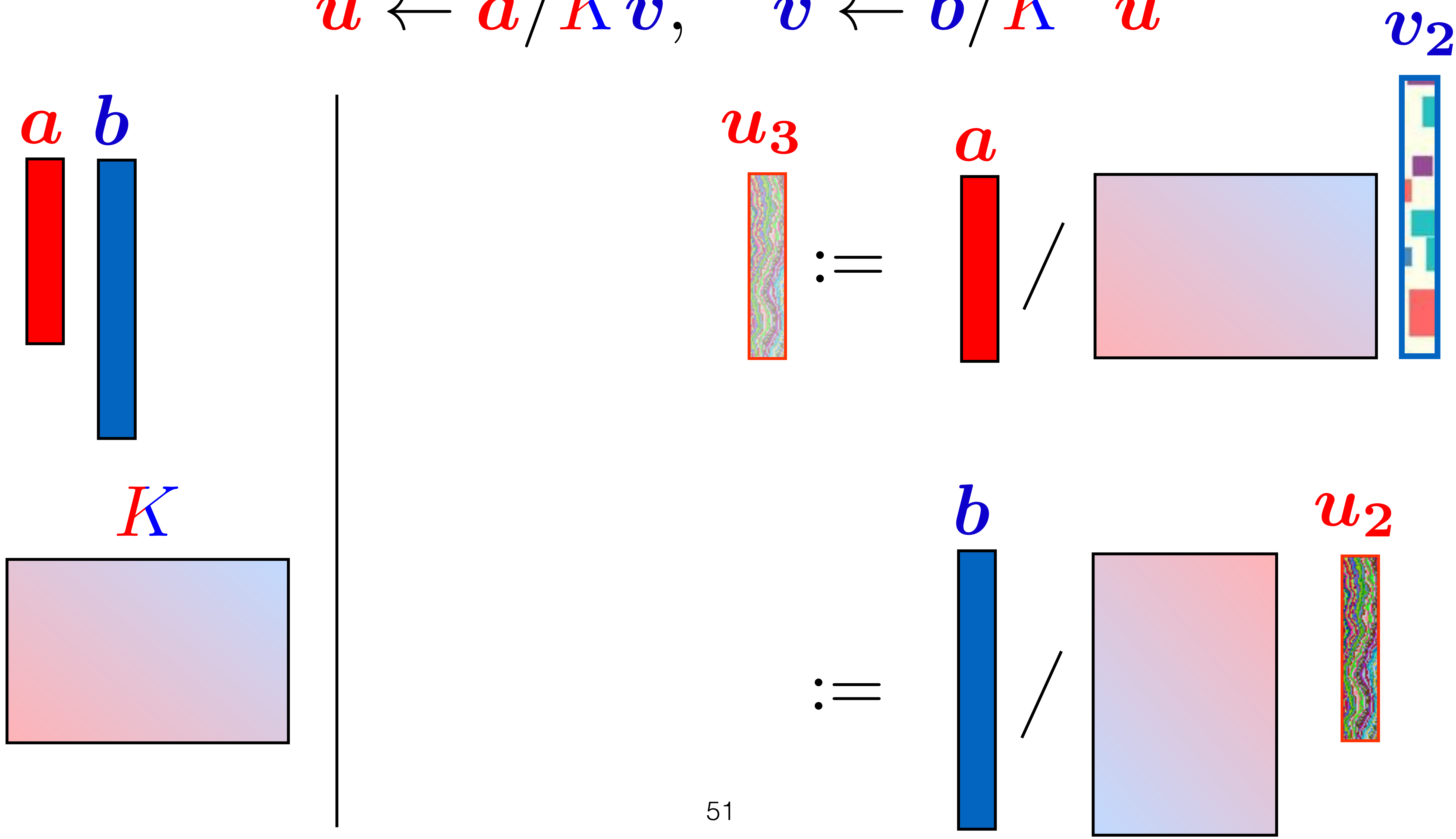
$$\boldsymbol{u} \leftarrow \boldsymbol{a}/K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b}/K^T\boldsymbol{u}$$



Fast & Scalable Algorithm

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

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

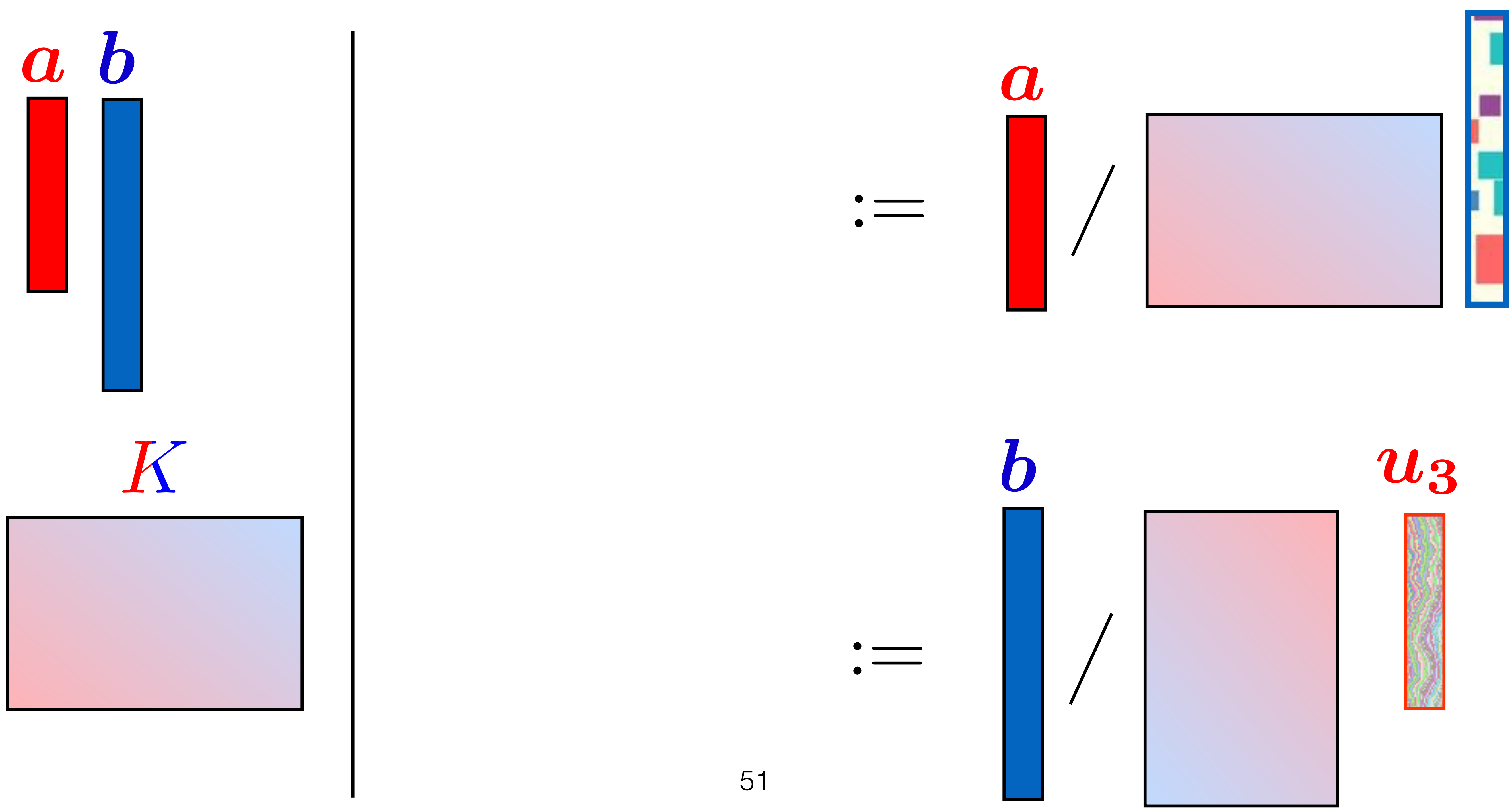


Fast & Scalable Algorithm

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

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

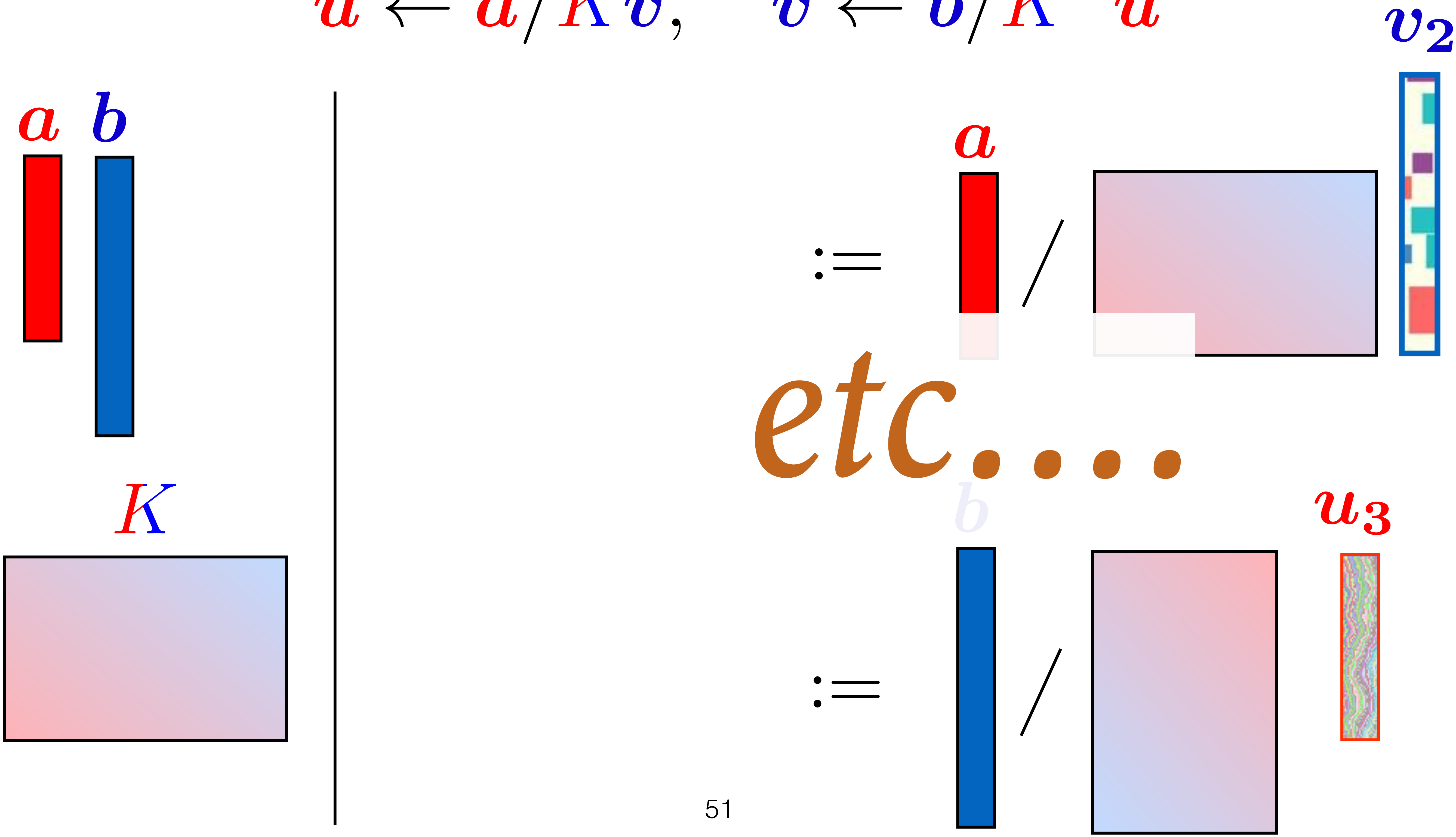
$\boldsymbol{v_2}$



Fast & Scalable Algorithm

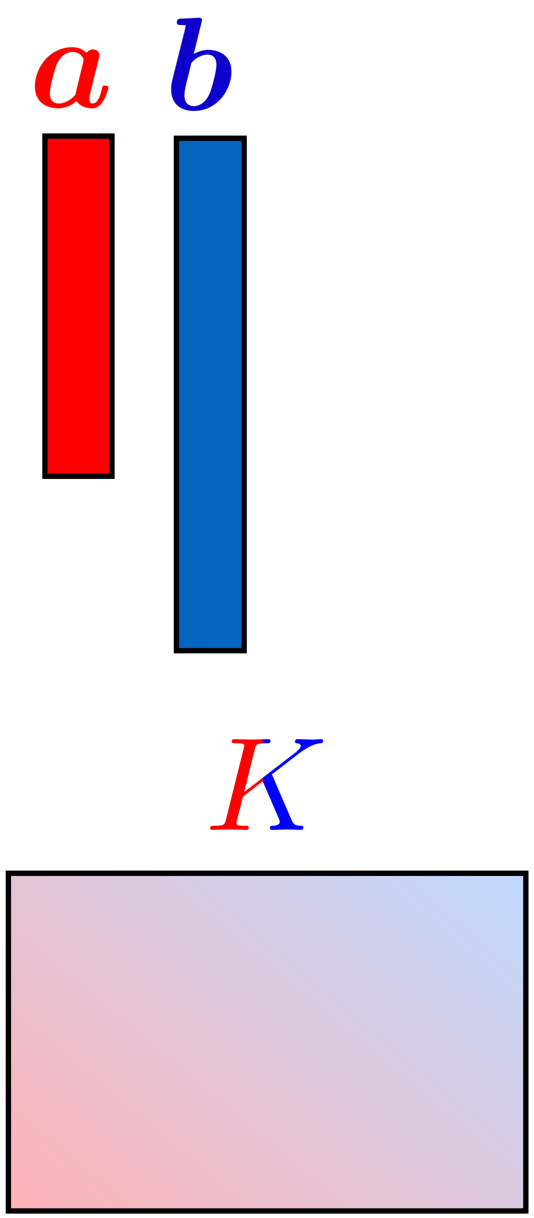
- **[Sinkhorn'64]** fixed-point iterations for $(\boldsymbol{u}, \boldsymbol{v})$

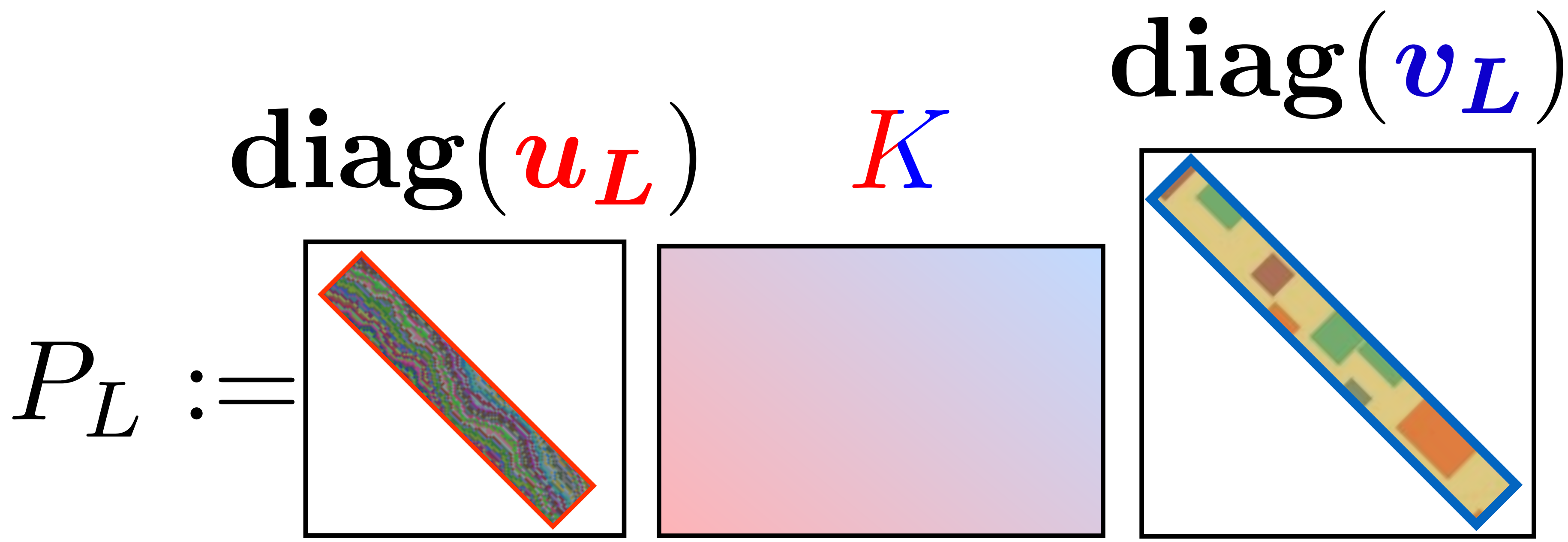
$$\boldsymbol{u} \leftarrow \boldsymbol{a}/K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b}/K^T\boldsymbol{u}$$



Fast & Scalable Algorithm

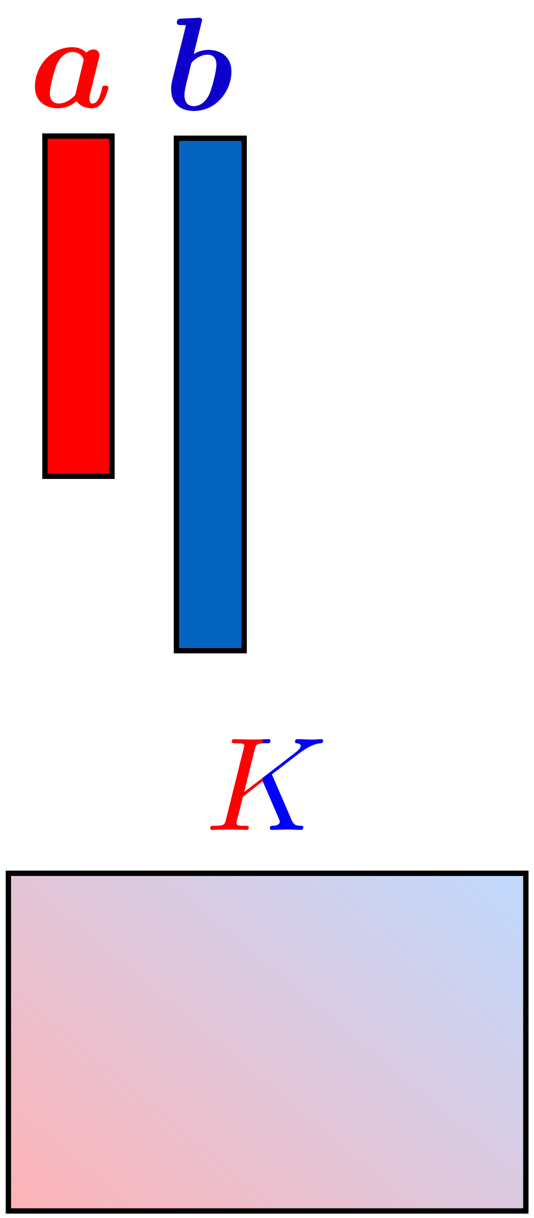
- **[Sinkhorn'64]** fixed-point iterations.



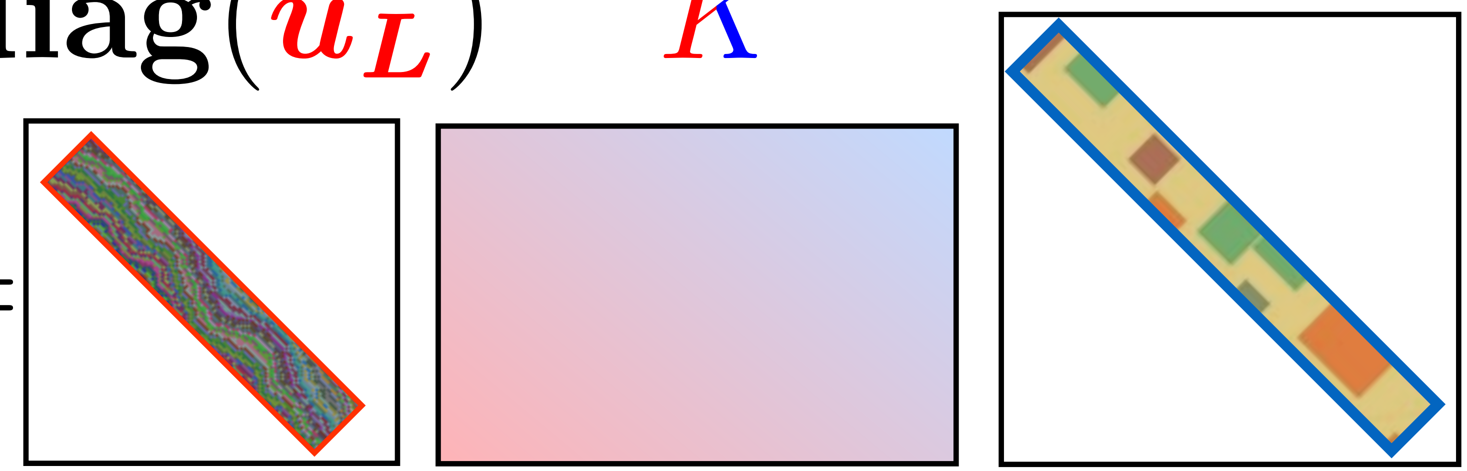


Fast & Scalable Algorithm

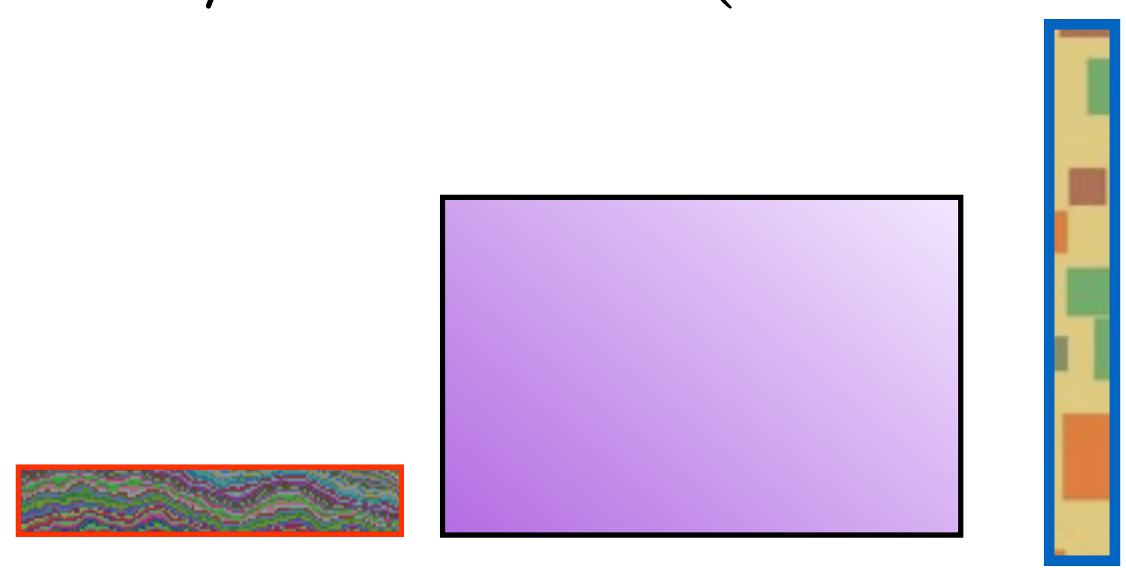
- **[Sinkhorn'64]** fixed-point iterations.



$$P_L := \mathbf{diag}(\boldsymbol{u_L})\, K\, \mathbf{diag}(\boldsymbol{v_L})$$

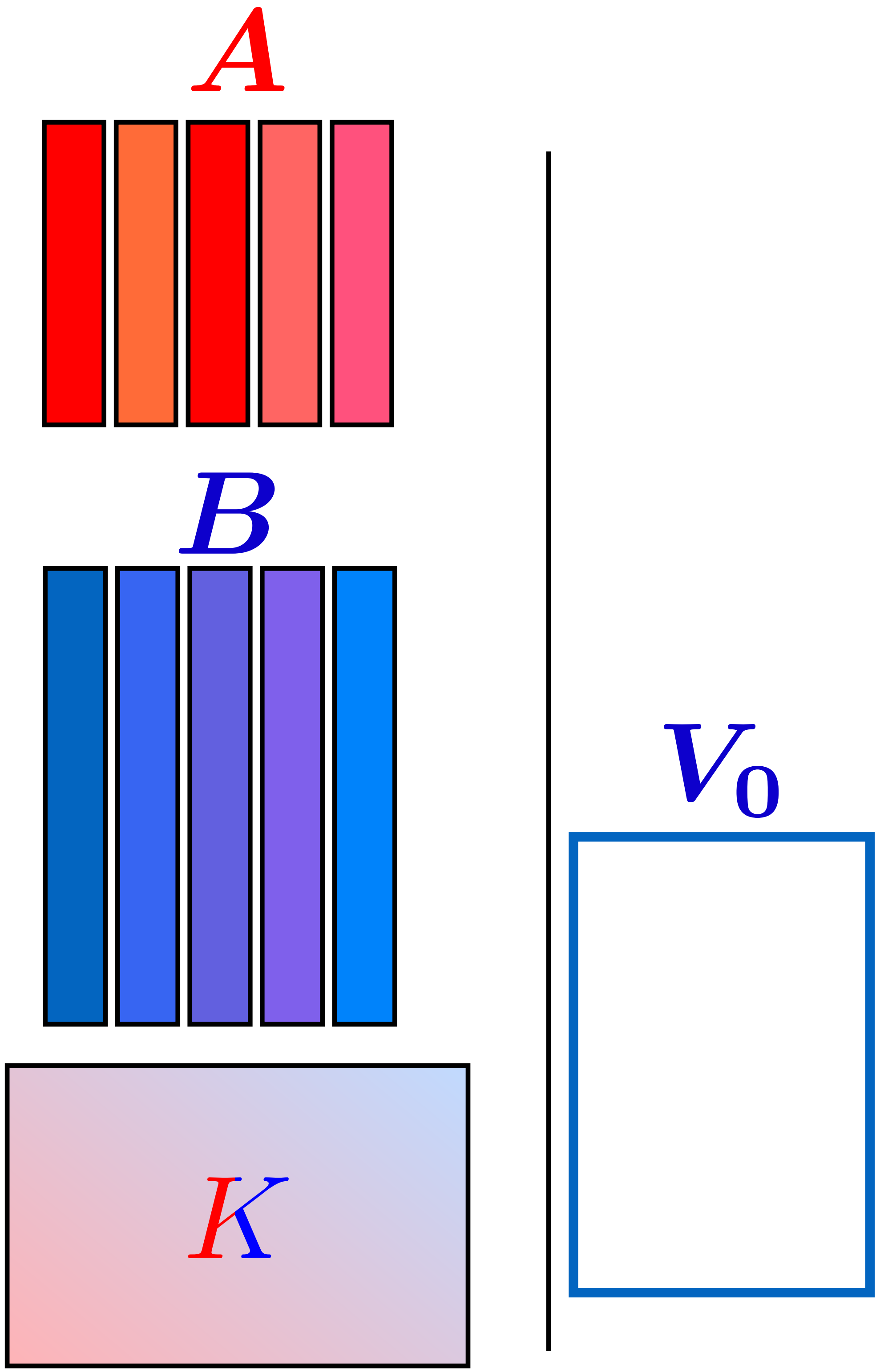


$$\langle P_L, M_{\boldsymbol{XY}} \rangle = \boldsymbol{u_L}^T (K \odot M_{\boldsymbol{XY}}) \boldsymbol{v_L}$$



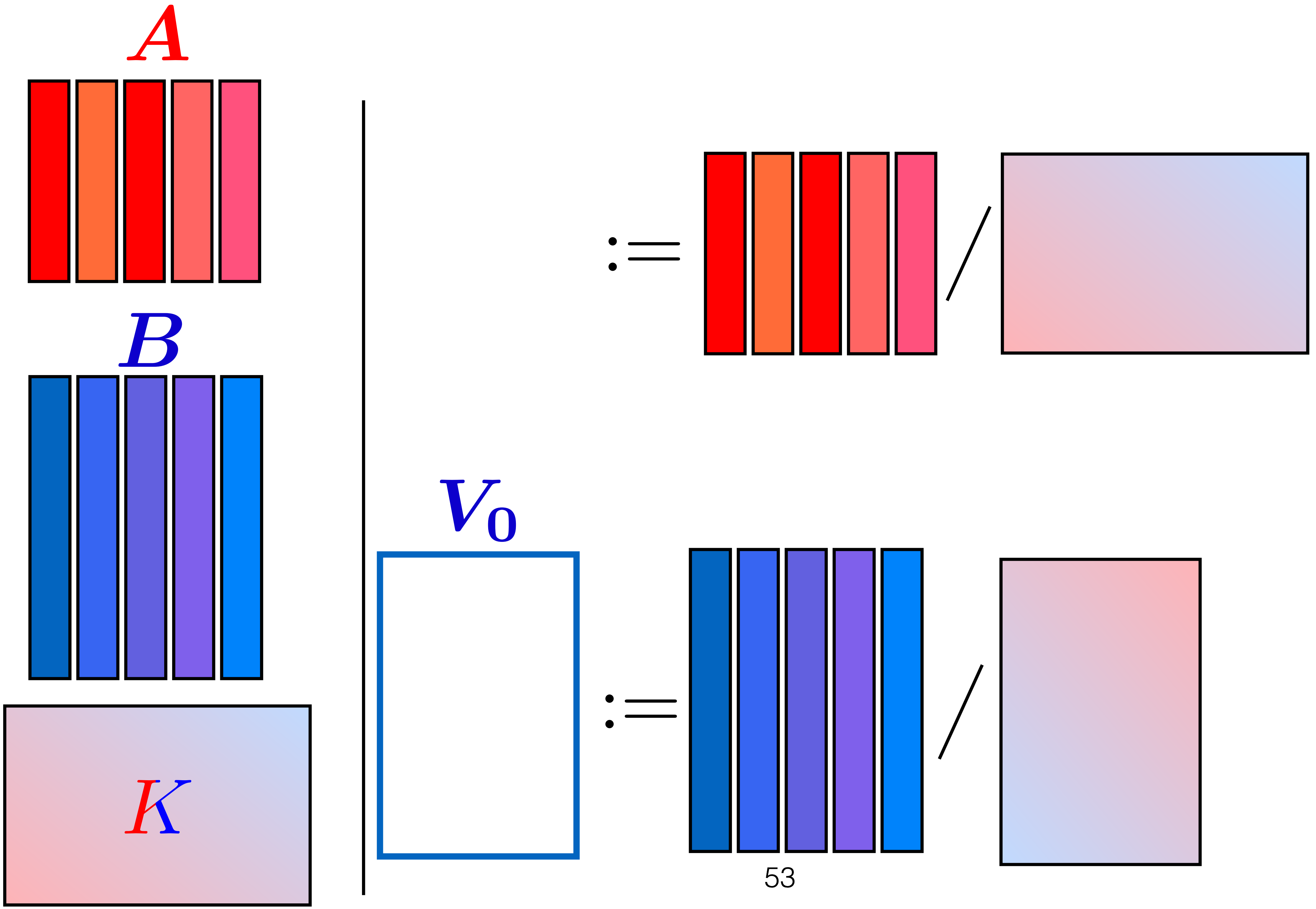
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



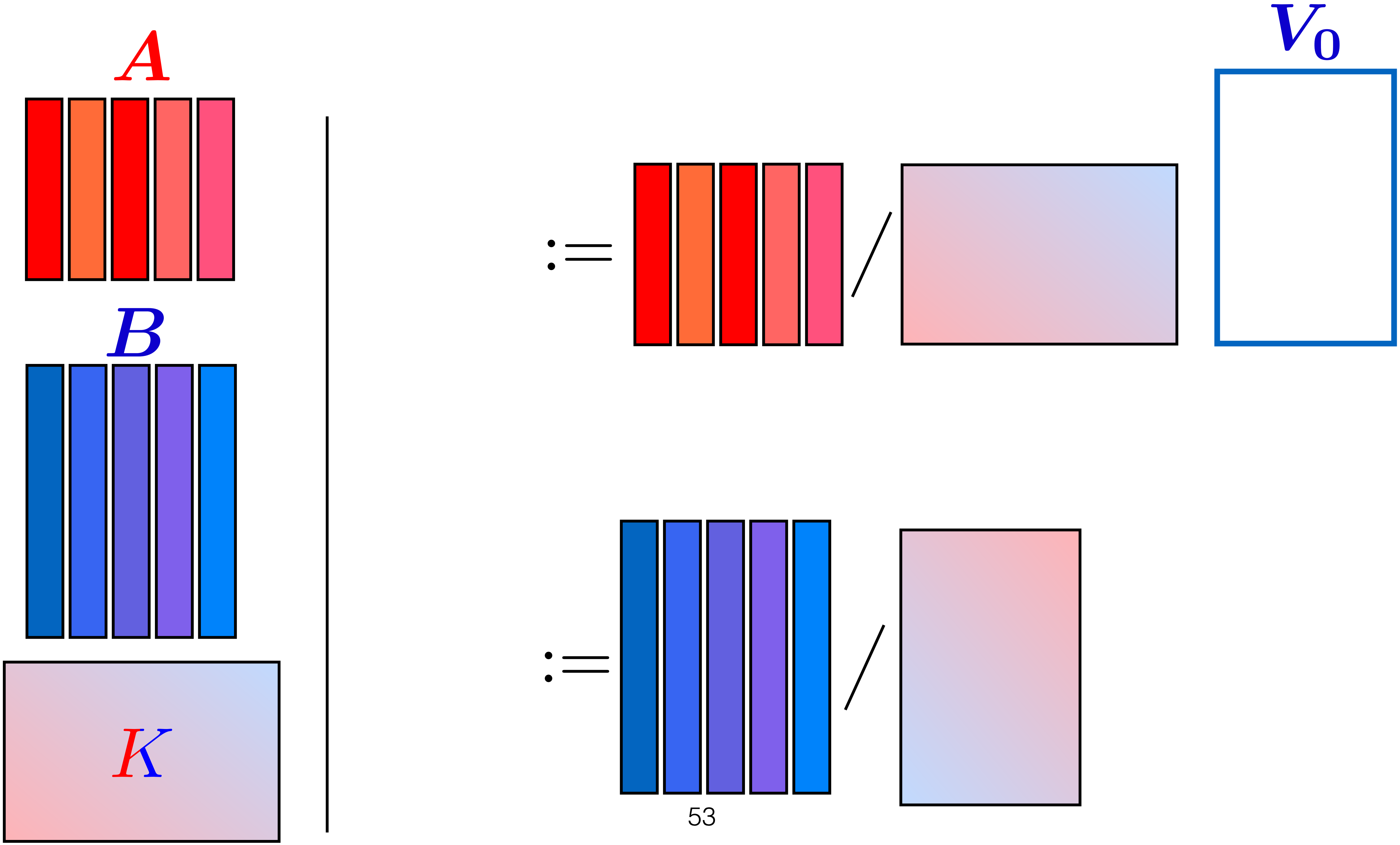
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



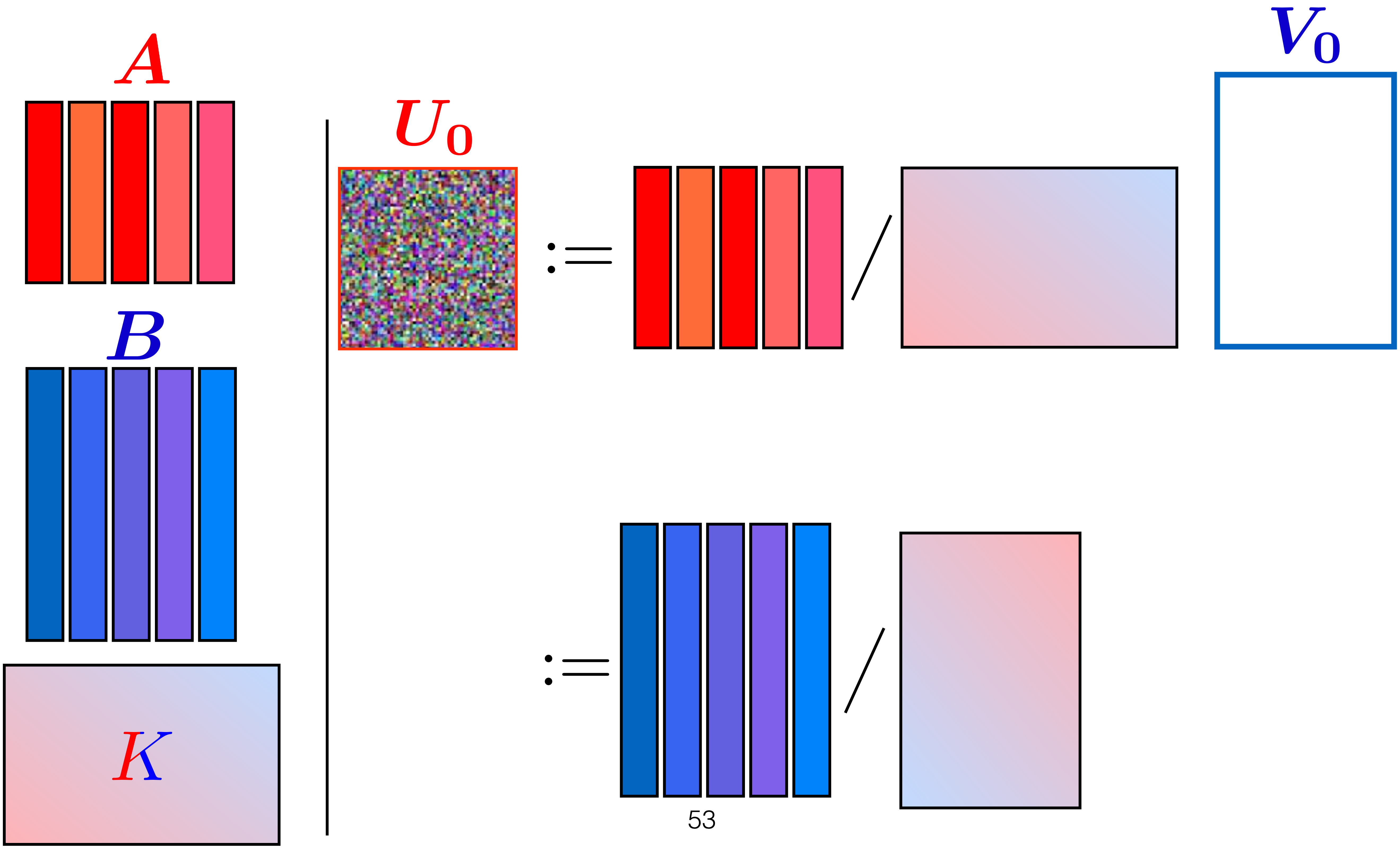
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



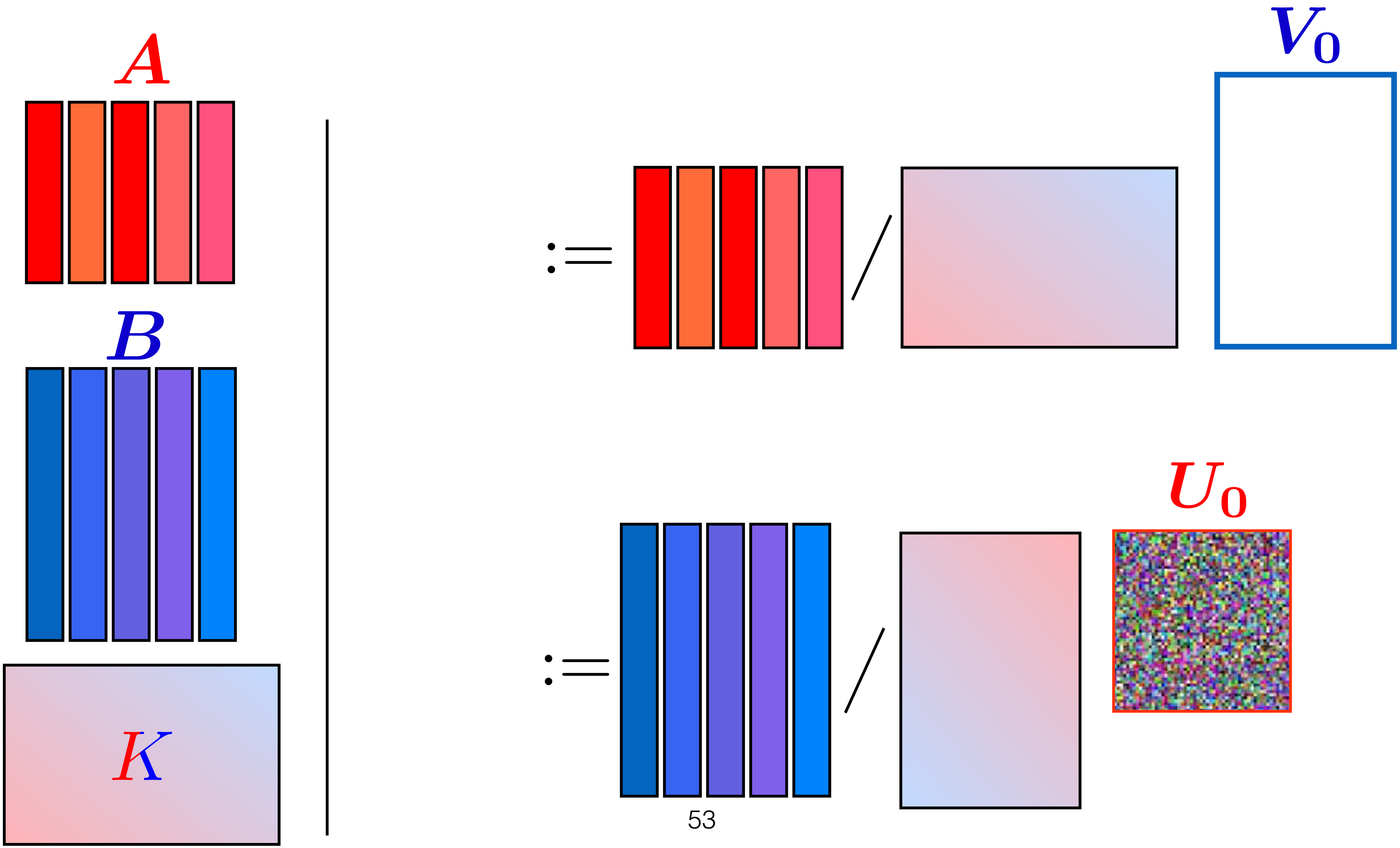
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



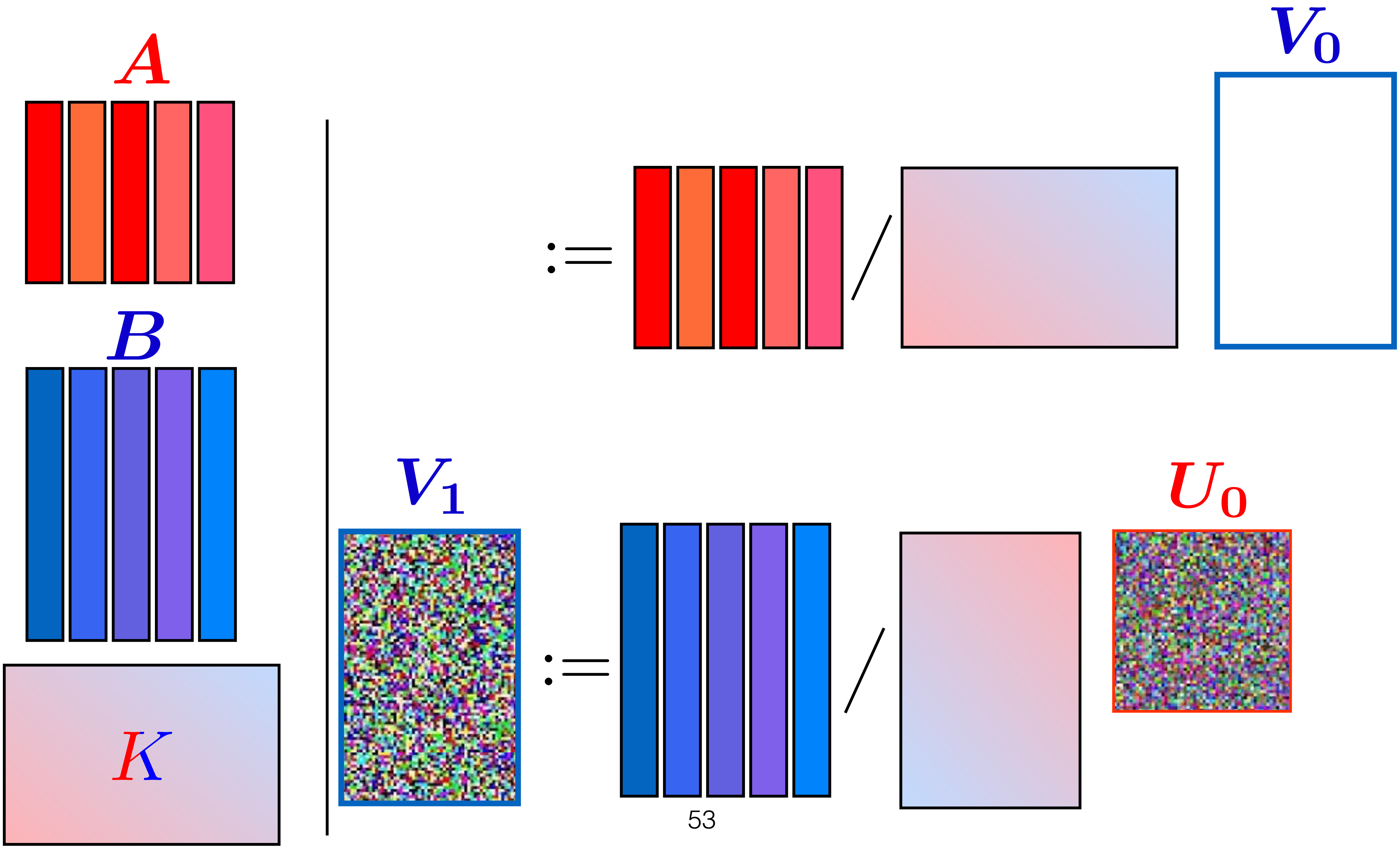
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



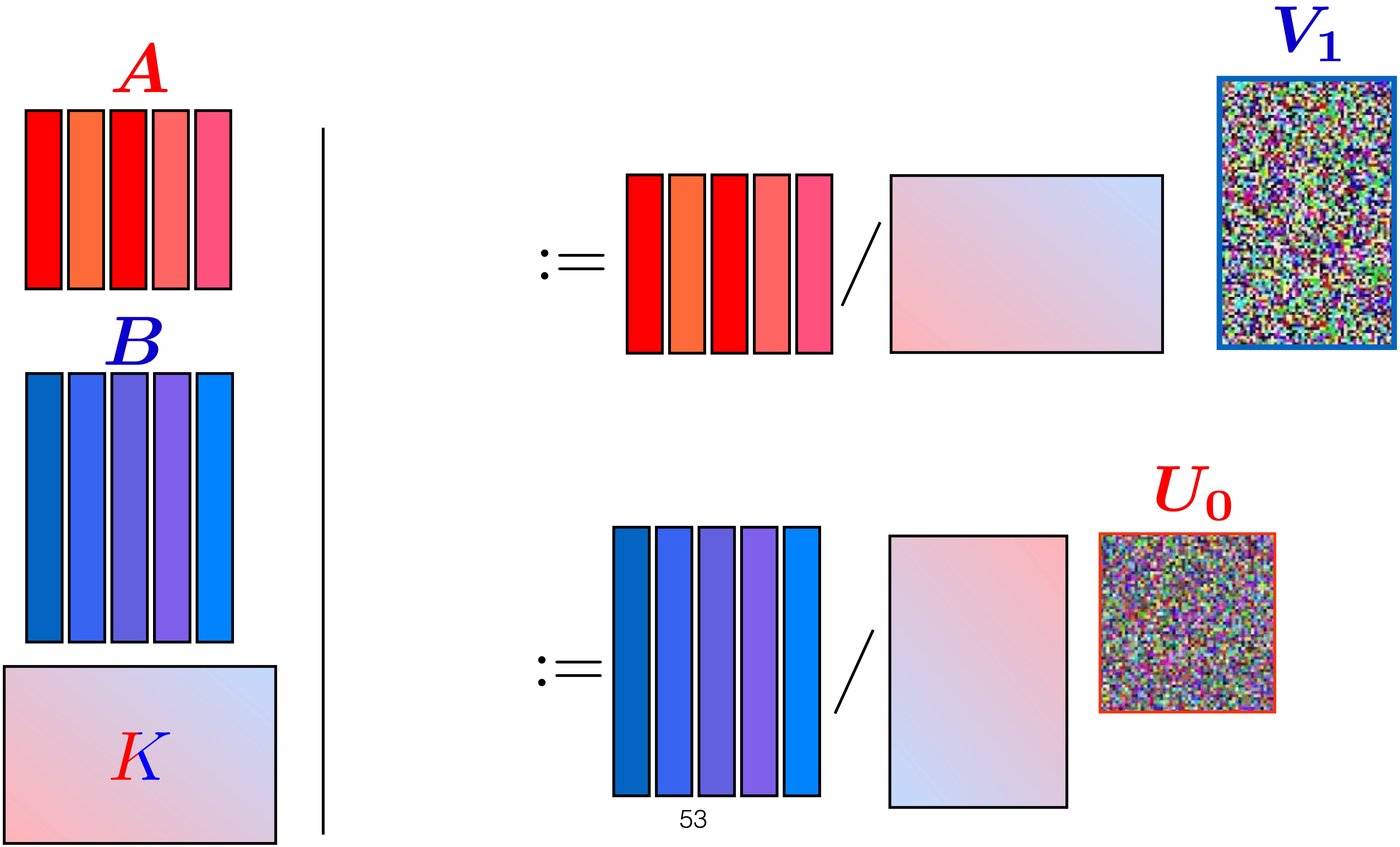
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



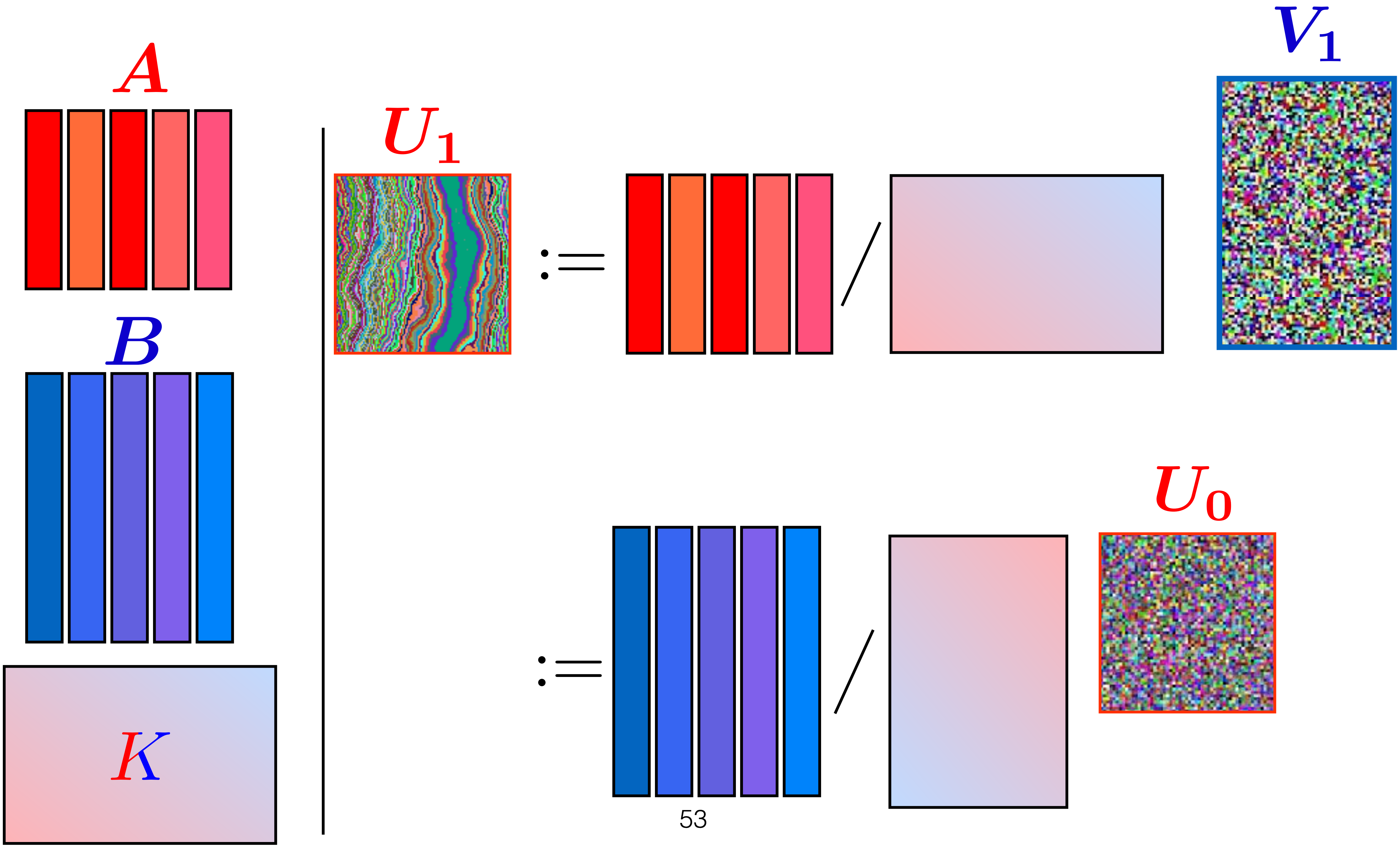
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



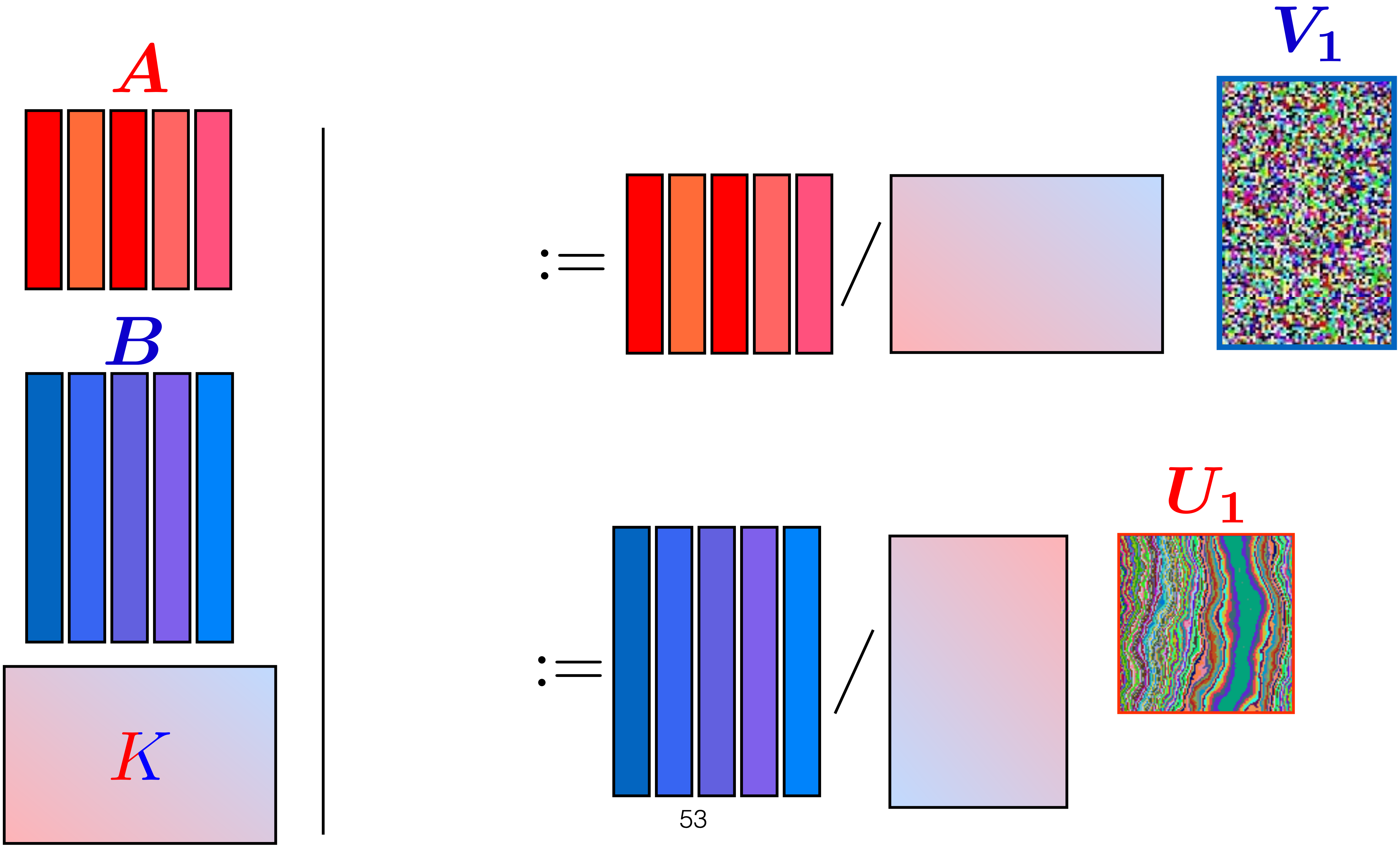
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



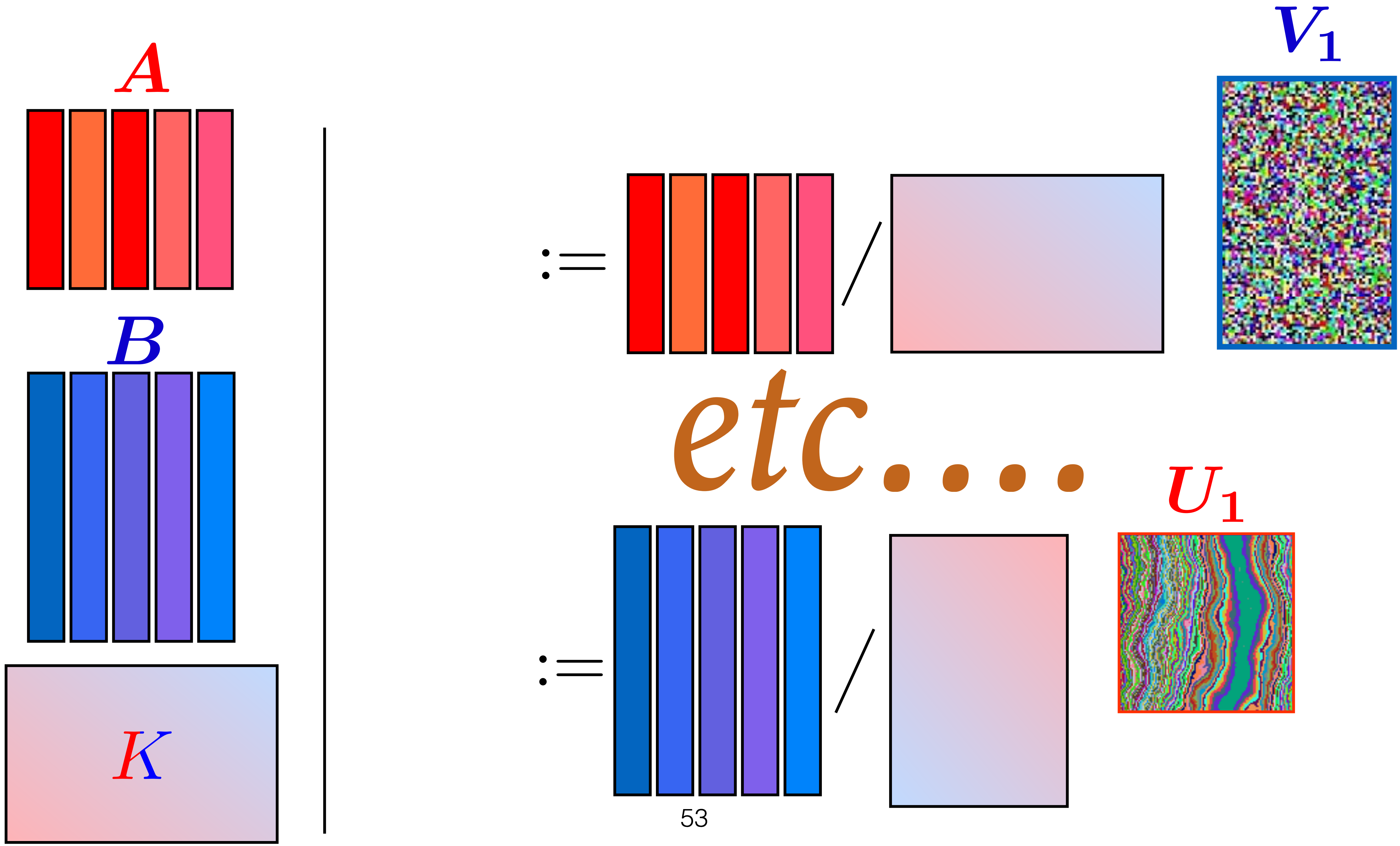
Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations

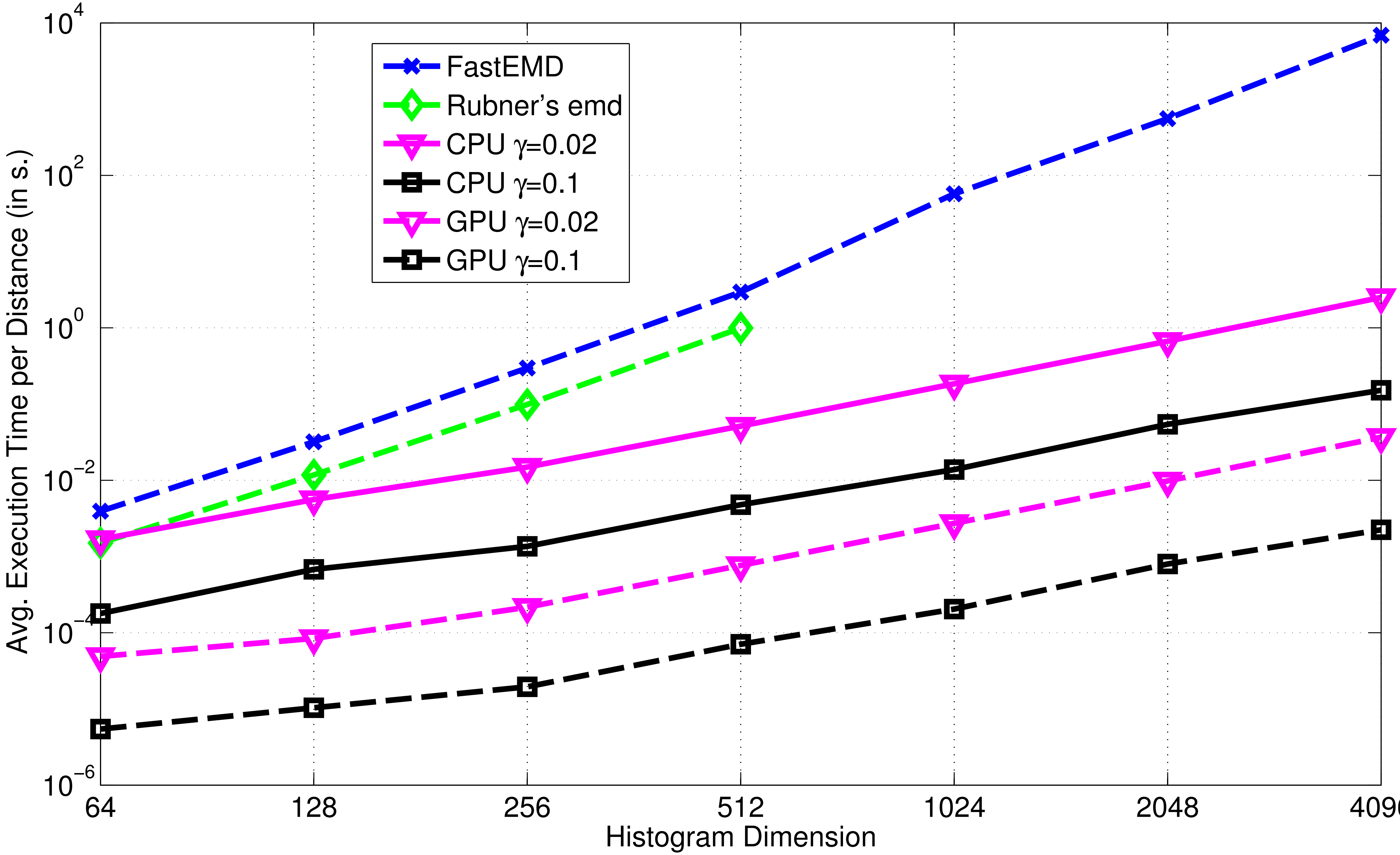


Also embarrassingly parallel

- **[Sinkhorn'64]** with *matrix* fixed-point iterations



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}.

Sinkhorn as a Dual Algorithm

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

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

REGULARIZED DISCRETE PRIMAL

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

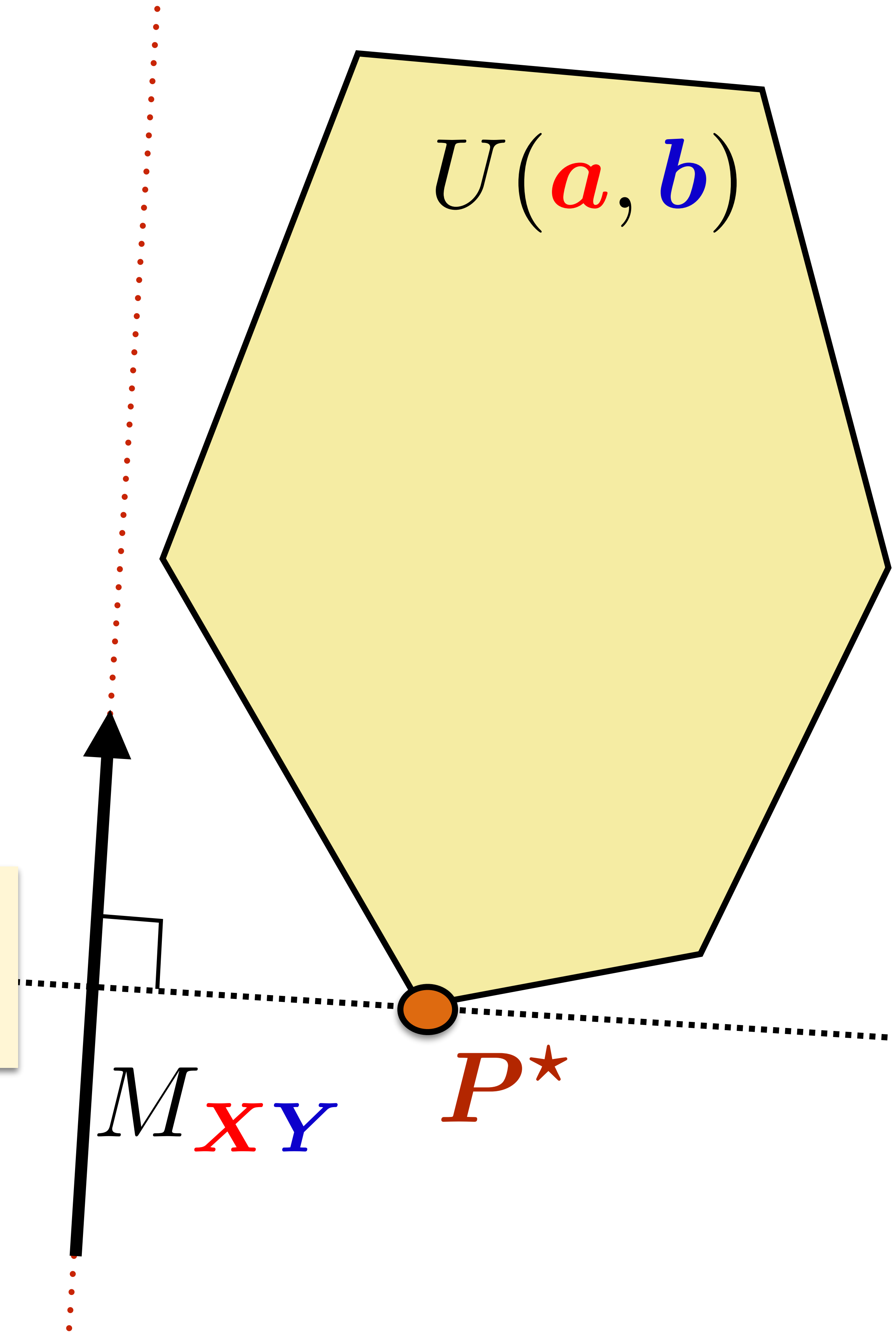
REGULARIZED DISCRETE DUAL

Sinkhorn = *Block Coordinate Ascent* on Dual

Sinkhorn in between W and MMD

$$\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a}_i \delta_{\boldsymbol{x}_i} \qquad \boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b}_j \delta_{\boldsymbol{y}_j}$$

$$W^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\mathbf{XY}} \rangle$$

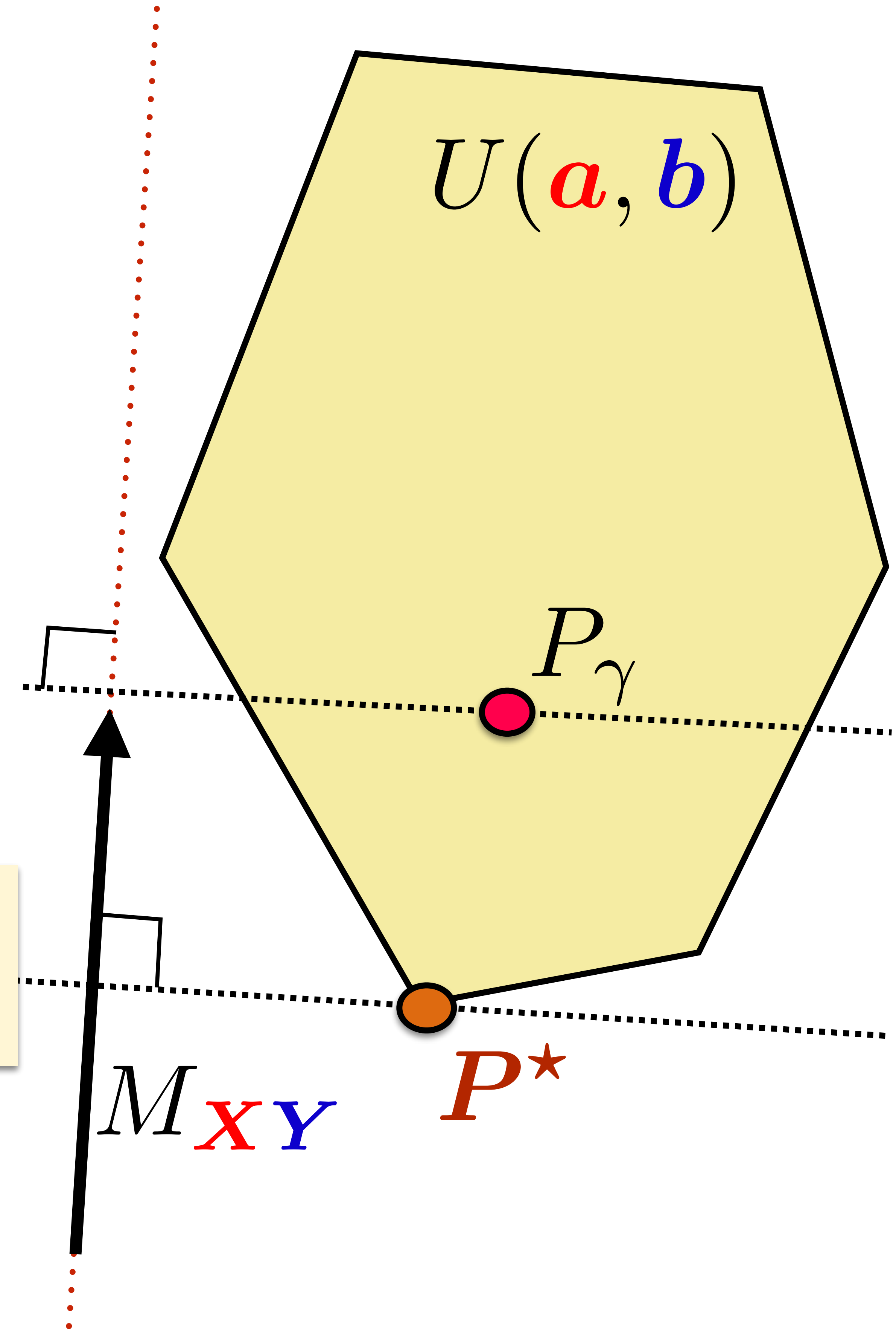


Sinkhorn in between W and MMD

$$\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a}_i \delta_{\boldsymbol{x}_i} \qquad \boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b}_j \delta_{\boldsymbol{y}_j}$$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle P_\gamma, M_{\mathbf{XY}} \rangle$$

$$W^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\mathbf{XY}} \rangle$$



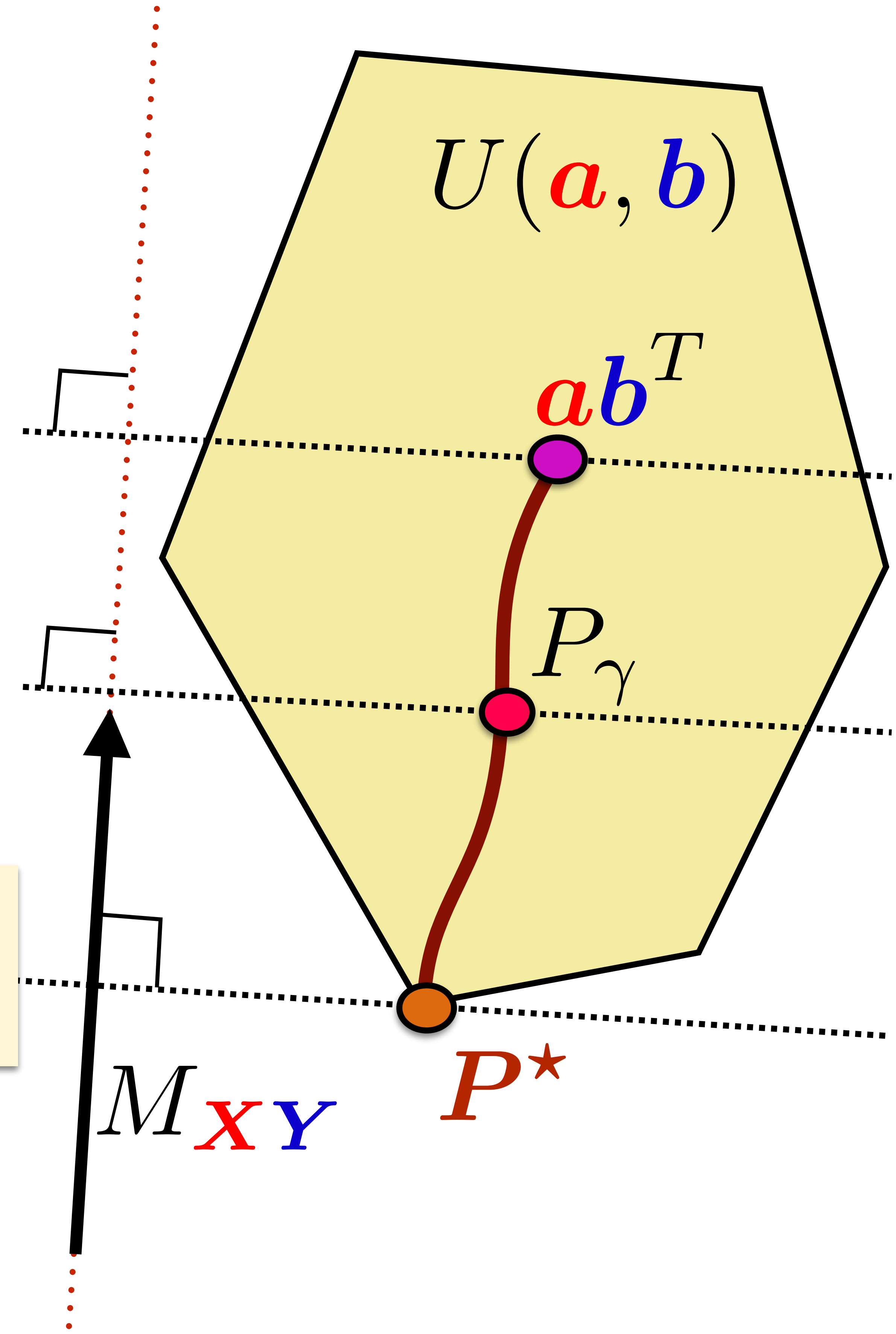
Sinkhorn in between W and MMD

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

$$\mathcal{E}(\mu, \nu) = \langle \mathbf{a}\mathbf{b}^T, M_{\mathbf{XY}} \rangle$$

$$W_\gamma(\mu, \nu) = \langle P_\gamma, M_{\mathbf{XY}} \rangle$$

$$W^p(\mu, \nu) = \langle P^\star, M_{\mathbf{XY}} \rangle$$



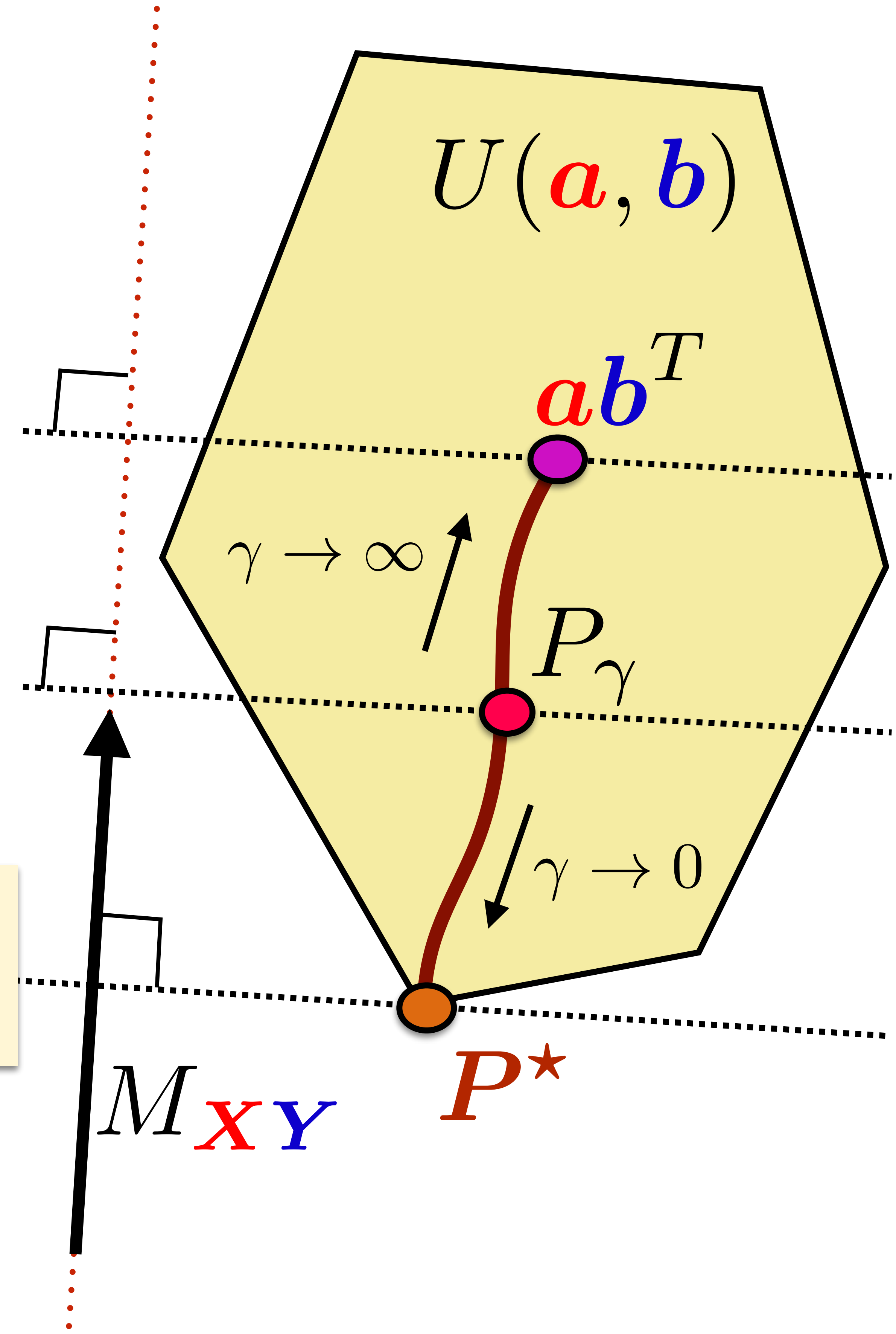
Sinkhorn in between W and MMD

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

$$\mathcal{E}(\mu, \nu) = \langle \mathbf{a}\mathbf{b}^T, M_{\mathbf{XY}} \rangle$$

$$W_\gamma(\mu, \nu) = \langle P_\gamma, M_{\mathbf{XY}} \rangle$$

$$\boxed{W^p(\mu, \nu) = \langle P^\star, M_{\mathbf{XY}} \rangle}$$



Sinkhorn in between W and MMD

$$\mathcal{E}(\boldsymbol{\mu},\boldsymbol{\nu}) = \langle \boldsymbol{a}\boldsymbol{b}^T, M_{\mathbf{X}\mathbf{Y}} \rangle$$

$$\mathcal{M}\mathcal{M}\mathcal{D}(\boldsymbol{\mu},\boldsymbol{\nu}) = \mathcal{E}(\boldsymbol{\mu},\boldsymbol{\nu}) - \frac{1}{2}(\mathcal{E}(\boldsymbol{\mu},\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\nu},\boldsymbol{\nu}))$$

$$W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = \langle P_\gamma, M_{\mathbf{X}\mathbf{Y}} \rangle$$

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

$$W^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\mathbf{X}\mathbf{Y}} \rangle$$

Sinkhorn in between W and MMD

$$\mathcal{MMD}(\boldsymbol{\mu},\boldsymbol{\nu}) = \mathcal{E}(\boldsymbol{\mu},\boldsymbol{\nu}) - \frac{1}{2}(\mathcal{E}(\boldsymbol{\mu},\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\nu},\boldsymbol{\nu}))$$

$$\Big\uparrow \quad \gamma \to \infty$$

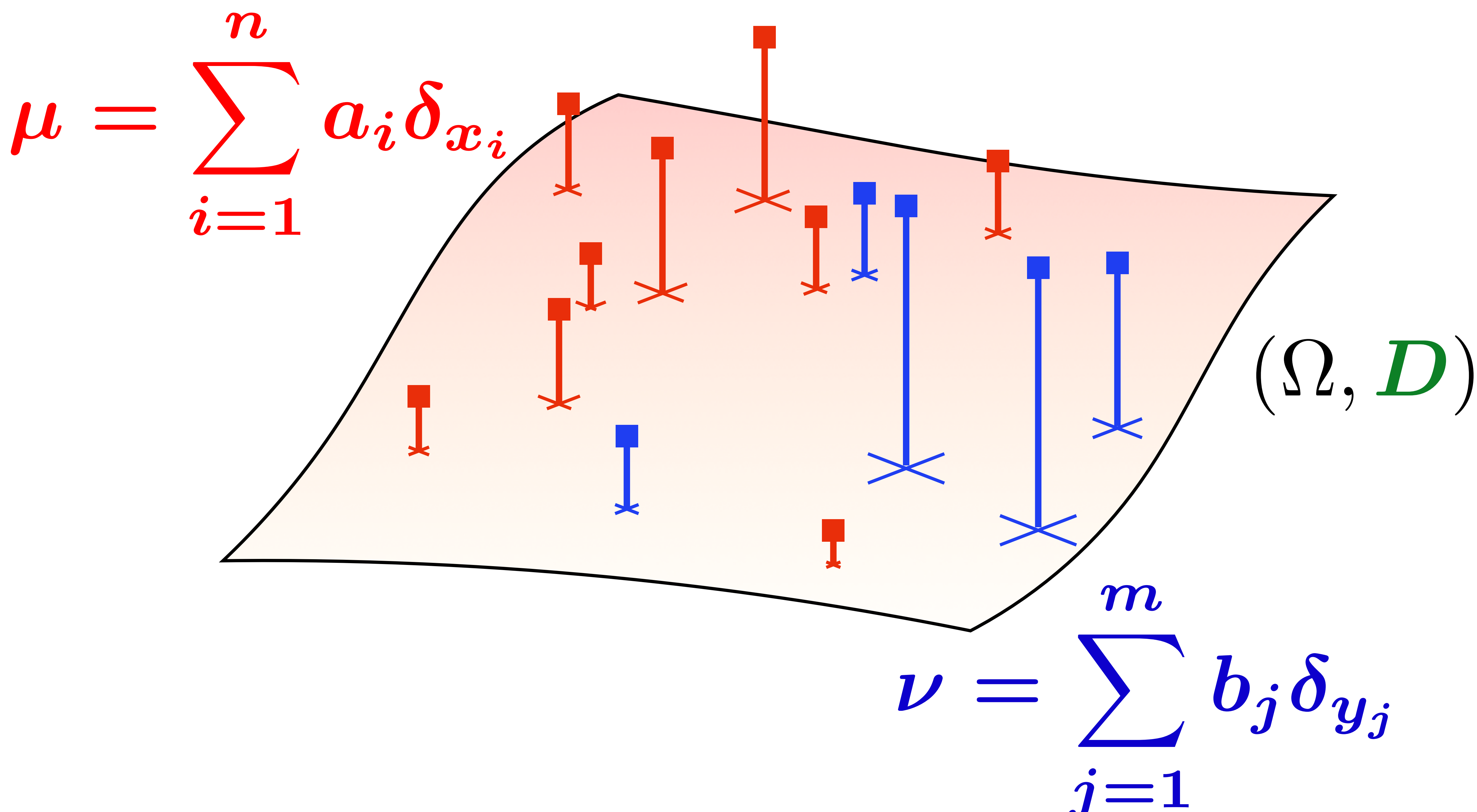
$$\bar{W}_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) = W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu}) - \frac{1}{2}(W_\gamma(\boldsymbol{\mu},\boldsymbol{\mu}) + W_\gamma(\boldsymbol{\nu},\boldsymbol{\nu}))$$

$$\Big\downarrow \quad \gamma \to 0$$

$$W^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

Differentiability of W

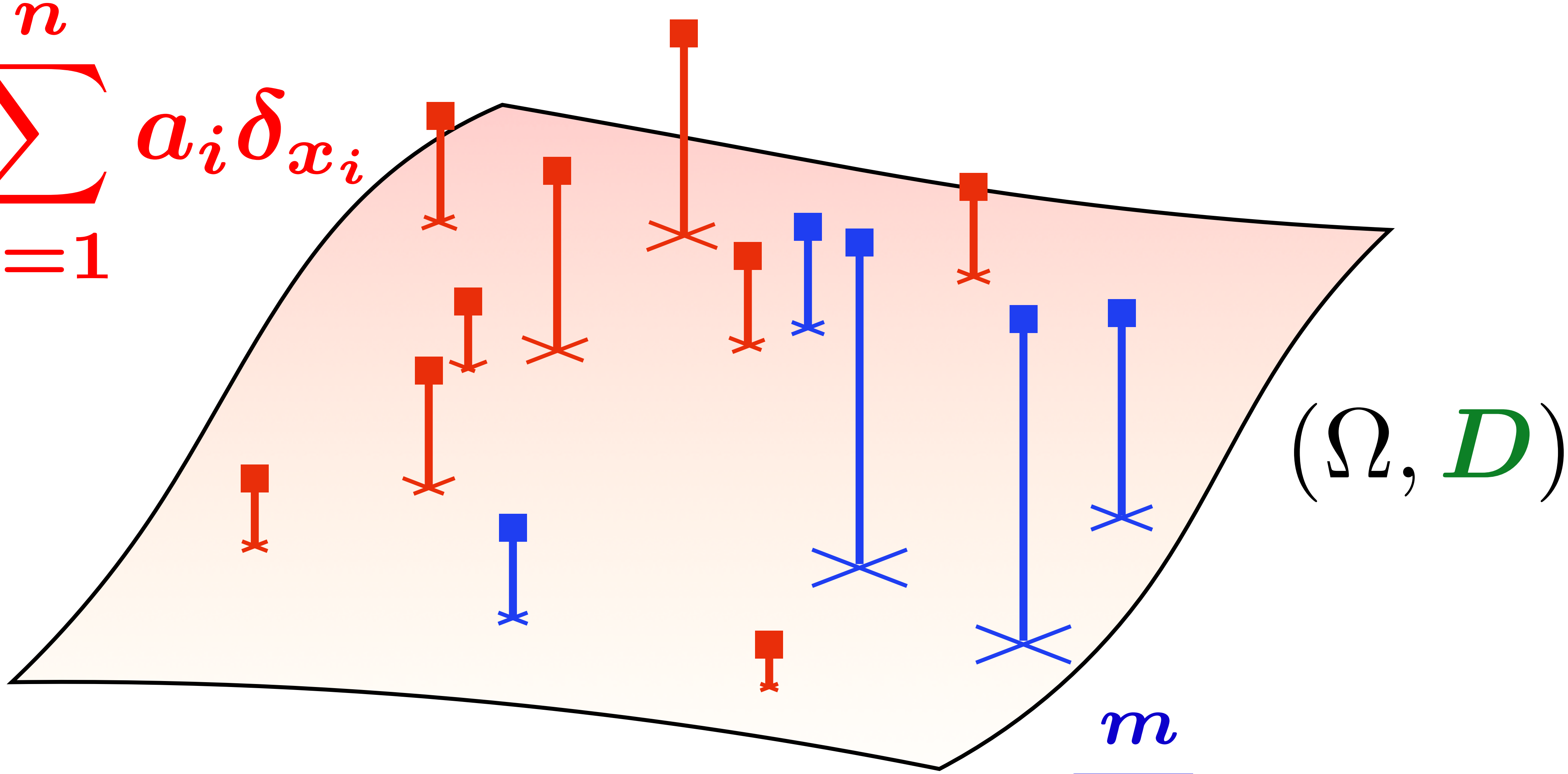
$$W(\boldsymbol{(a, X)}, \boldsymbol{(b, Y)})$$



Differentiability of W

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

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

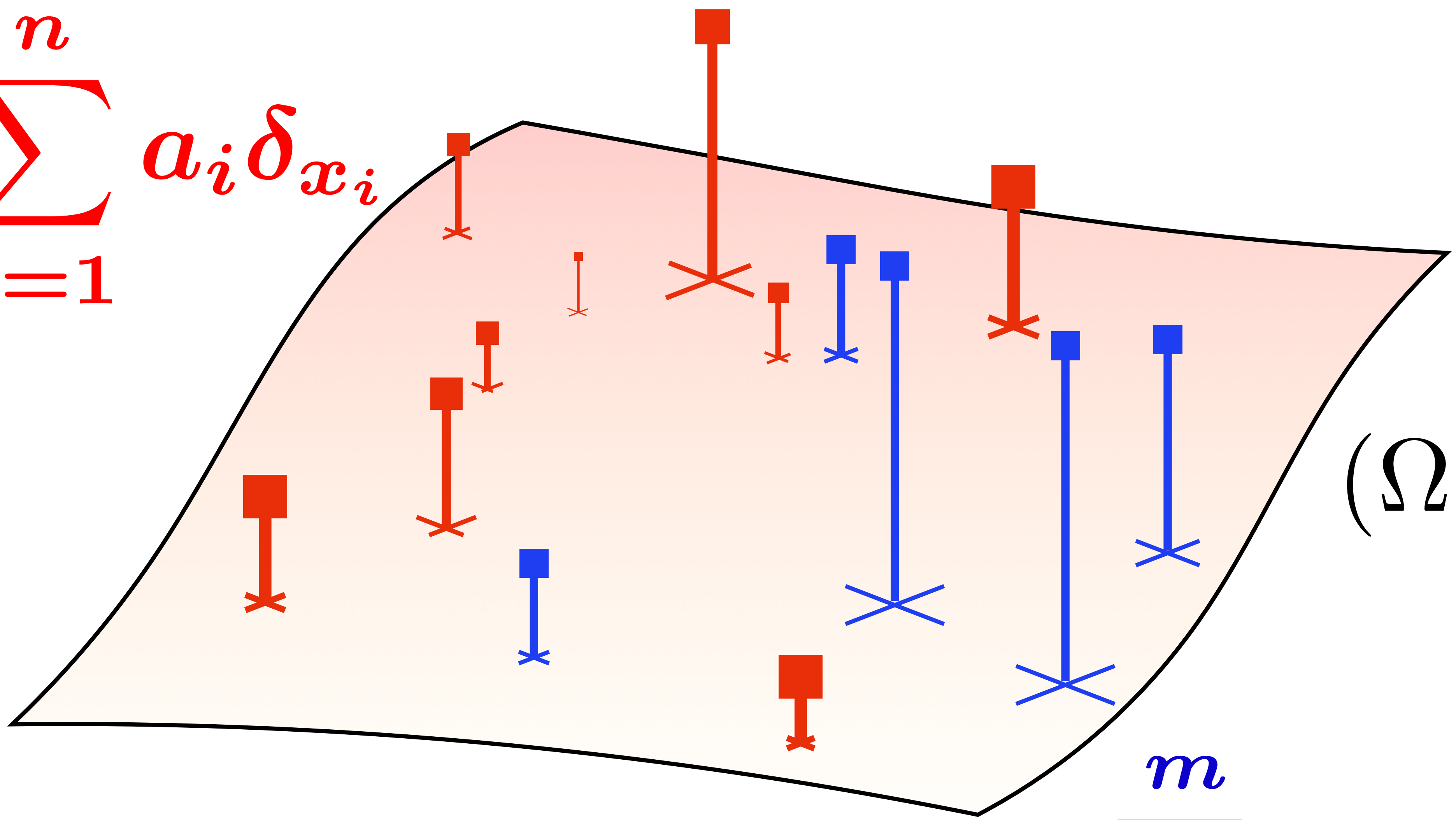


$$\boldsymbol{\nu} = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Differentiability of W

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

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



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

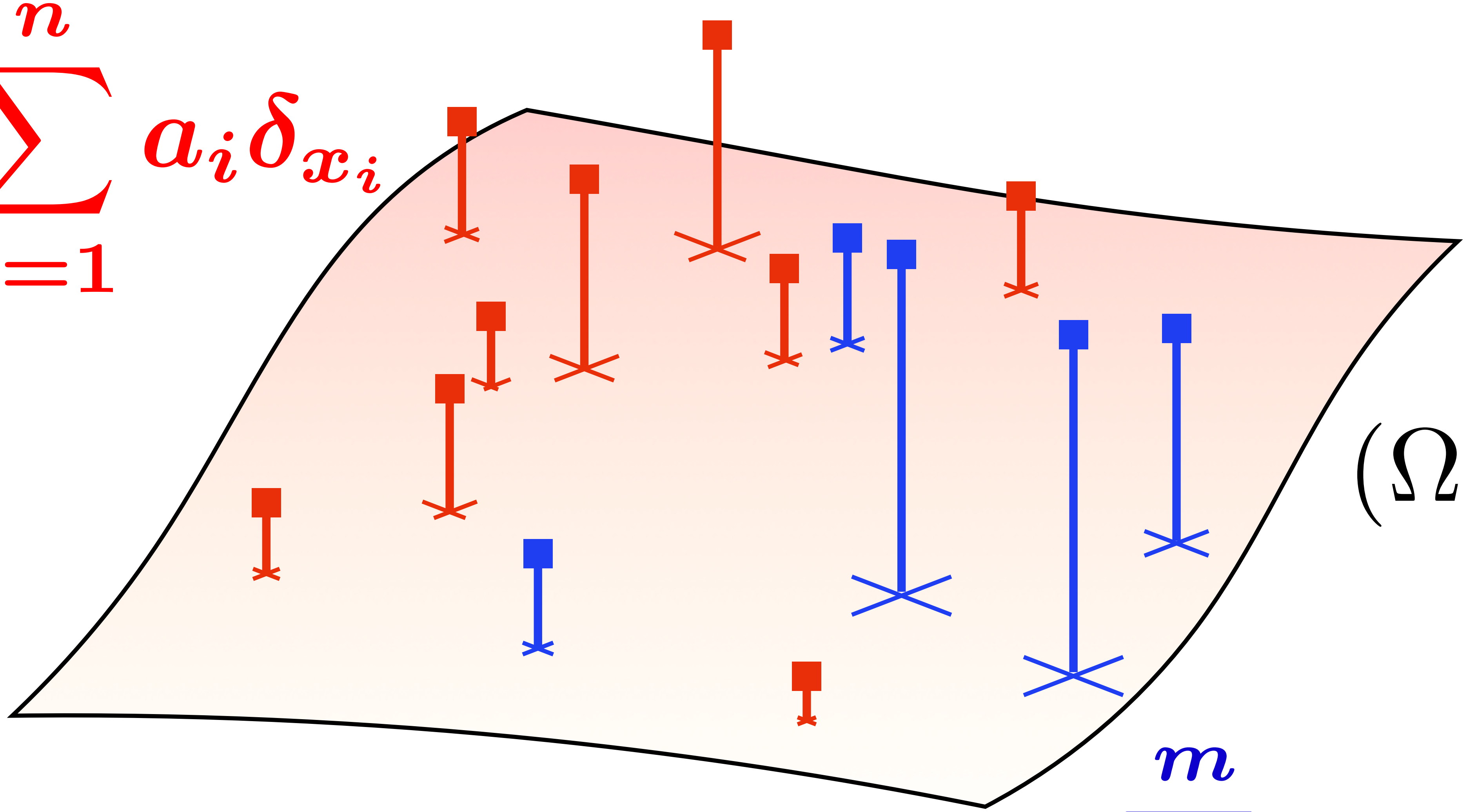
$$\boldsymbol{a} \leftarrow \boldsymbol{a} + \boldsymbol{\Delta a}$$

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

Sinkhorn ⇝ *Differentiability*

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

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



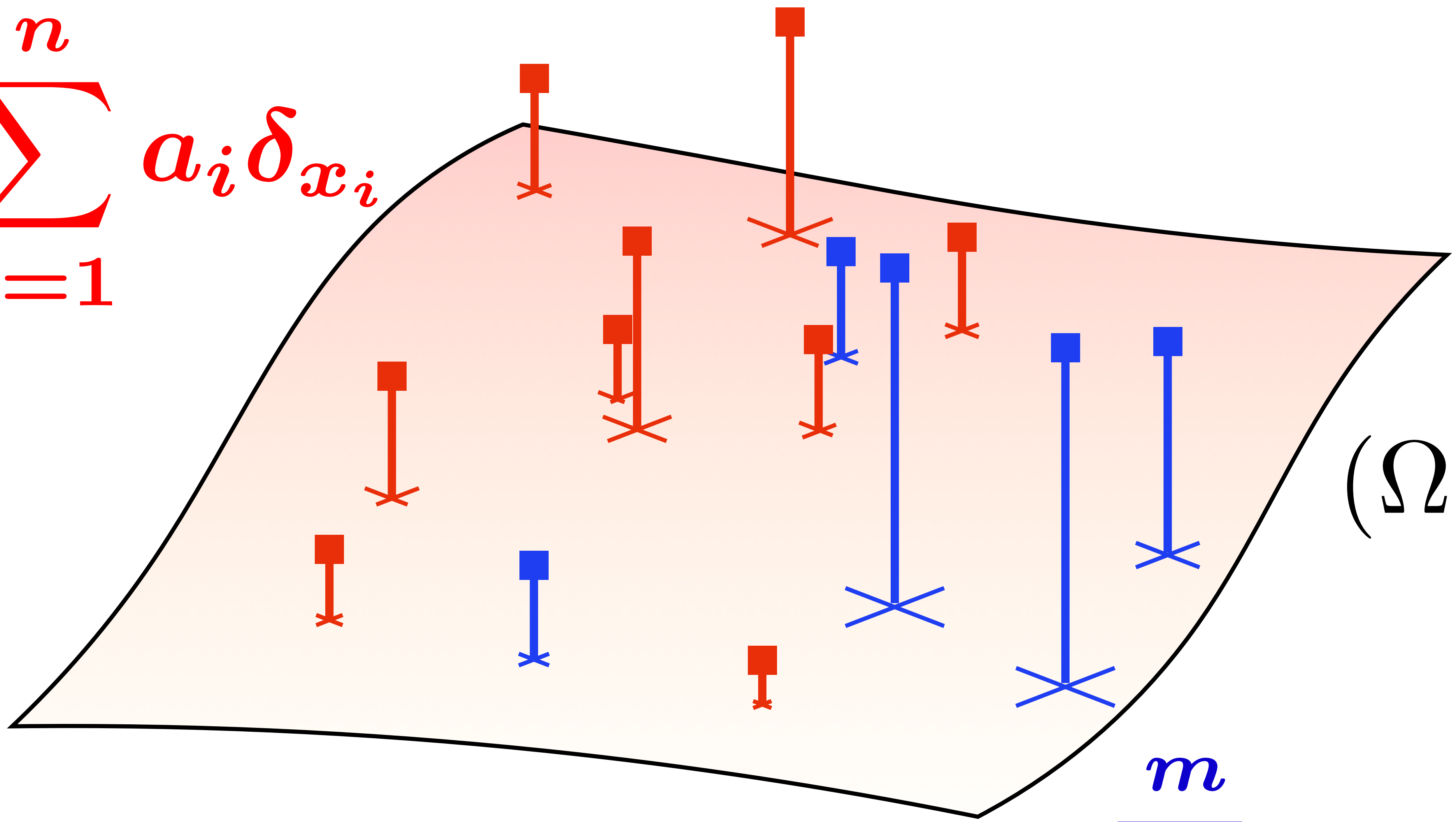
$$(\Omega, \boldsymbol{D})$$

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

Sinkhorn ⇢ *Differentiability*

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

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



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

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

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

How to decrease W? change weights

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n,\boldsymbol{\beta}\in\mathbb{R}^m\\ \boldsymbol{\alpha_i}+\boldsymbol{\beta_j}\leq D(\boldsymbol{x_i},\boldsymbol{y_j})^p}} \alpha^T\boldsymbol{a} + \beta^T\boldsymbol{b}$$

DUAL

Prop. $W(\boldsymbol{\mu},\boldsymbol{\nu})$ is convex w.r.t. $\boldsymbol{a}$, $\partial_{\boldsymbol{a}}W = \alpha^\star$

Prop. $W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu})$ is convex and **differentiable** w.r.t. $\boldsymbol{a}$, $\nabla_{\boldsymbol{a}}W_\gamma = \gamma\log\boldsymbol{u}$

How to decrease W? change locations

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}\\ \boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}} \langle \boldsymbol{P}, \mathbf{1}_n\mathbf{1}_d^T\boldsymbol{X}^2 + \boldsymbol{Y}^{2T}\mathbf{1}_d\mathbf{1}_m - 2\boldsymbol{X}^T\boldsymbol{Y}\rangle$$

PRIMAL

Prop. $p=2, \Omega=\mathbb{R}^d$. $W(\boldsymbol{\mu},\boldsymbol{\nu})$ decreases if

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

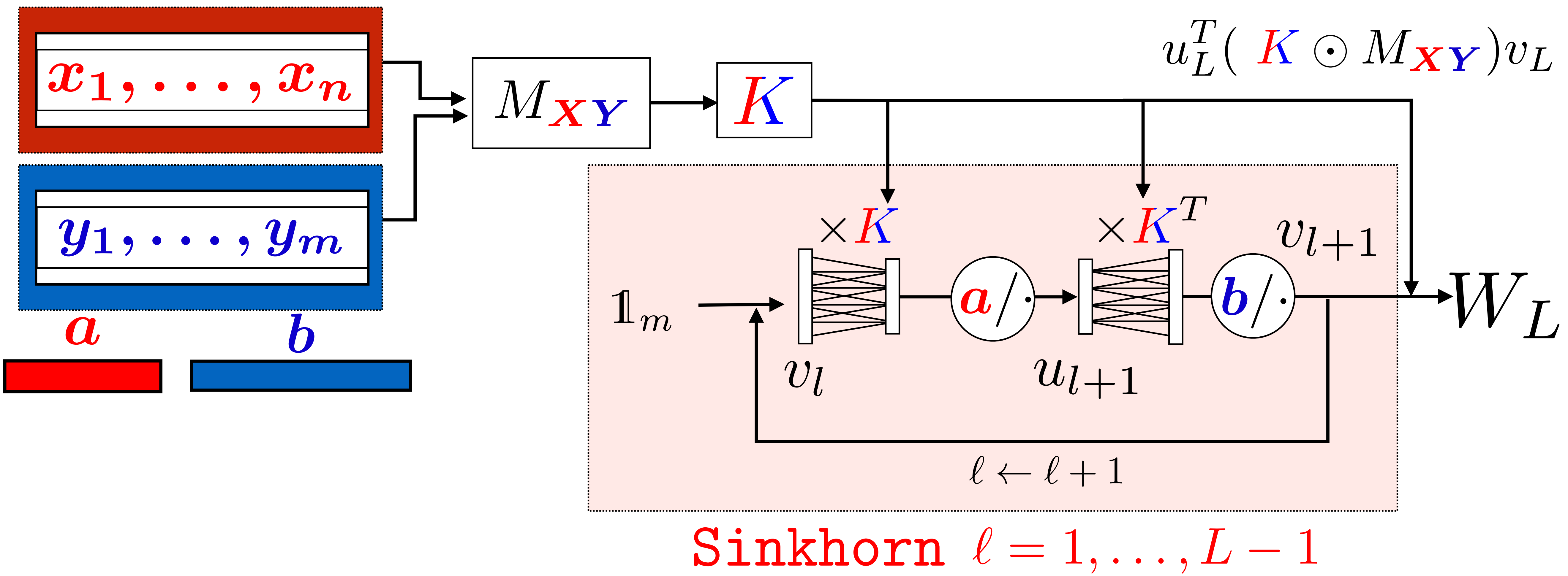
Prop. $p=2, \Omega=\mathbb{R}^d$. $W_\gamma(\boldsymbol{\mu},\boldsymbol{\nu})$ is **differentiable** w.r.t. $\boldsymbol{X}$, with

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

Sinkhorn: A Programmer View

Def. For $L \geq 1$, define

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



Sinkhorn: A Programmer View

Def. For $L \geq 1$, define

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

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

The Programmer's Way

Def. For $L \geq 1$, define

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

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

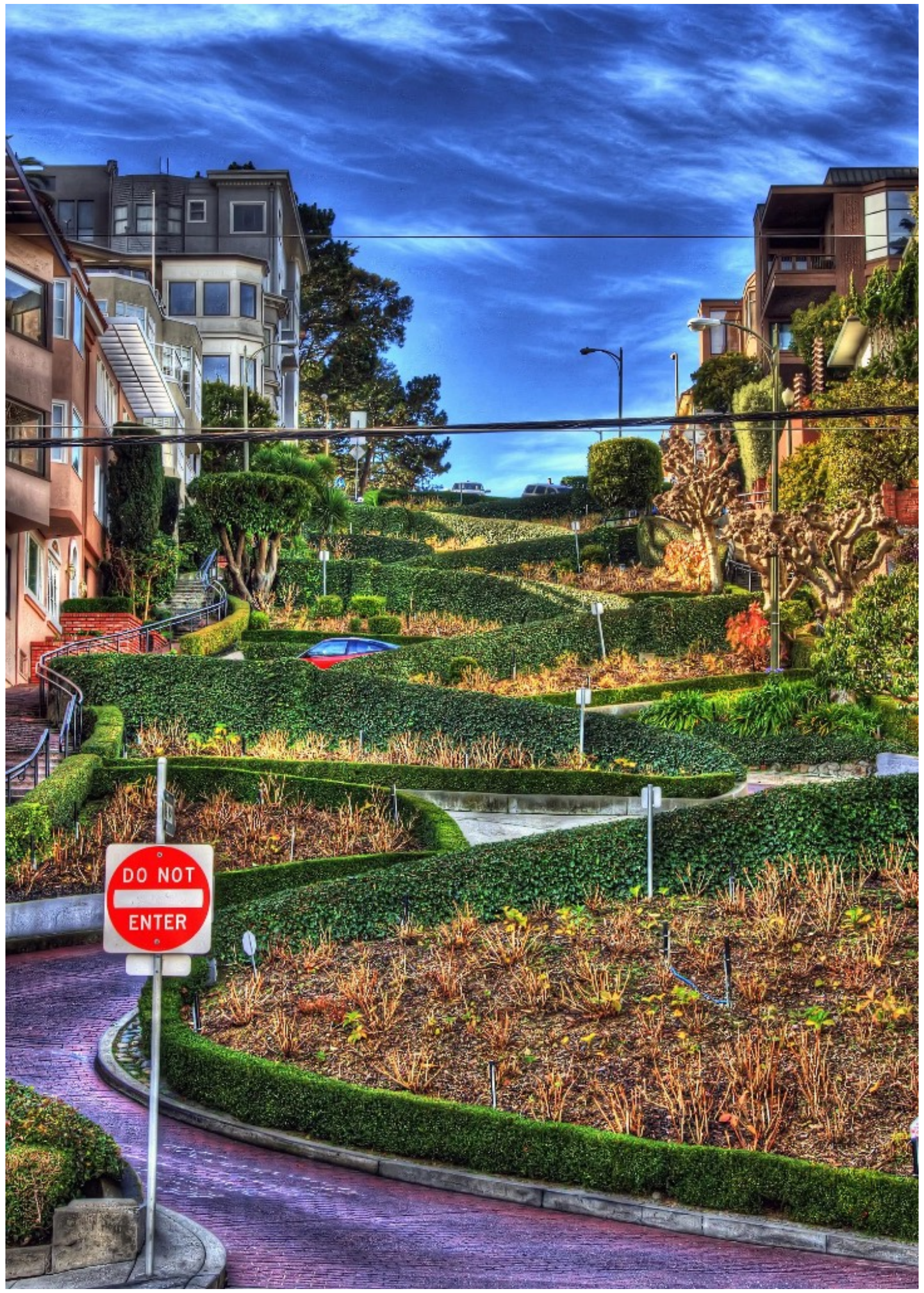
$\boldsymbol{v_0} = \mathbf{1}_m; l \geq 0, \boldsymbol{u_l} \stackrel{\text{def}}{=} \boldsymbol{a}/K\boldsymbol{v_l}, \boldsymbol{v_{l+1}} \stackrel{\text{def}}{=} \boldsymbol{b}/K^T\boldsymbol{u_l}.$

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

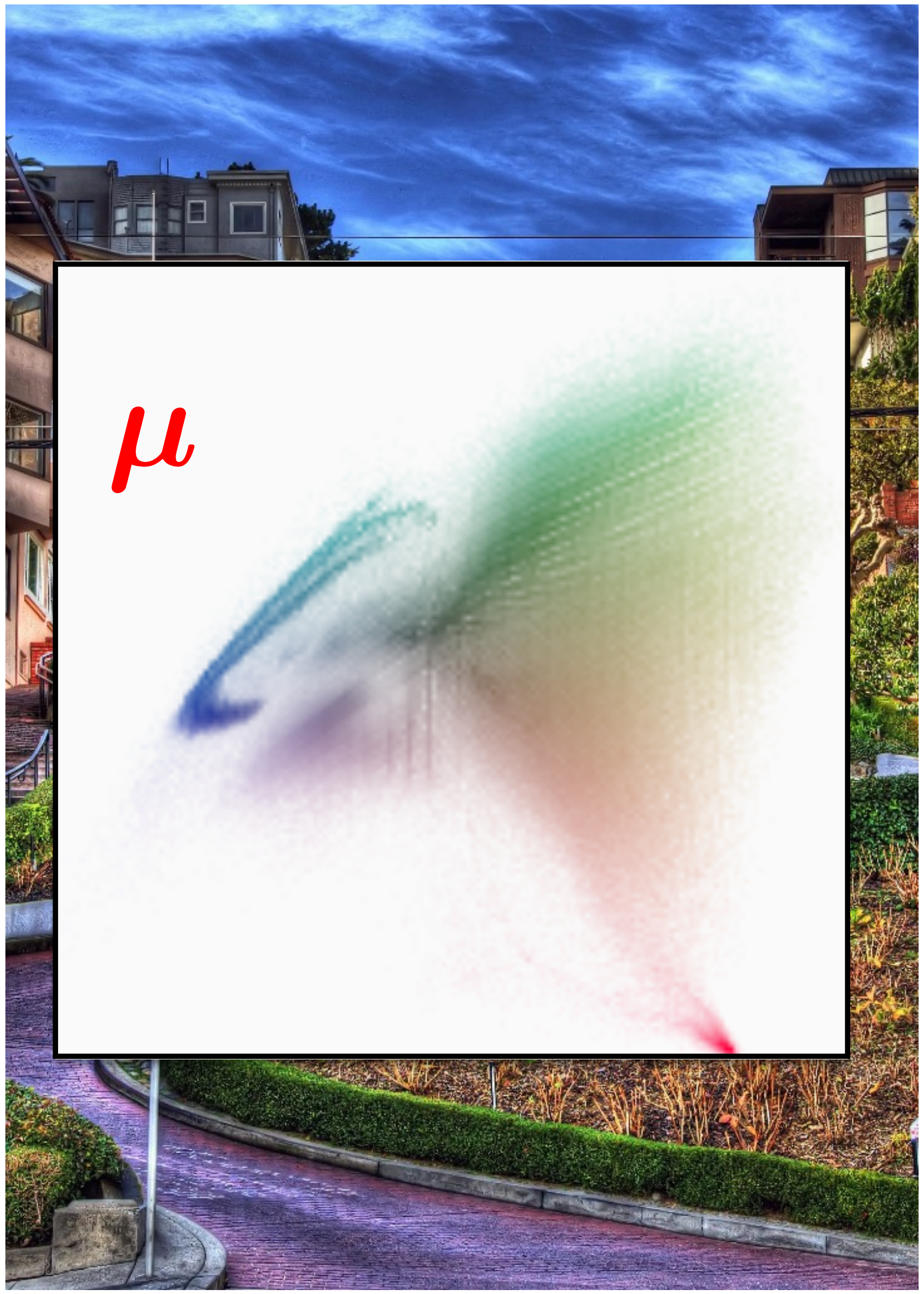
3. Applications

- Wasserstein distances for retrieval
- Wasserstein barycenters
- W for unsupervised learning
- W inverse problems
- W to learn parameters and generative models

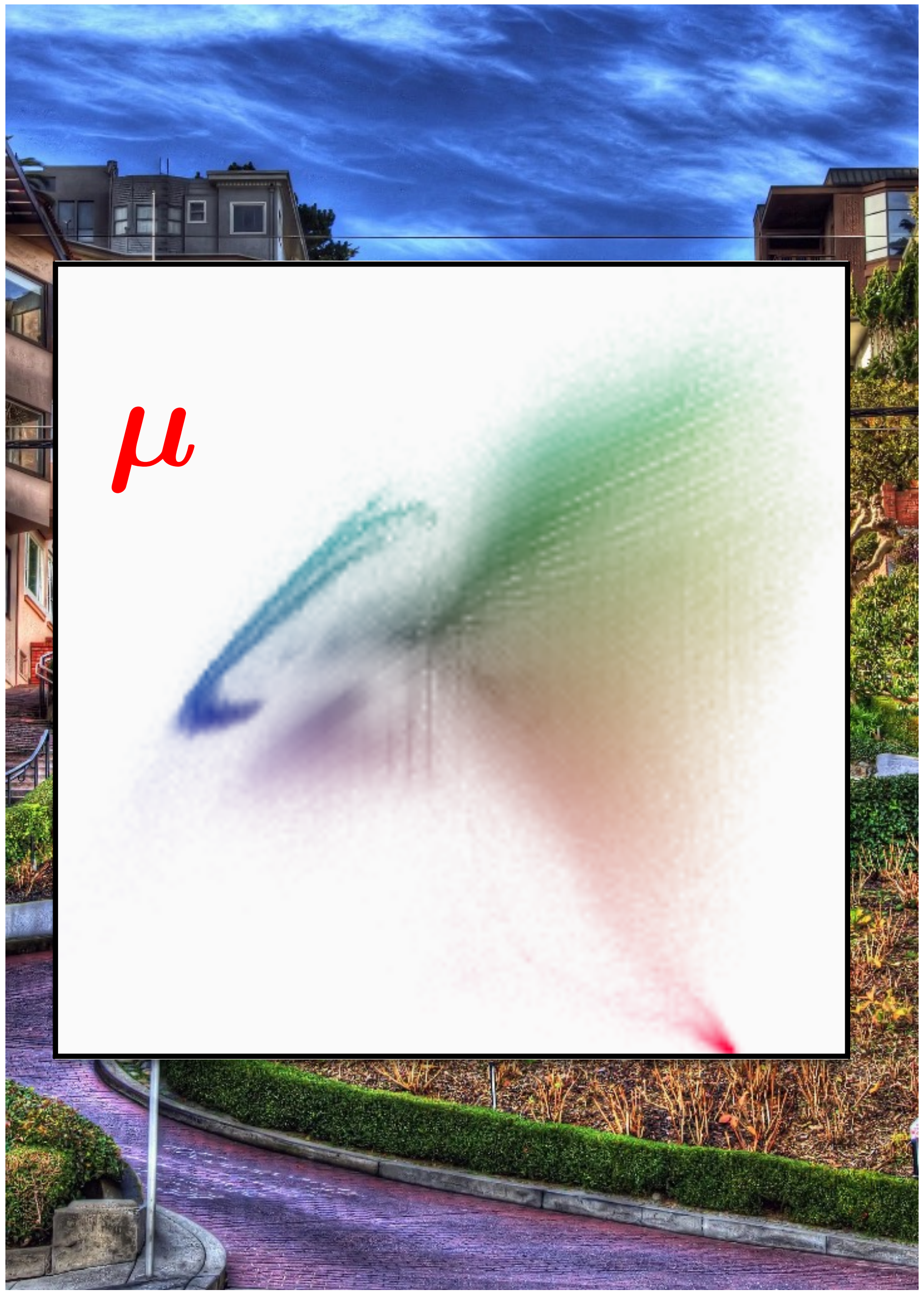
The Earth Mover's Distance



The Earth Mover's Distance

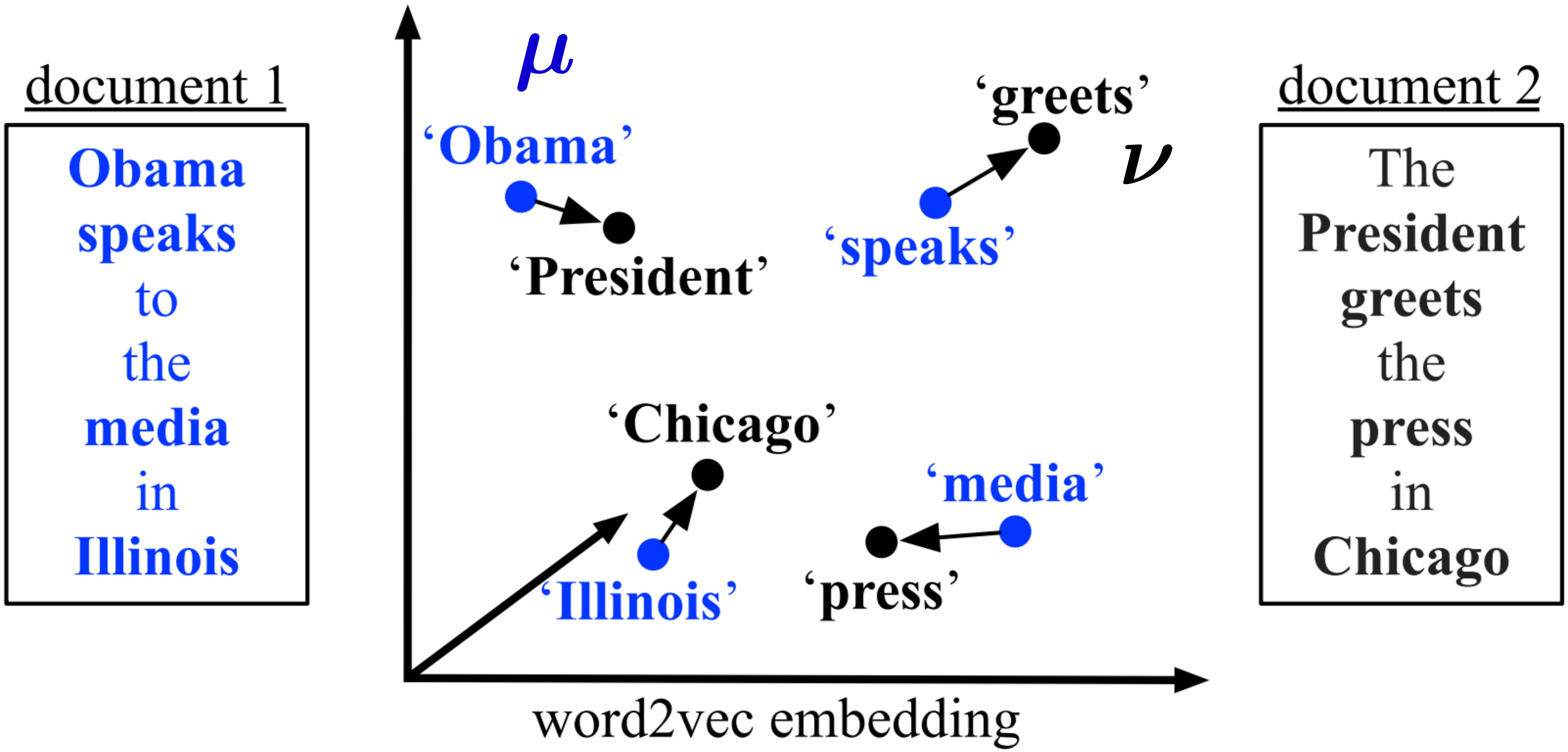


The Earth Mover's Distance



[Rubner'98] $\mathrm{dist}(I_1, I_2) = W_1(\boldsymbol{\mu}, \boldsymbol{\nu})$

The Word Mover's Distance



[Kusner'15] $\mathrm{dist}(D_1, D_2) = W_2(\boldsymbol{\mu}, \boldsymbol{\nu})$

Recall

Up to 2010: OT solvers $\quad W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = ?$

Goal now: use OT as a **loss or fidelity** term

$$\underset{\boldsymbol{\mu} \in \mathcal{P}(\Omega)}{\operatorname{argmin}} \; F(W_p(\boldsymbol{\mu}, \boldsymbol{\nu_1}), W_p(\boldsymbol{\mu}, \boldsymbol{\nu_2}), \ldots, \boldsymbol{\mu}) = ?$$

$$\nabla_{\boldsymbol{\mu}} W_p(\boldsymbol{\mu}, \boldsymbol{\nu_1}) = ?$$

Wassersteinization

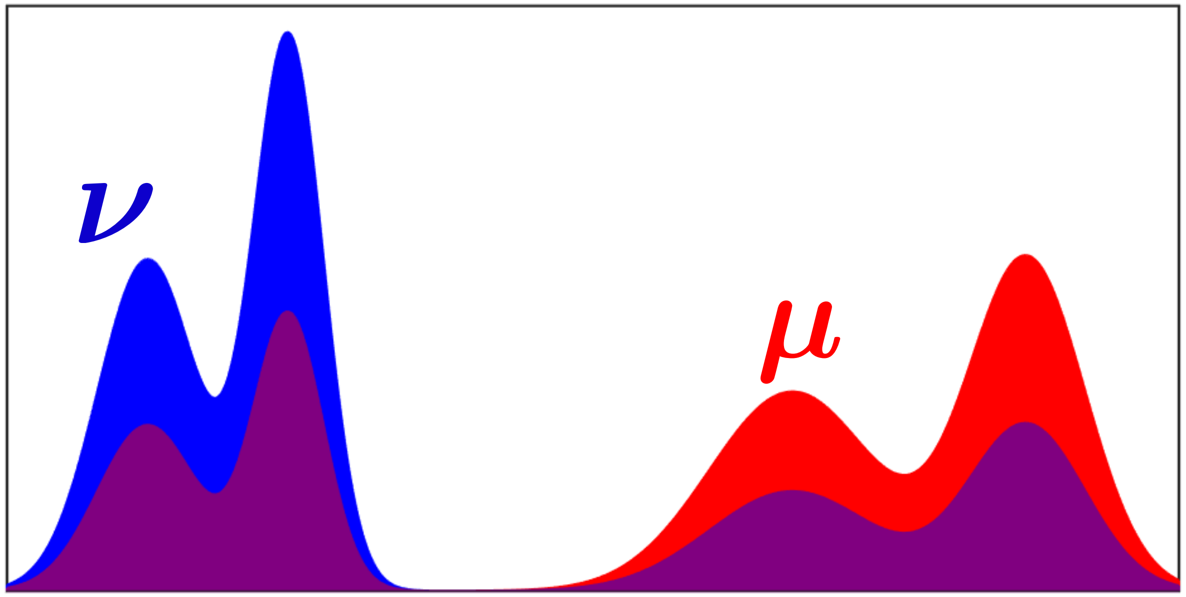
[wos-ur-stahyn-ahy-sey-sh*uh*-n]
noun.

Introduction of optimal transport into an optimization or learning problem.

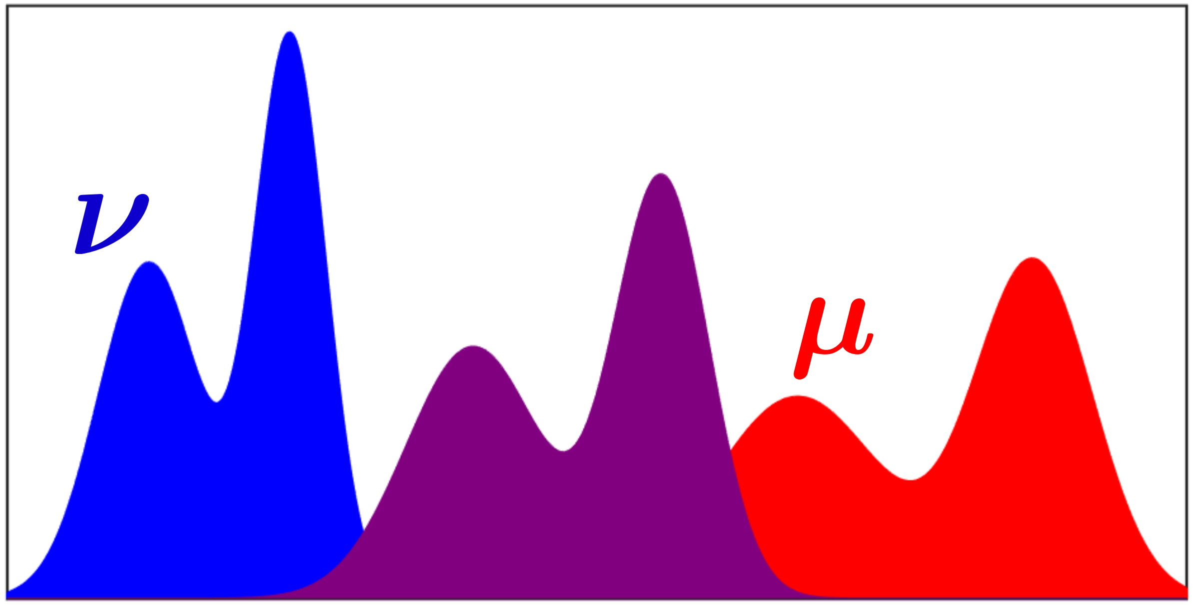
cf. least-squarification, L_1ification, deep-netification, kernelization

Averaging Measures

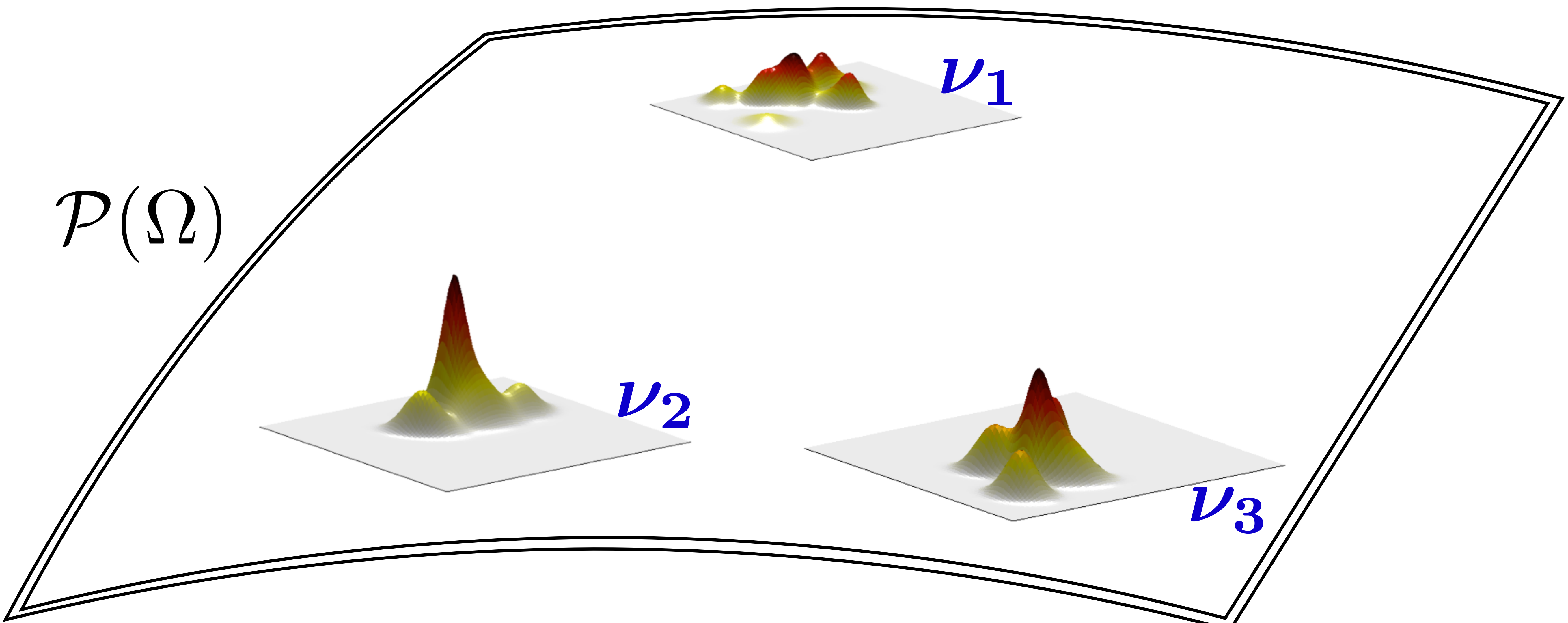
L_2 average



W average

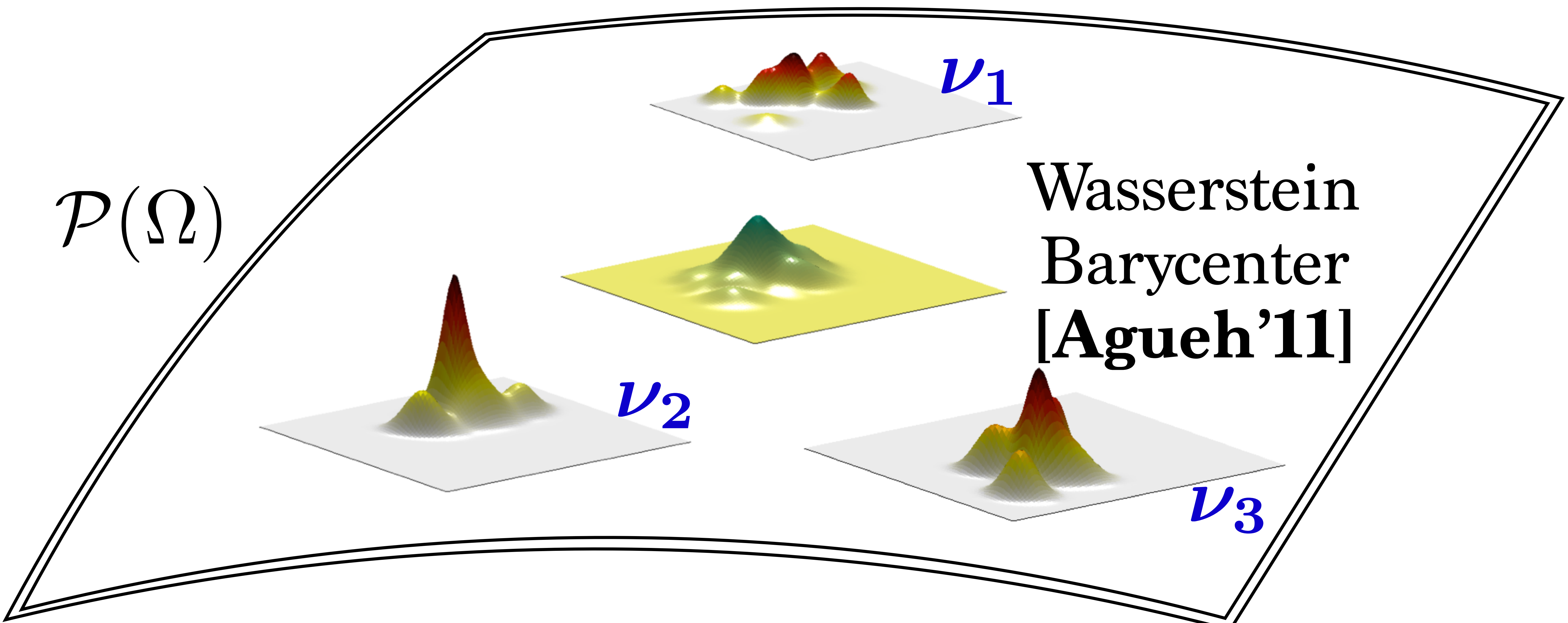


Barycenter for Measures?

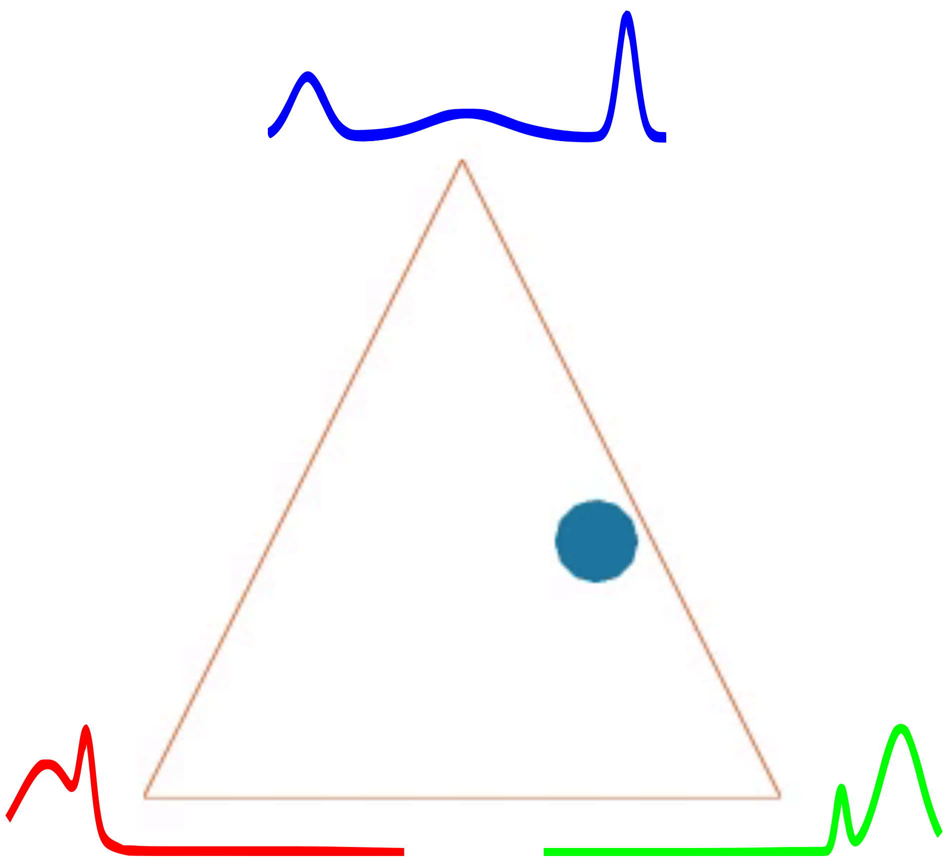


Barycenter for Measures?

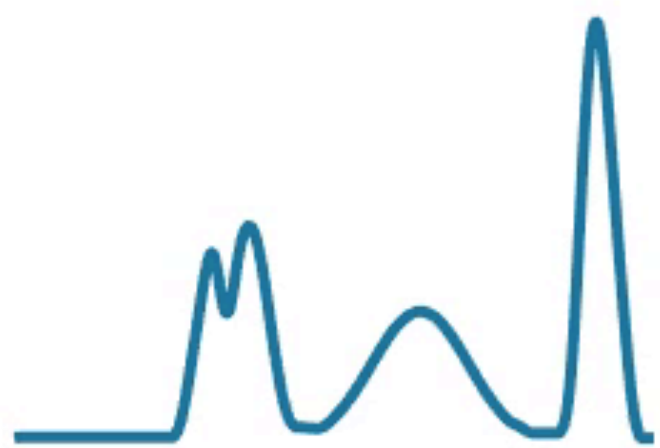
$$\min_{\mu \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$



Barycenter for Measures?



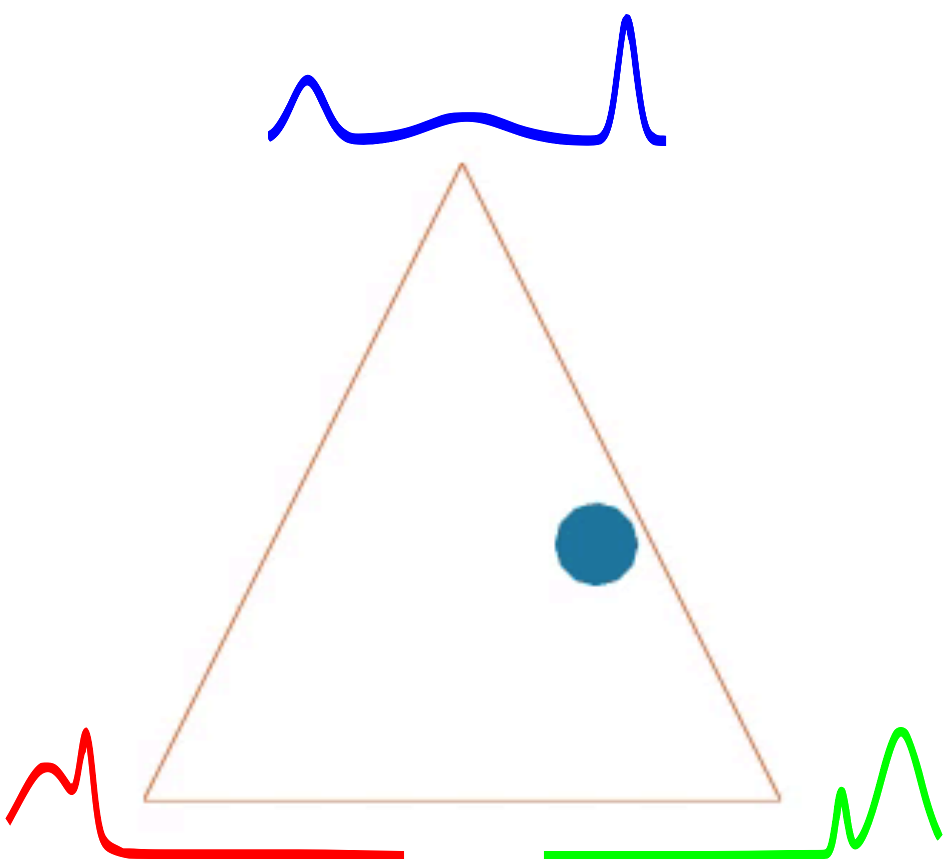
$\lambda \in \Sigma_3$



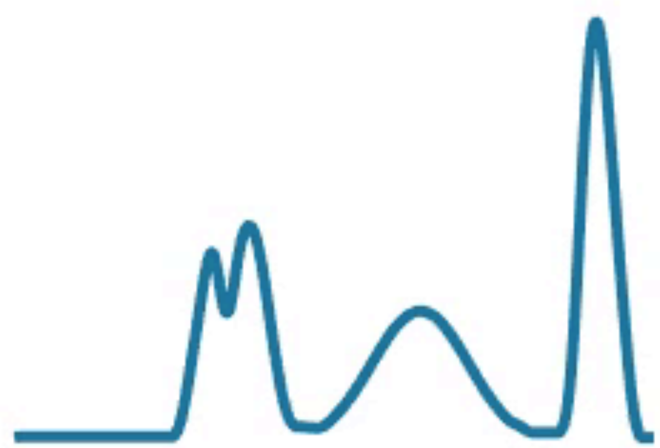
Wasserstein mean

L_2 mean

Barycenter for Measures?



$\lambda \in \Sigma_3$



Wasserstein mean

L_2 mean

Averaging Histograms is a LP

When Ω is a finite metric space defined by M.

$$\min_{\boldsymbol{a}\in\Sigma_n} \sum_i \lambda_i W_M(\boldsymbol{a}, \boldsymbol{b_i})$$

Averaging Histograms is a LP

When Ω is a finite metric space defined by *M*.

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

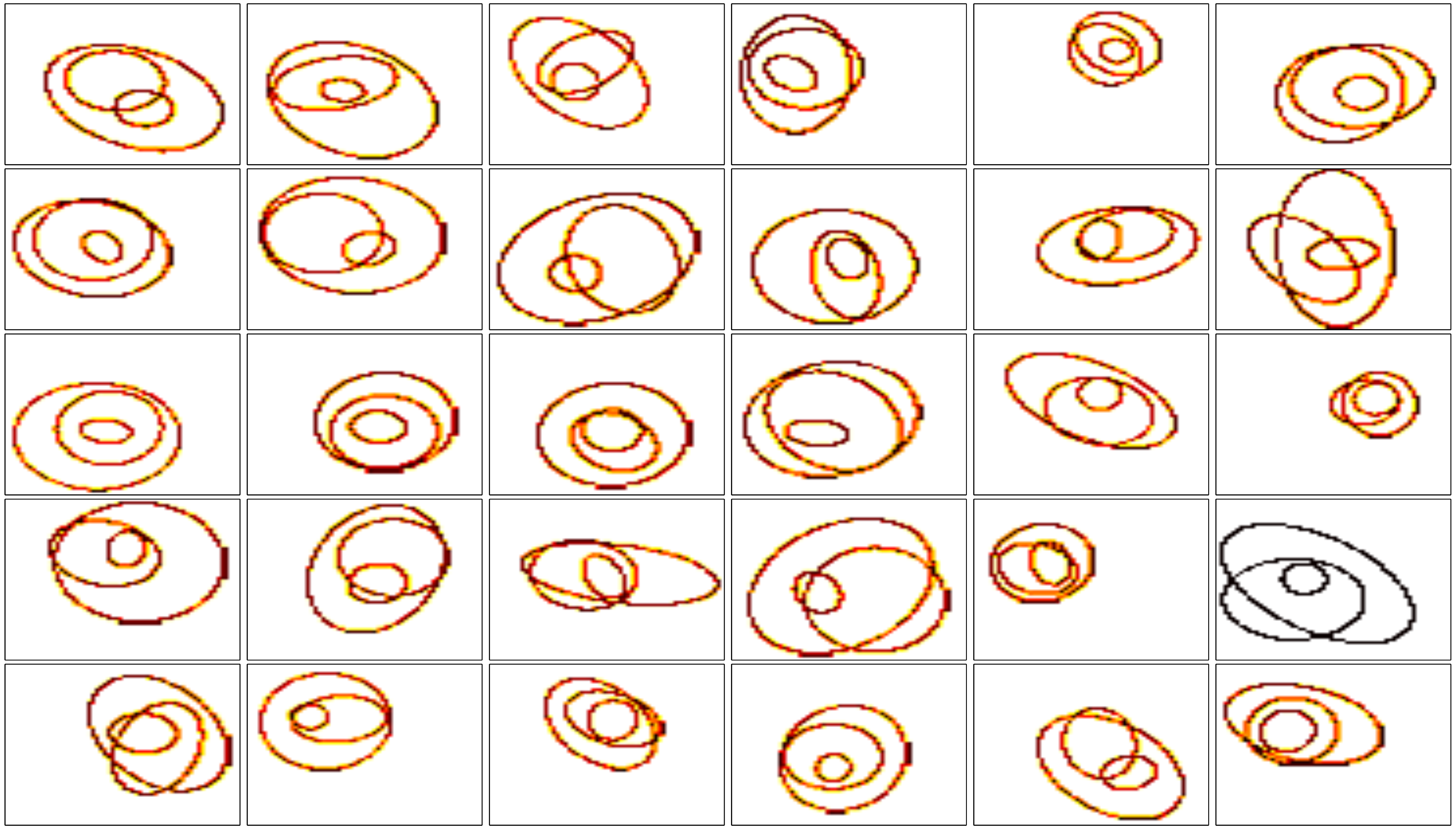
$$\text{s.t. } \boldsymbol{P_i}^T \mathbf{1}_n = \boldsymbol{b_i}, \forall i \leq N,$$

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

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

Primal Descent on Regularized W

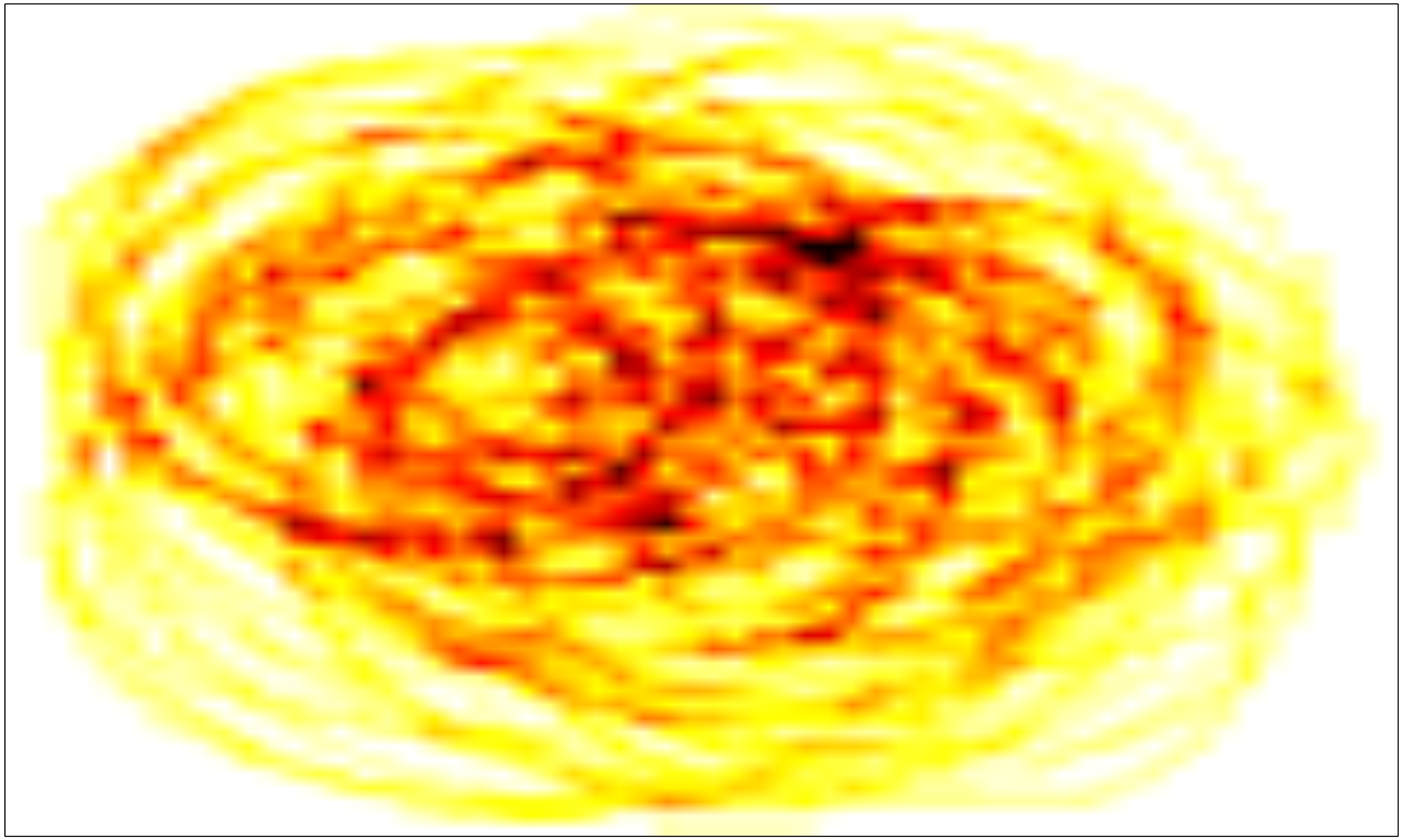
$$\min_{\boldsymbol{a}\in\Sigma_{h\times h}} \sum_{i=1}^{N} \lambda_i W_{\gamma}(\boldsymbol{a}, \boldsymbol{b_i})$$



[Cuturi'14]

Primal Descent on Regularized W

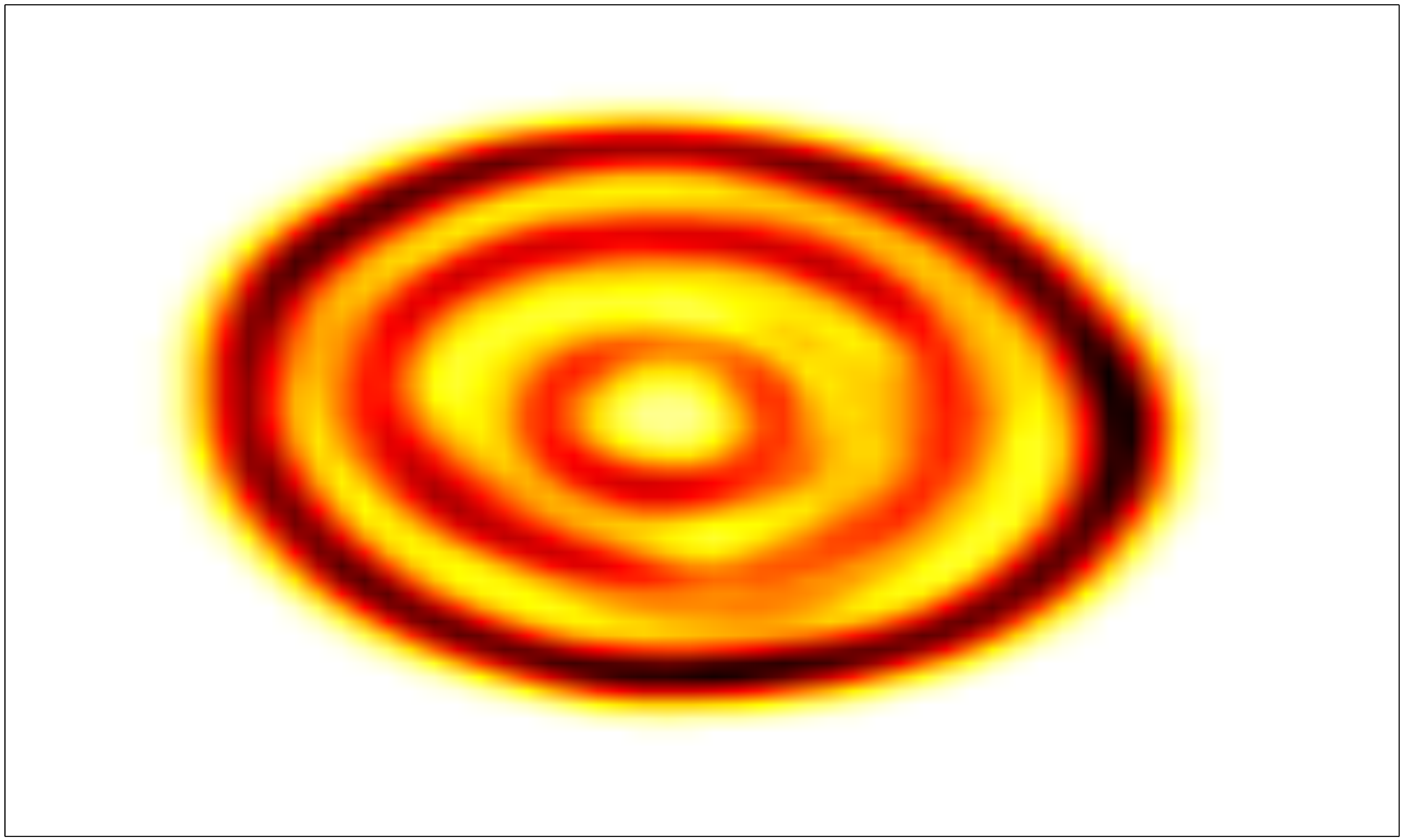
$$\min_{\boldsymbol{a}\in\Sigma_{h\times h}} \sum_{i=1}^{N} \lambda_i W_{\gamma}(\boldsymbol{a}, \boldsymbol{b_i})$$



[Cuturi'14]

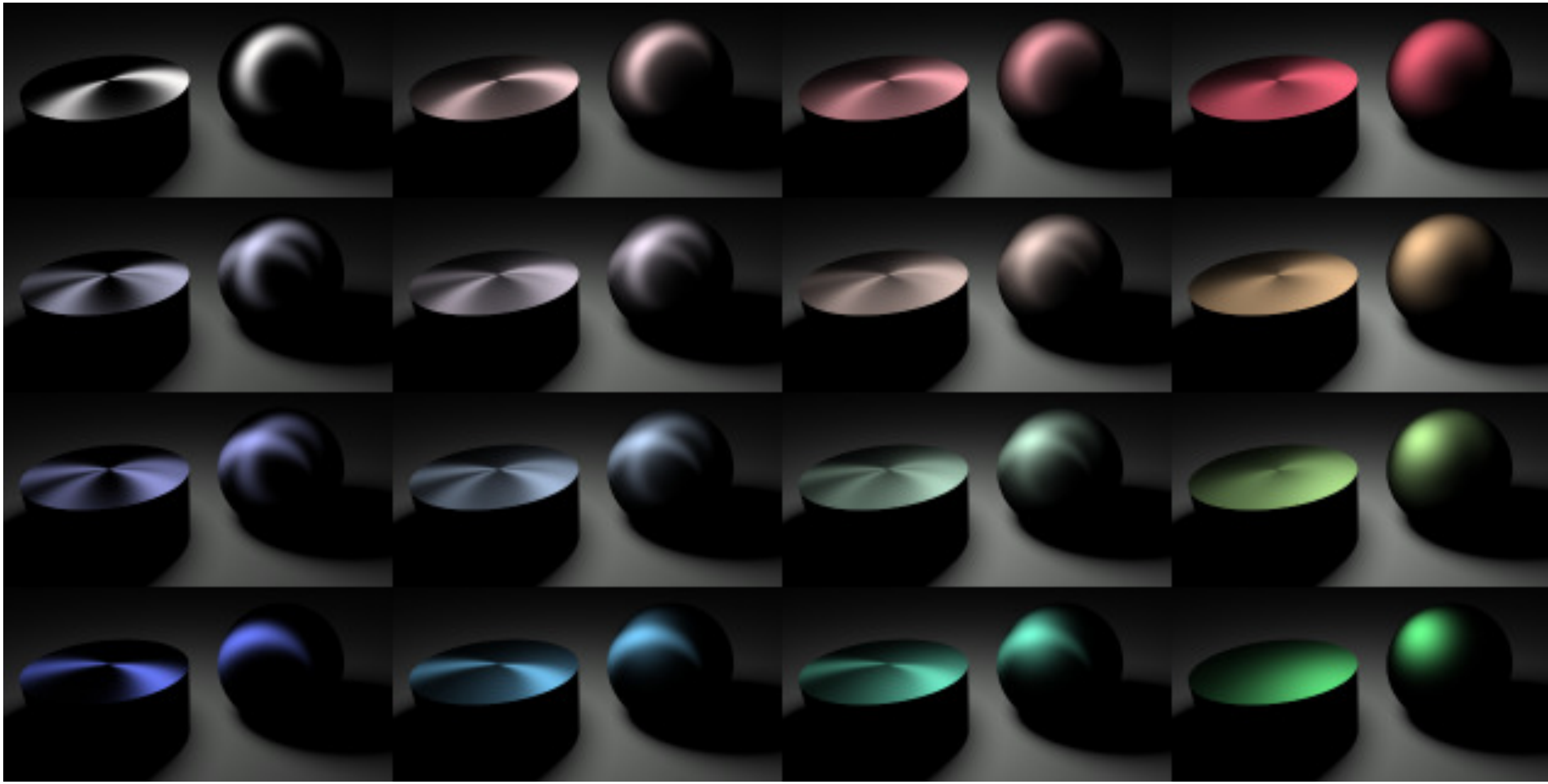
Primal Descent on Regularized W

$$\min_{\boldsymbol{a}\in\Sigma_{h\times h}} \sum_{i=1}^{N} \lambda_i W_{\gamma}(\boldsymbol{a}, \boldsymbol{b_i})$$



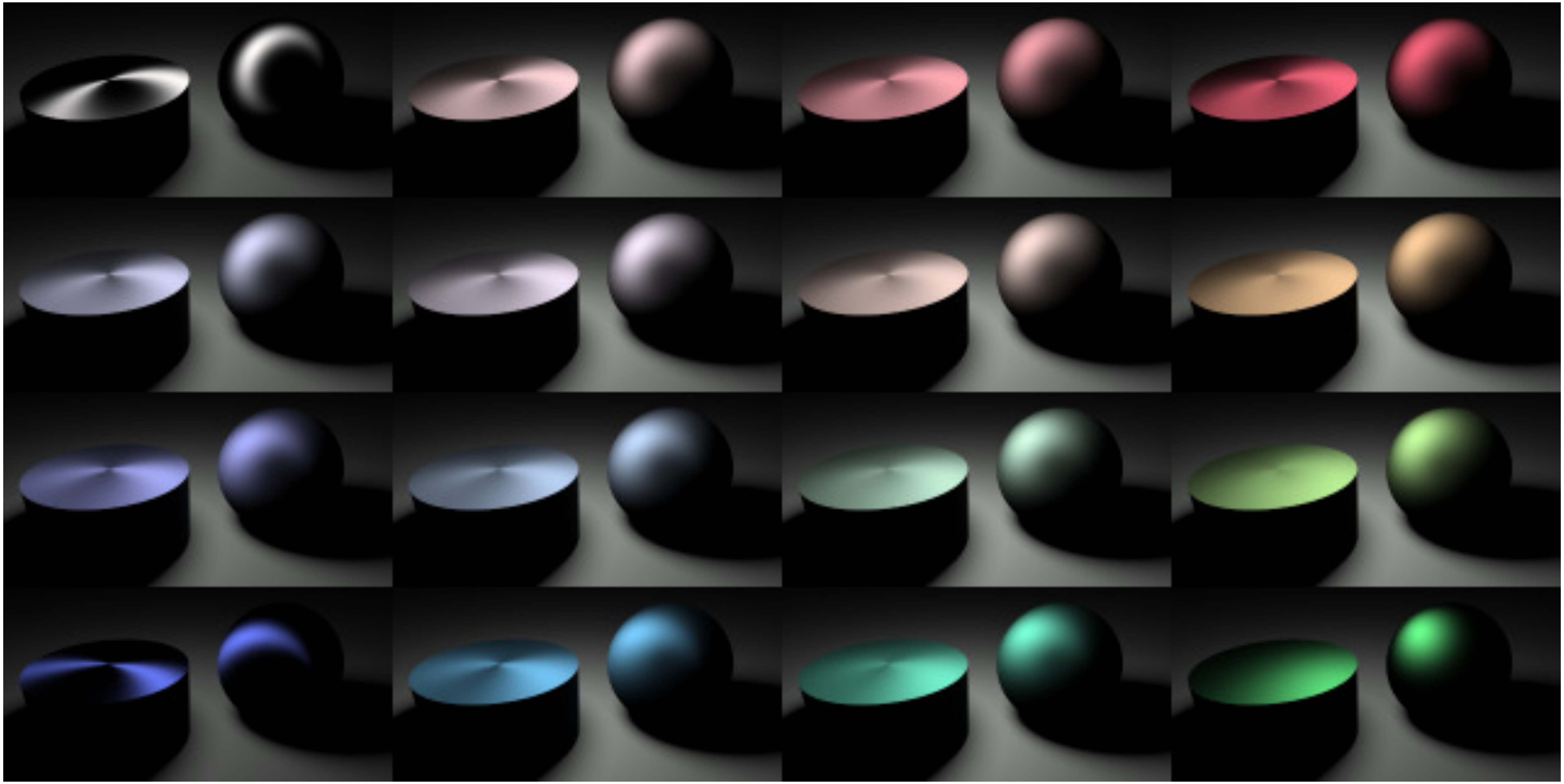
[Cuturi'14]

Applications in Imaging



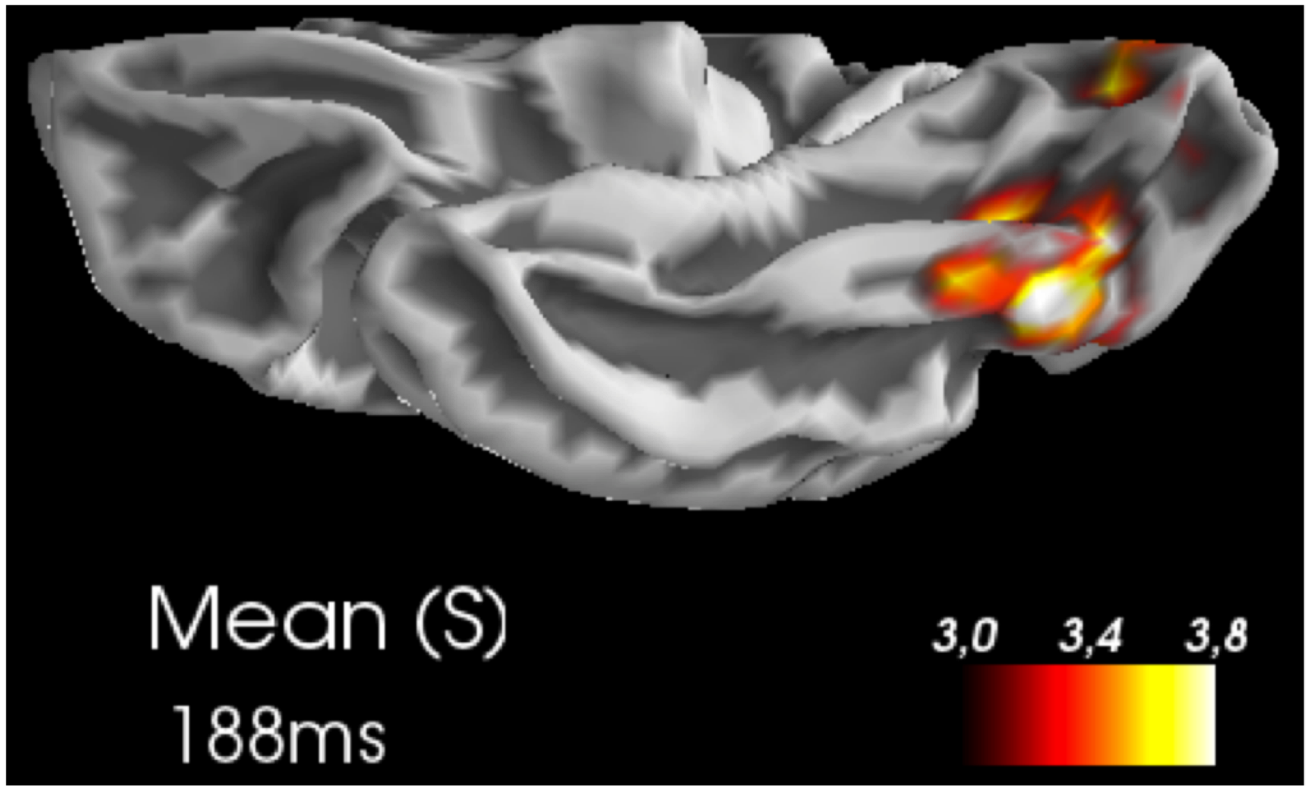
[Solomon'15]

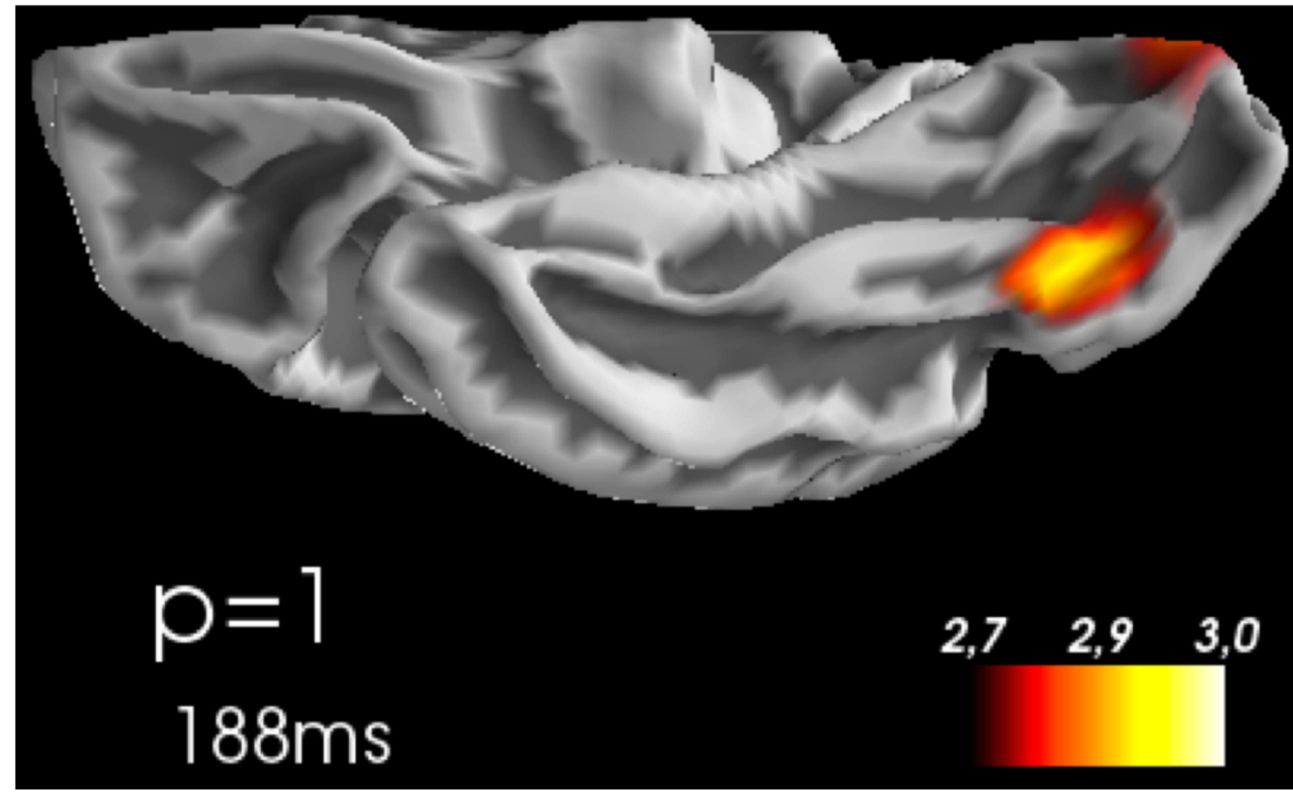
Applications in Imaging



[Solomon'15]

Applications: Brain Imaging





Extension to non-normalized data!
Applied to MEG and fMRI.

[Gramfort'16]

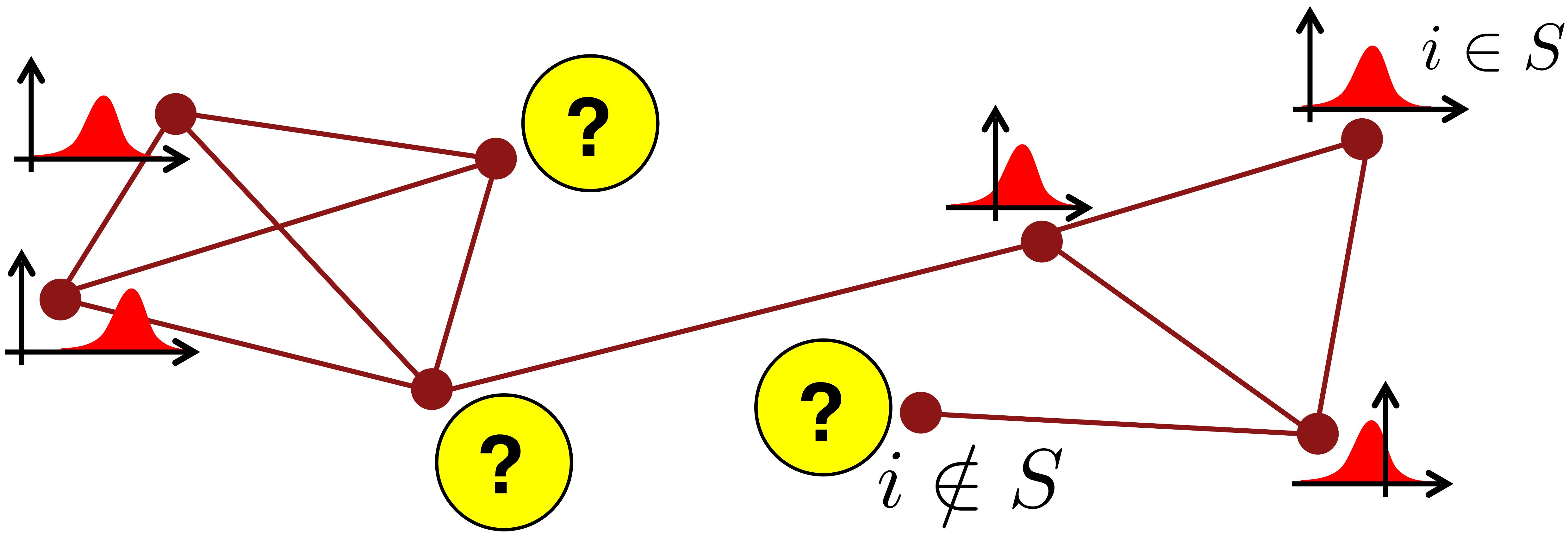
Wasserstein Posterior (WASP)

Merge Bayesian subset posteriors.

1. Split data into J subsets $S_1, \ldots, S_J$
2. Distribute to J machines.
3. In parallel, sample from $\{p(\theta)|S_i\}_{i=1}^{J}$ using MCMC.
4. Aggregate using Wasserstein barycenters

[Srivastava'15]

Wasserstein Propagation

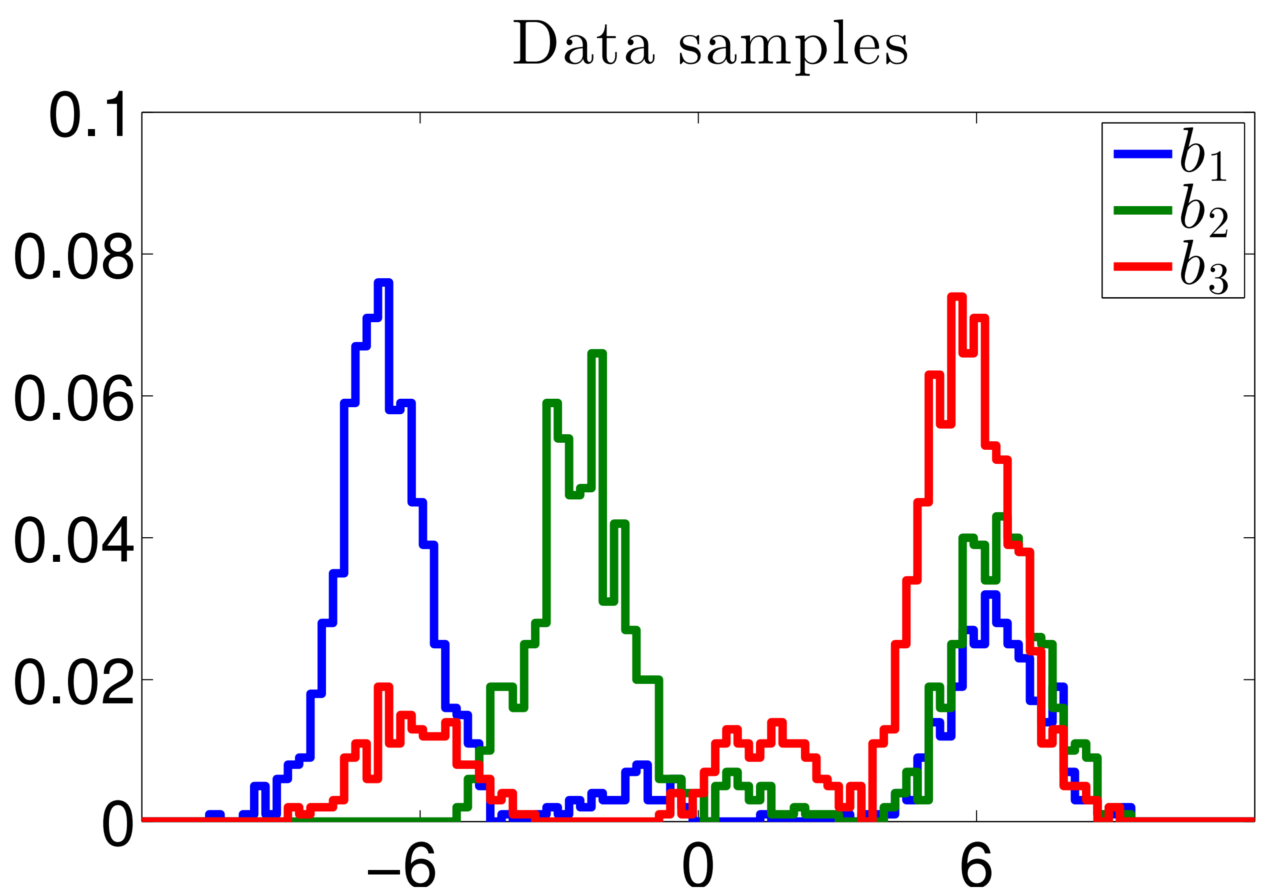


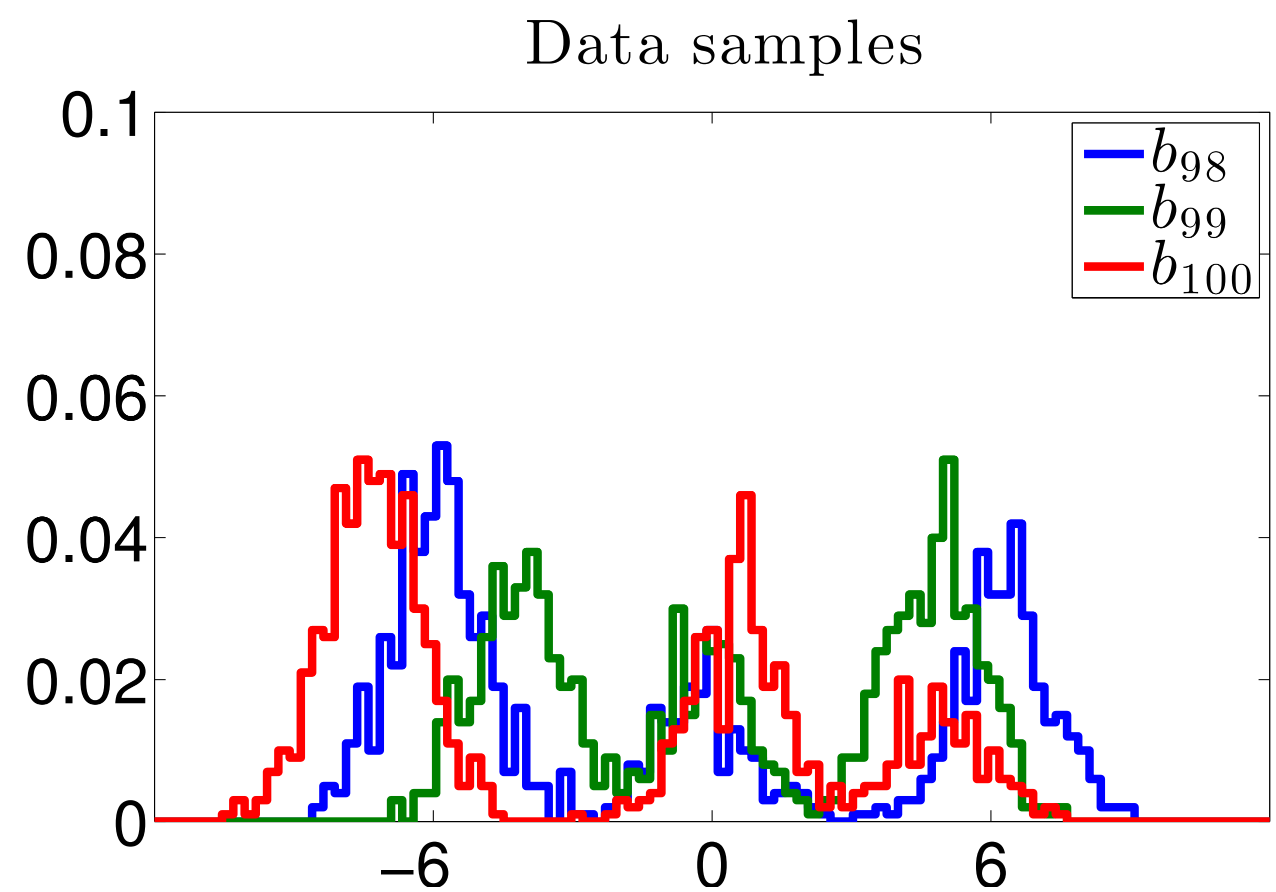
$$\min_{\substack{\mu_i \in \mathcal{P}(\Omega) \\ \mu_i \text{ fixed for } i \in S}} \sum_{(e_1, e_2) \in E} W_2^2(\mu_{e_1}, \mu_{e_2})$$

[Solomon'14]

Dictionary Learning

$$\min_{A \in (\Sigma_n)^K, \Lambda \in (\Sigma_K)^N} \sum_{i=1}^{N} W\left(b_i, \sum_{k=1}^{K} \Lambda_k^i a_k\right)$$

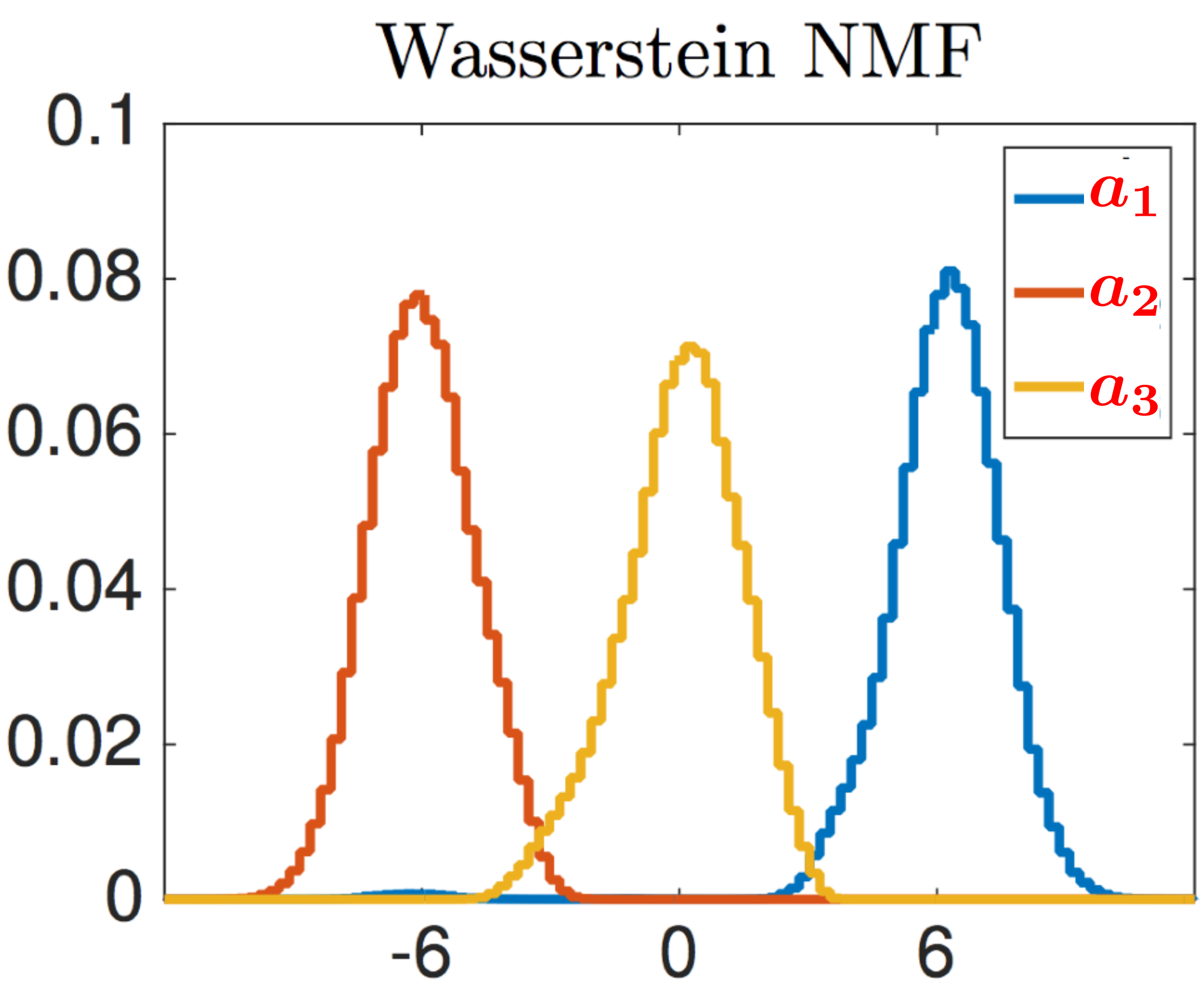


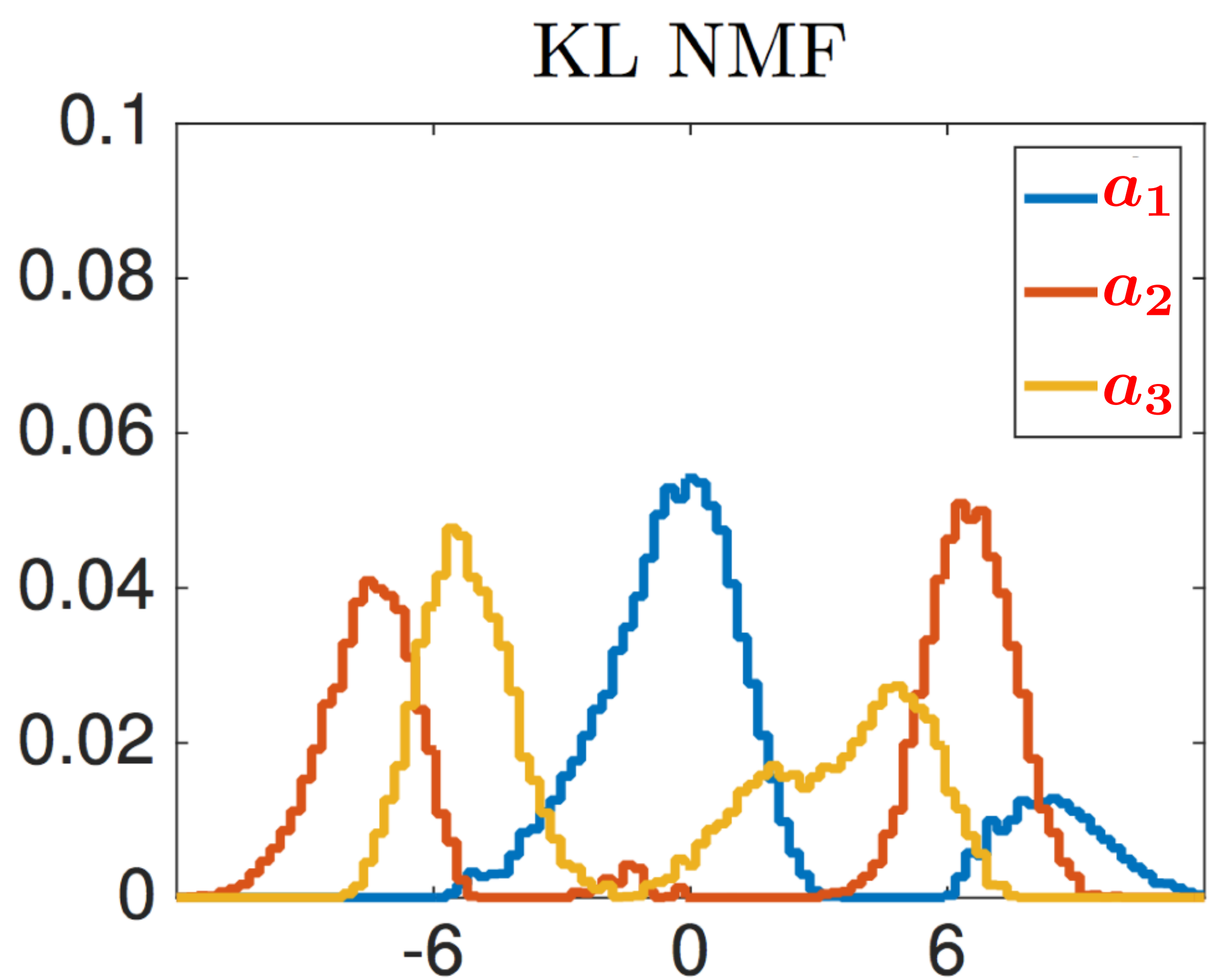


[Sandler'11] [Zen'14] [Rolet'16]

Dictionary Learning

$$\min_{\boldsymbol{A}\in(\Sigma_n)^K,\boldsymbol{\Lambda}\in(\Sigma_K)^N} \sum_{i=1}^{N} W\left(\boldsymbol{b_i}, \sum_{k=1}^{K} \boldsymbol{\Lambda}_{\boldsymbol{k}}^{\boldsymbol{i}} \boldsymbol{a_k}\right)$$





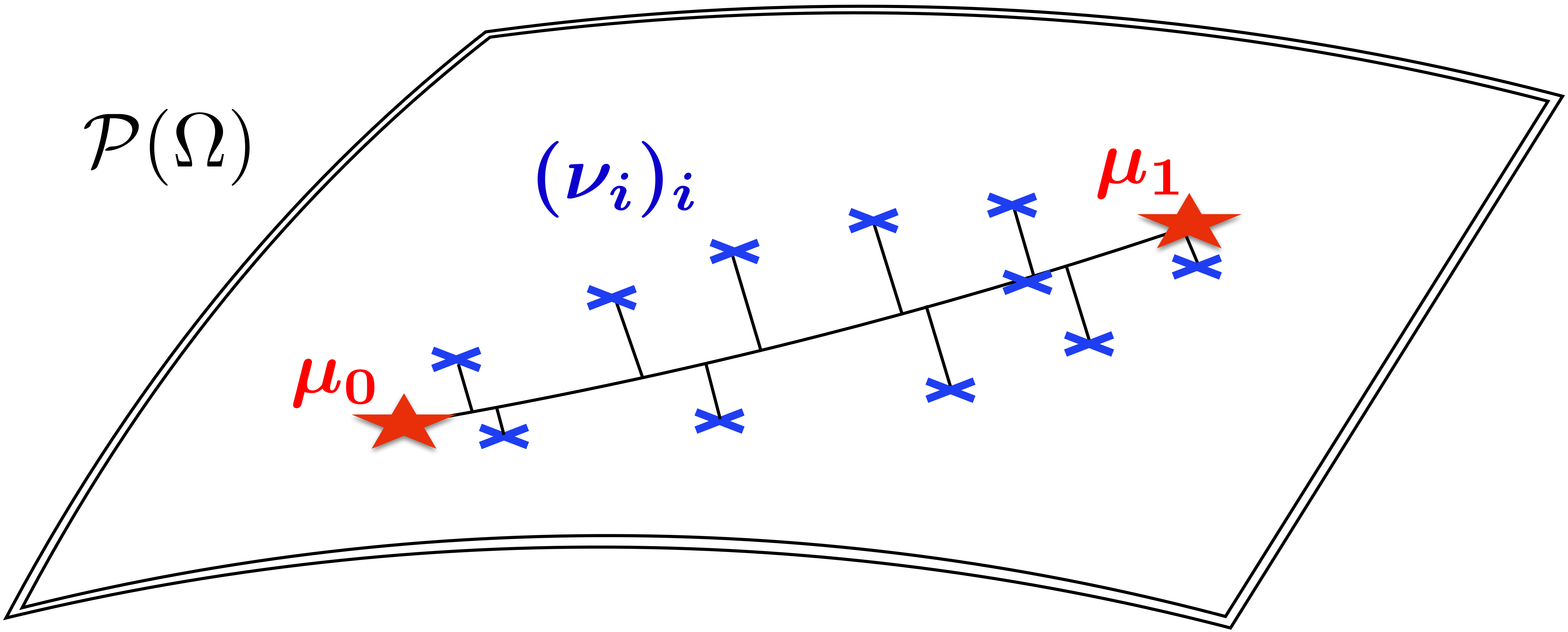
[Sandler'11] [Zen'14] [Rolet'16]

Topic Models

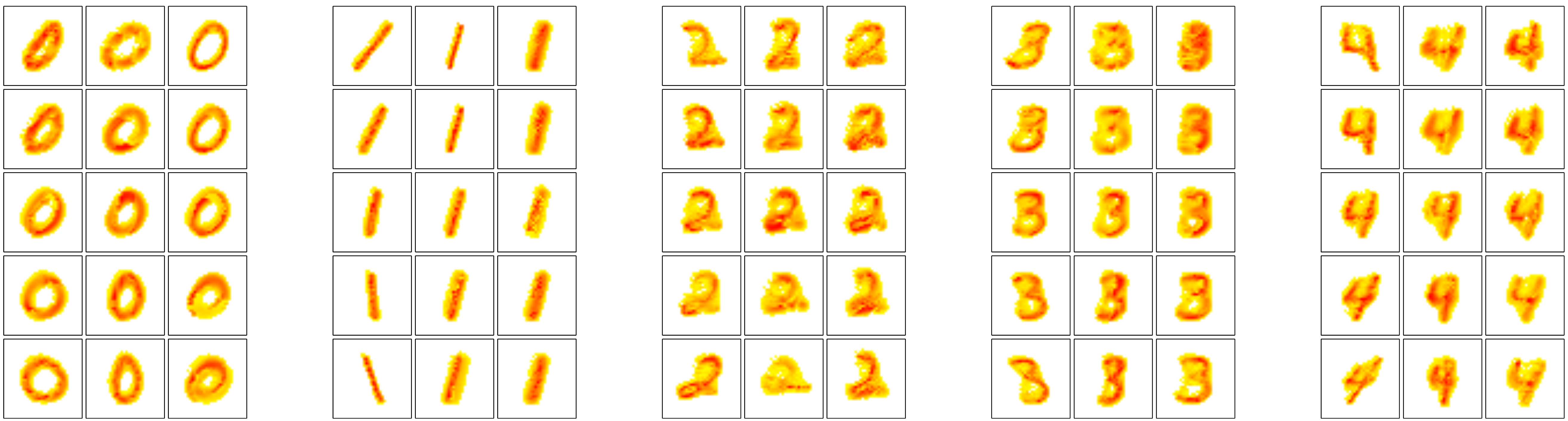
[Rolet'16]

Wasserstein PCA

$$\min_{\boldsymbol{\mu_0},\boldsymbol{\mu_1}} \sum_{i=1}^{N} \min_t W_2^2(\rho^t_{\boldsymbol{\mu_0}\to\boldsymbol{\mu_1}}, \boldsymbol{\nu_i})$$



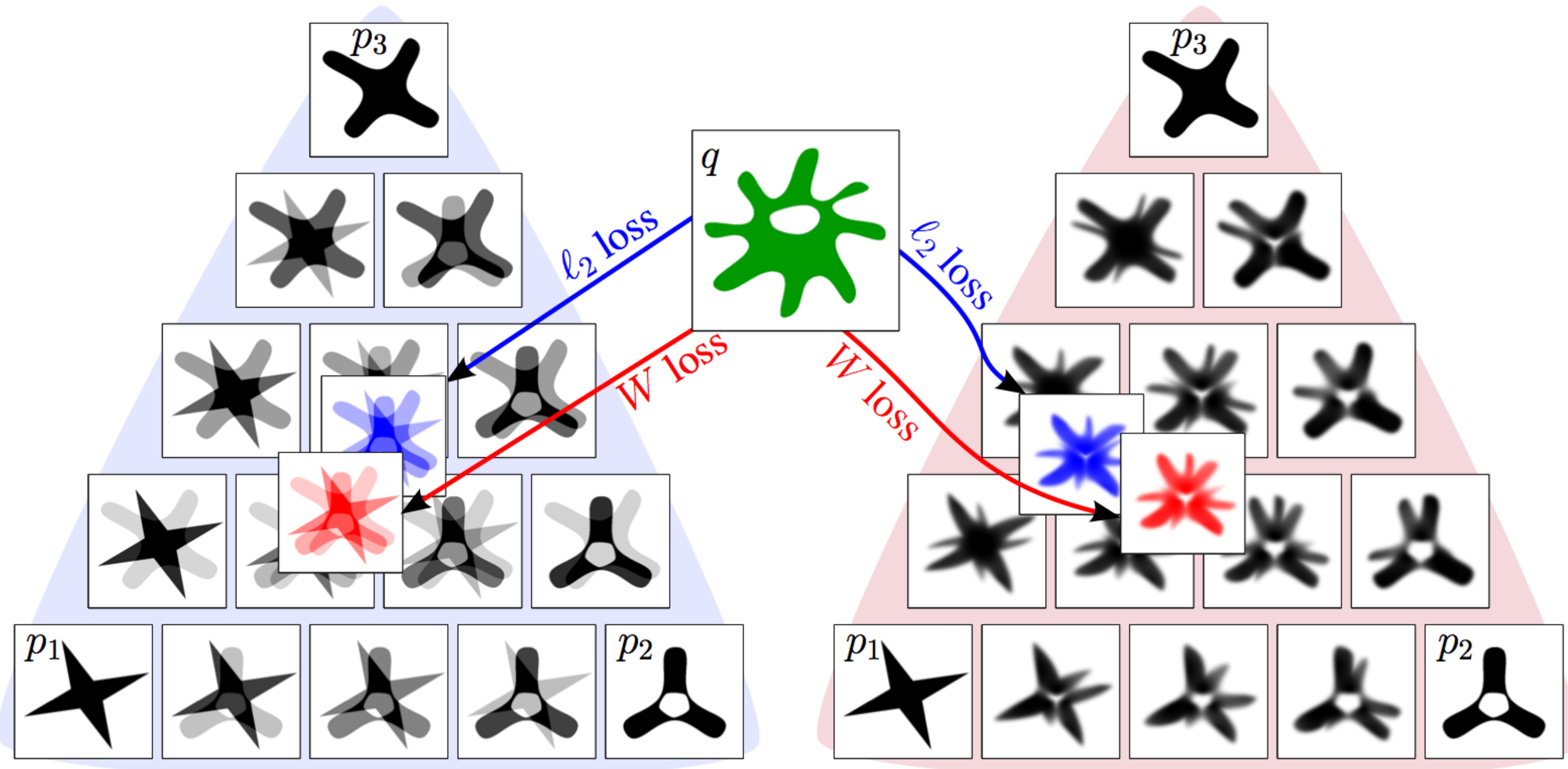
Generalized Principal Geodesics



For each digit, 1,000 MNIST images

[Seguy'15

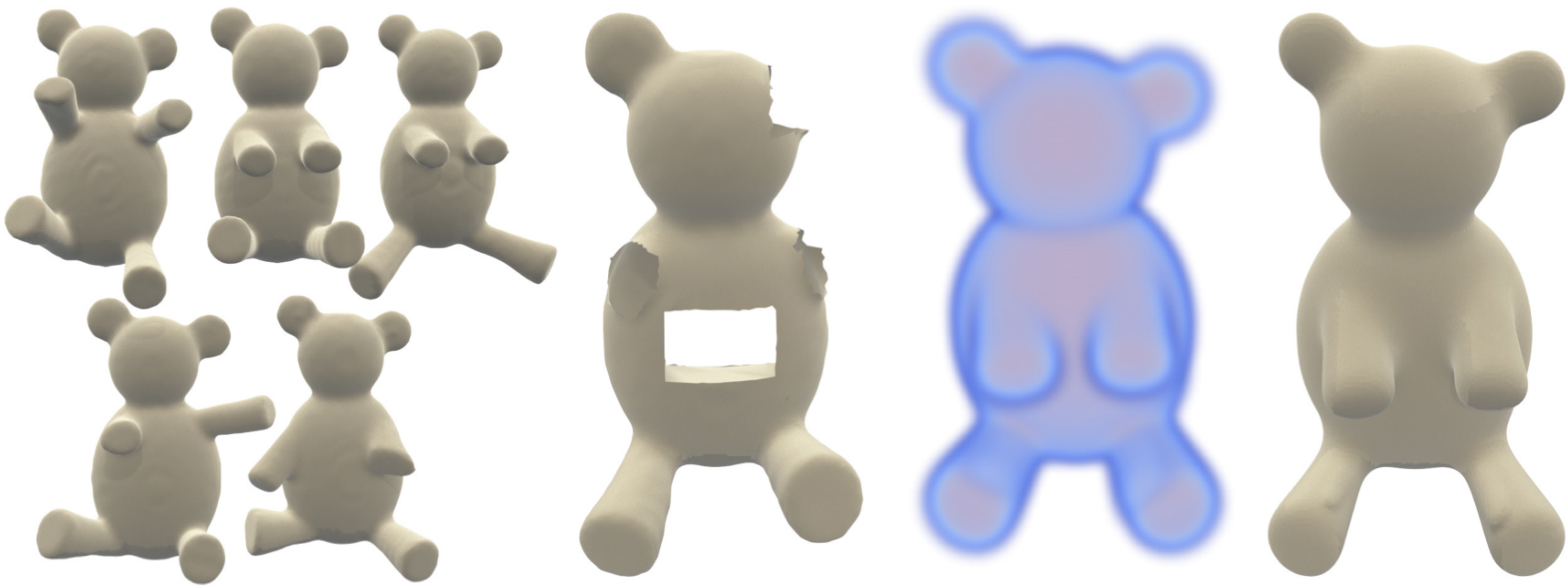
Wasserstein Inverse Problems



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

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

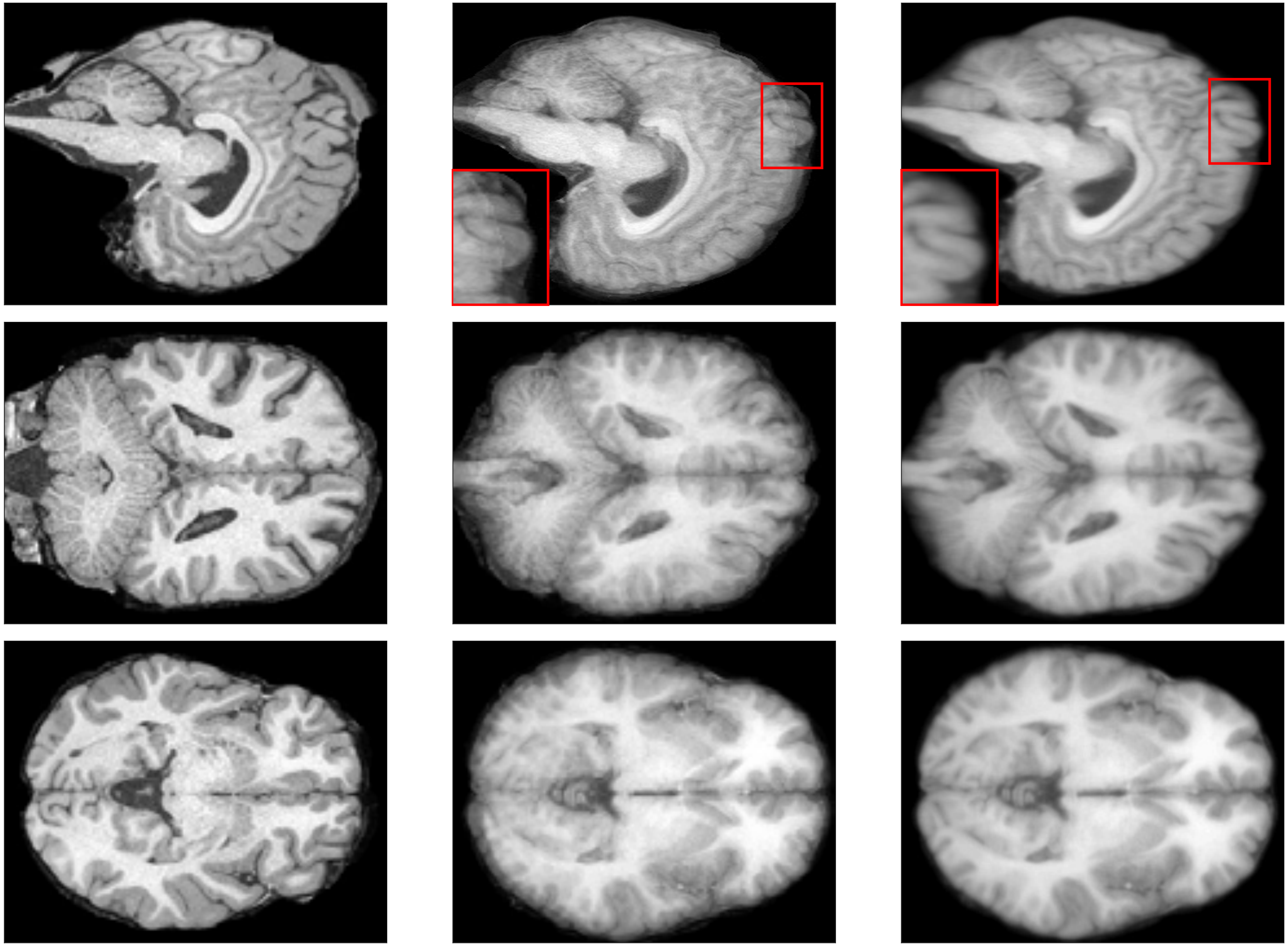
Application: Volume Reconstruction



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

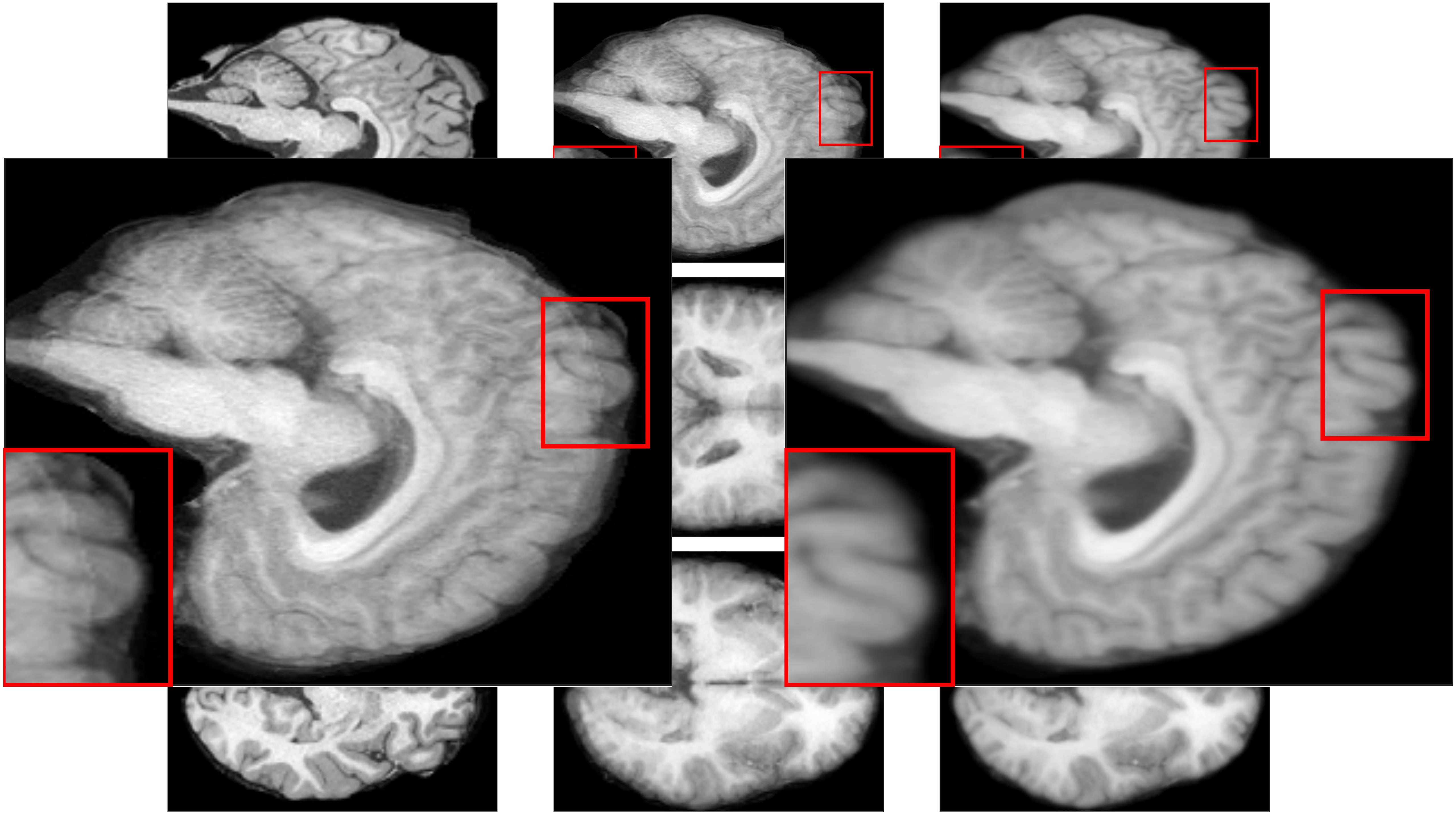
[Bonneel'16]

Application: Brain Mapping



Original | Euclidean projection | Wasserstein projection

Application: Brain Mapping



Original

Euclidean projection

Wasserstein projection

Distributionally Robust Learning

$$\boldsymbol{\nu}_{\text{data}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{\delta}_{(\boldsymbol{x}_i, \boldsymbol{y}_i)}$$

Supervised learning

$$\inf_{\theta \in \Theta} \mathbb{E}_{\boldsymbol{\nu}_{\text{data}}} [\mathcal{L}(f_\theta(X), Y)]$$

Learning with Wasserstein Ambiguity

$$\inf_{\theta \in \Theta} \sup_{\boldsymbol{\mu} : W_p(\boldsymbol{\nu}_{\text{data}}, \boldsymbol{\mu}) < \varepsilon} \mathbb{E}_{\boldsymbol{\mu}} [\mathcal{L}(f_\theta(X), Y)]$$

[Esvahani'17]

Distributionally Robust Learning

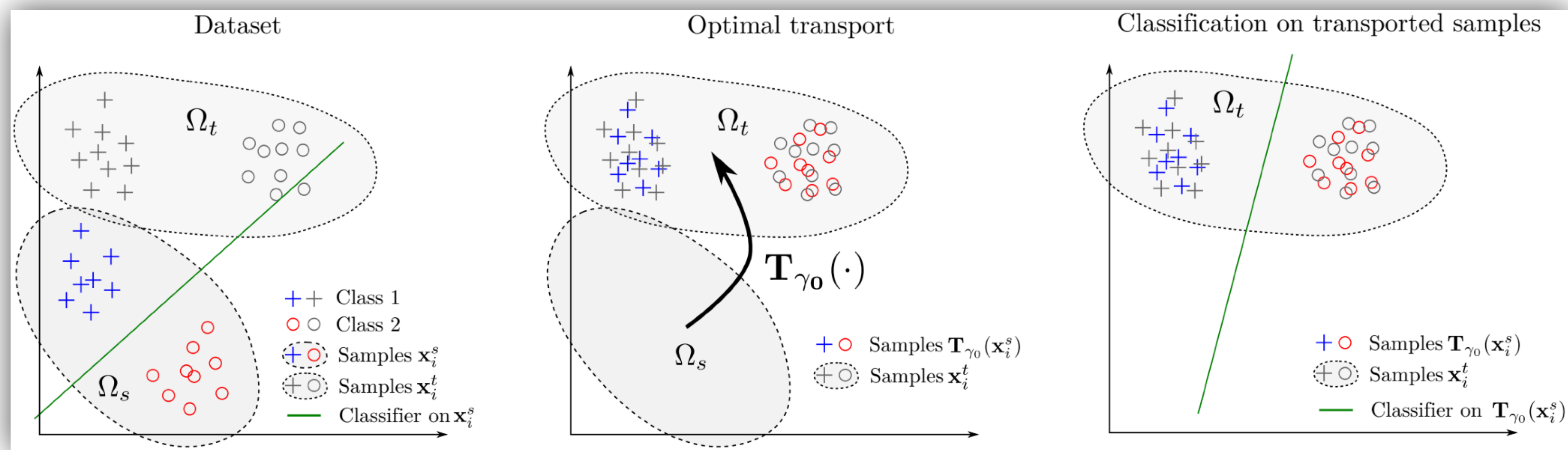
Learning with Wasserstein Ambiguity

$$\inf_{\theta \in \Theta} \sup_{\boldsymbol{\mu} : W_p(\boldsymbol{\nu}_{\mathrm{data}}, \boldsymbol{\mu}) < \varepsilon} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{L}(f_\theta(X), Y)]$$

Advantages:

- Bound on out-of-sample performance
- Converges as size of dataset increases
- Often reduces to a finite convex program (e.g. when f is element-wise max over elementary concave functions)

Domain Adaptation



1. Estimate transport map
2. Transport labeled samples to new domain
3. Train classifier on transported labeled samples

Learning with a Wasserstein Loss

$$\text{Dataset } \{(x_i, y_i)\}, x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^n_+$$

x_i

husky
snow
sled
slope
men

y_i

Goal is to find $f_{\boldsymbol{\theta}}$: Images $\mapsto$ Labels

Learning with a Wasserstein Loss

$$\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{N} \mathcal{L}(f_{\boldsymbol{\theta}}(x_i), y_i)$$

x_i

husky
snow
sled
slope
men

y_i

Which loss $\mathcal{L}$ could we use?

Learning with a Wasserstein Loss

$$\min_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{N} \mathcal{L}(f_{\boldsymbol{\theta}}(x_i), y_i)$$

dog driver winter ice $f_{\boldsymbol{\theta}}(x_i)$	husky snow sled slope men y_i

Which loss $\mathcal{L}$ could we use?

Learning with a Wasserstein Loss

$$\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{N} \mathcal{L}(f_{\boldsymbol{\theta}}(x_i), y_i)$$

$$\mathcal{L}(\boldsymbol{a}, \boldsymbol{b}) = \min_{\boldsymbol{P} \in \mathbb{R}^{nm}} \langle \boldsymbol{P}, M \rangle + \varepsilon \mathrm{KL}(\boldsymbol{P}\mathbf{1}, \boldsymbol{a}) + \varepsilon \mathrm{KL}(\boldsymbol{P}^T\mathbf{1}, \boldsymbol{b}) - \gamma E(\boldsymbol{P})$$

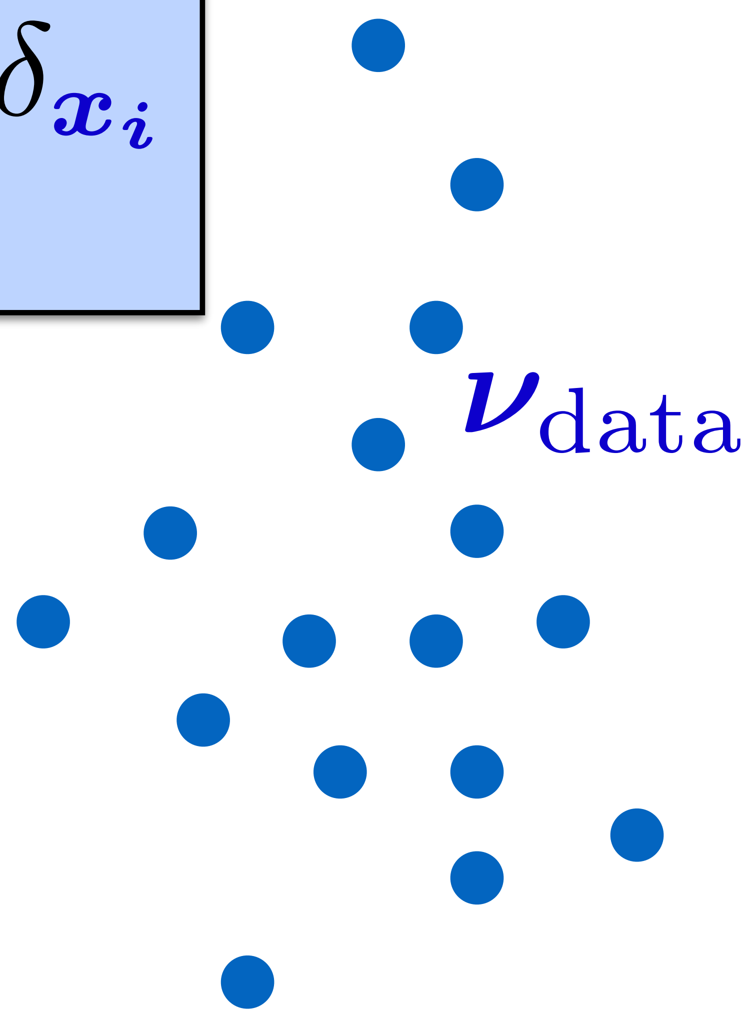
1. Generalizes Word Mover's to label clouds
2. Sinkhorn algorithm can be generalized

[Frogner'15] [Chizat'15][Chizat'16]

Statistics 0.1 : Density Fitting

We collect data

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



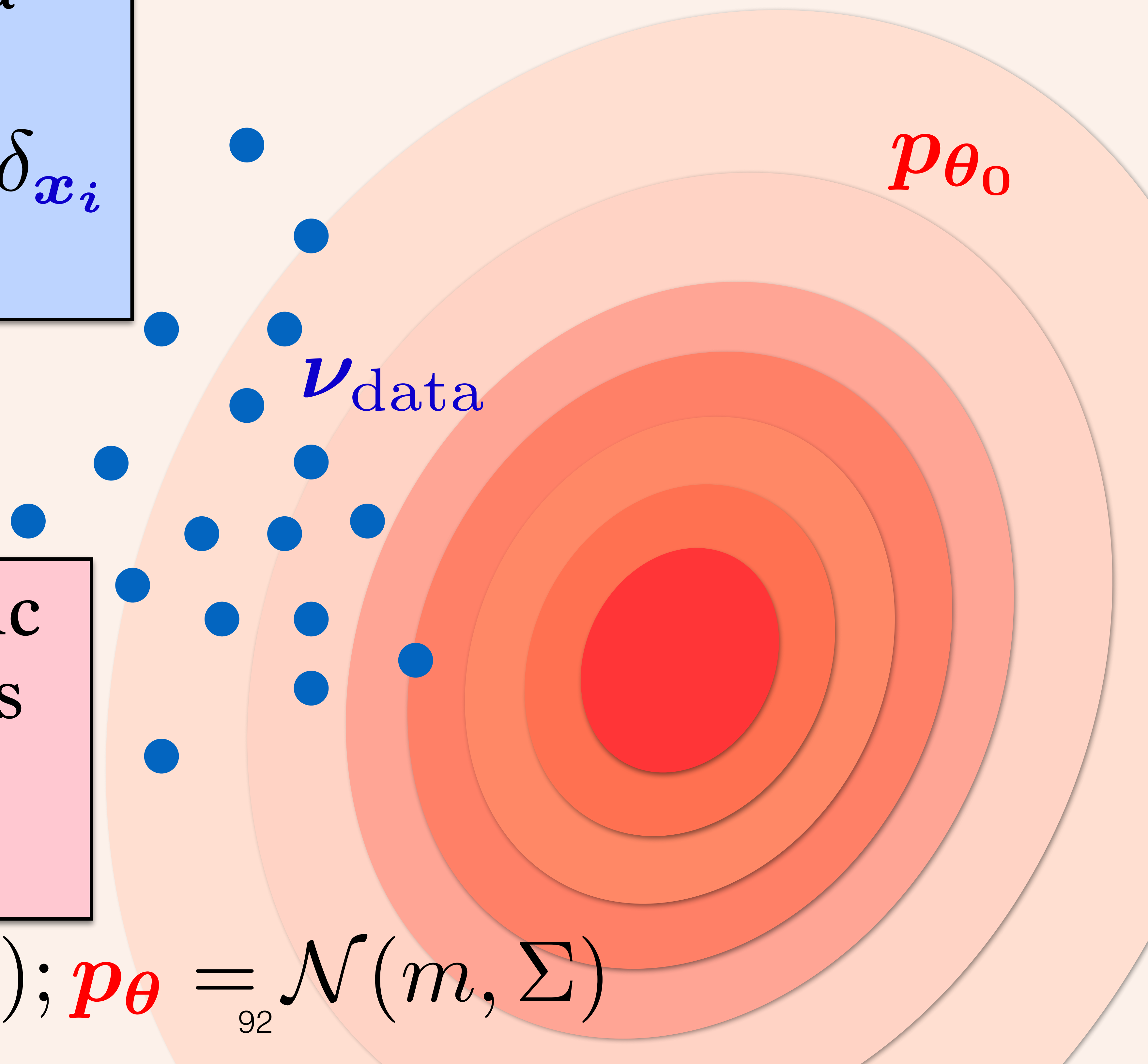
Statistics 0.1 : Density Fitting

We collect data

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

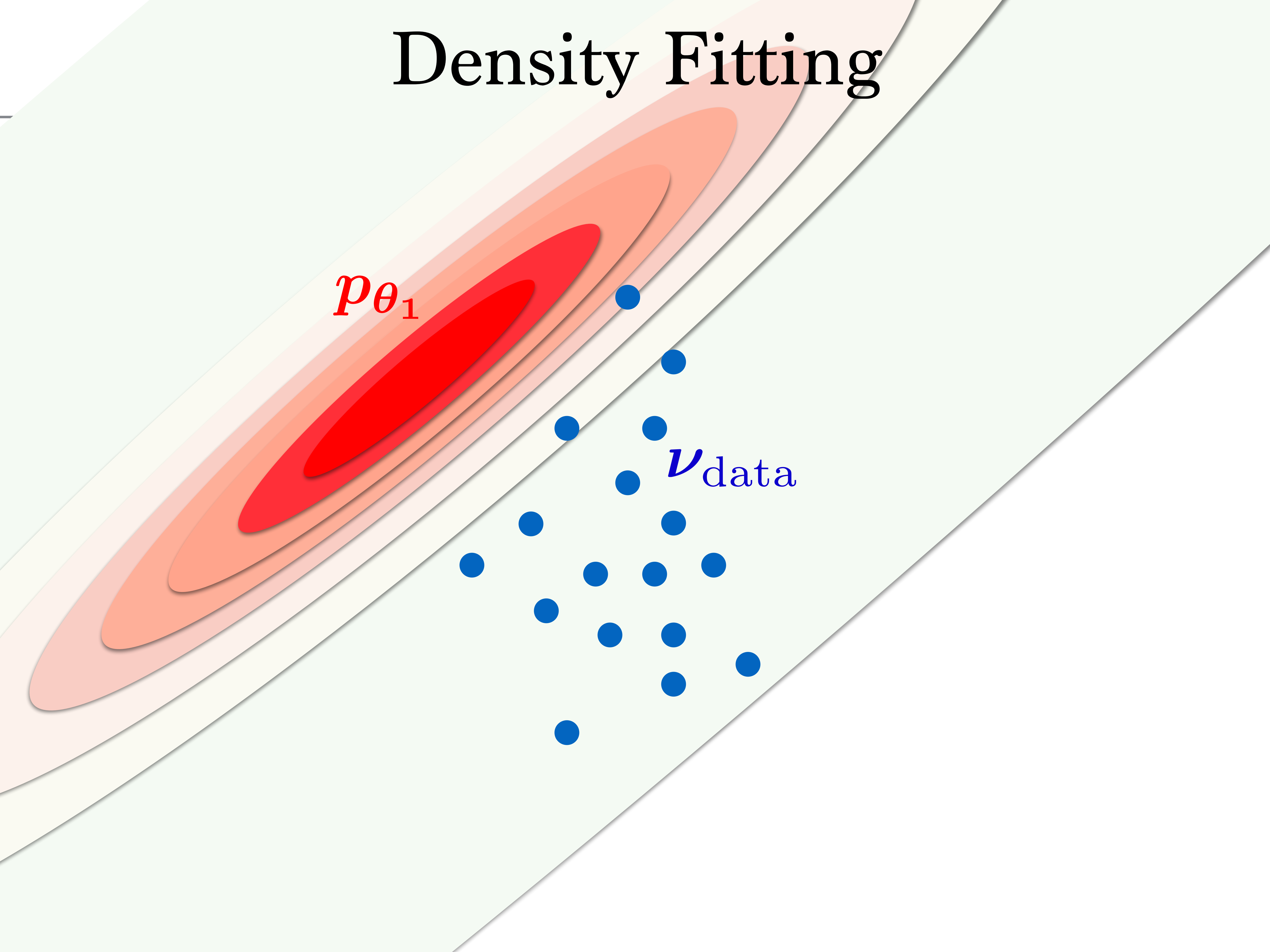
We fit a parametric family of densities

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

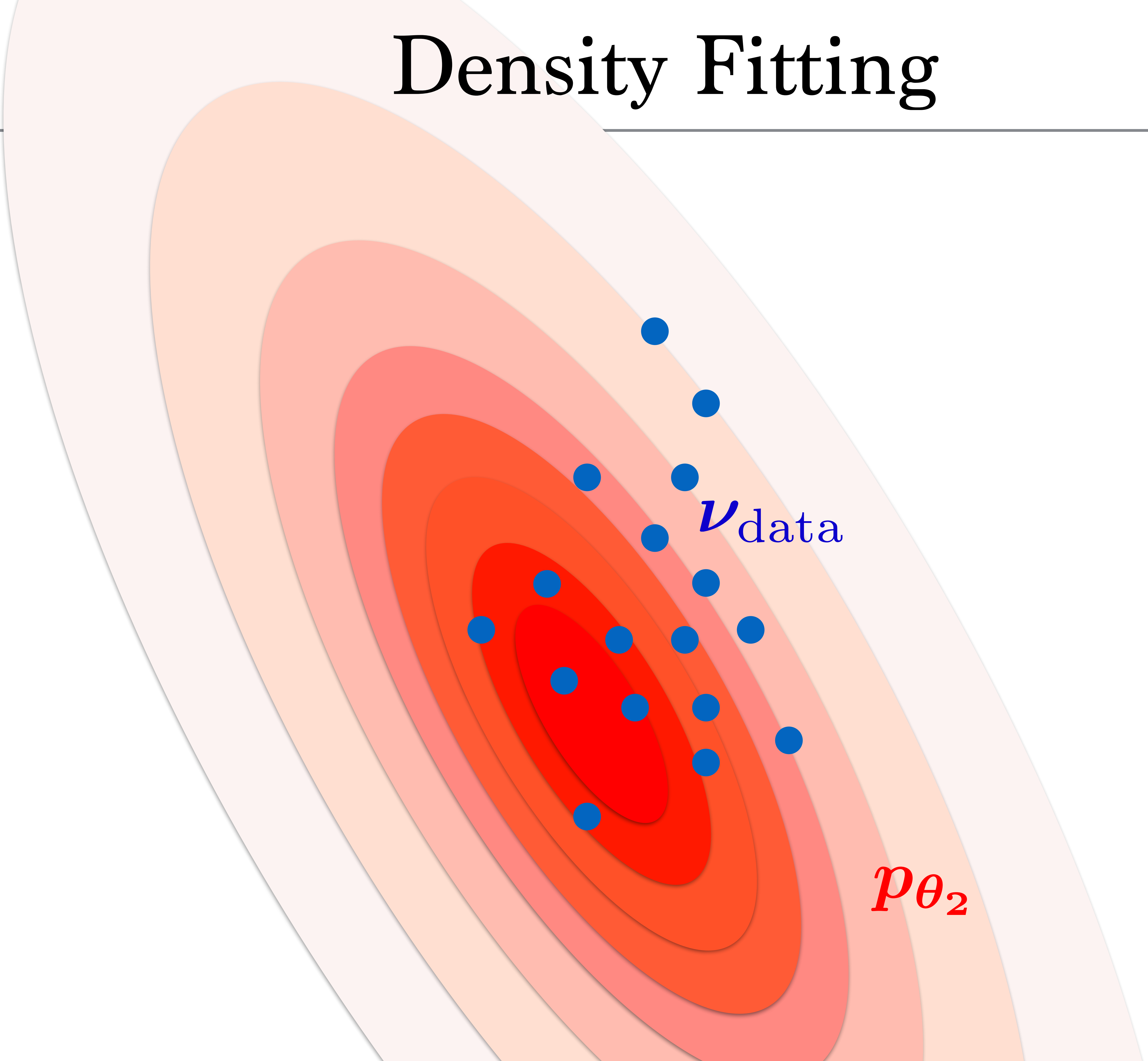


$$e.g.\ \boldsymbol{\theta} = (m, \Sigma); \boldsymbol{p_\theta} = \mathcal{N}(m, \Sigma)$$

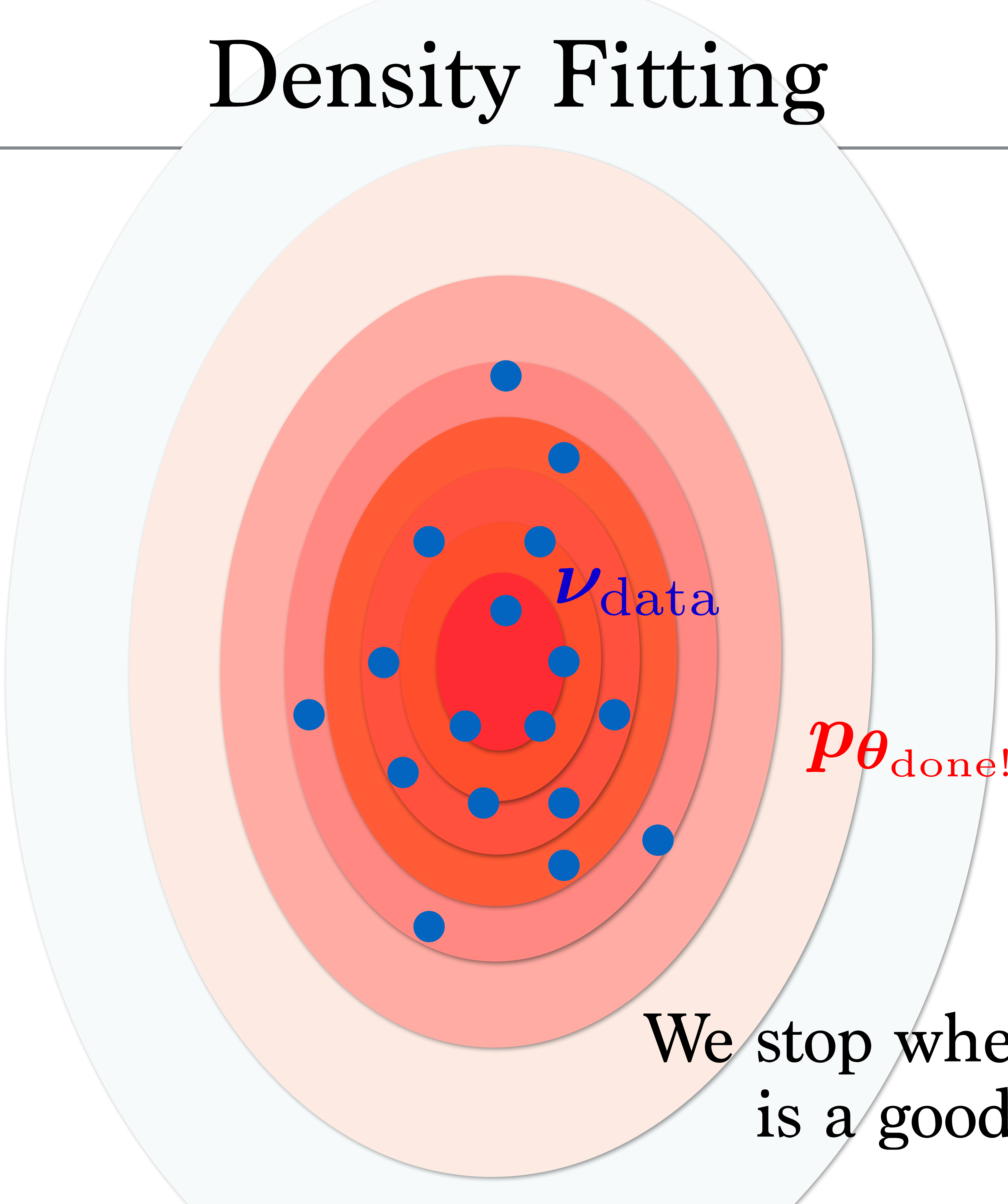
Density Fitting



Density Fitting

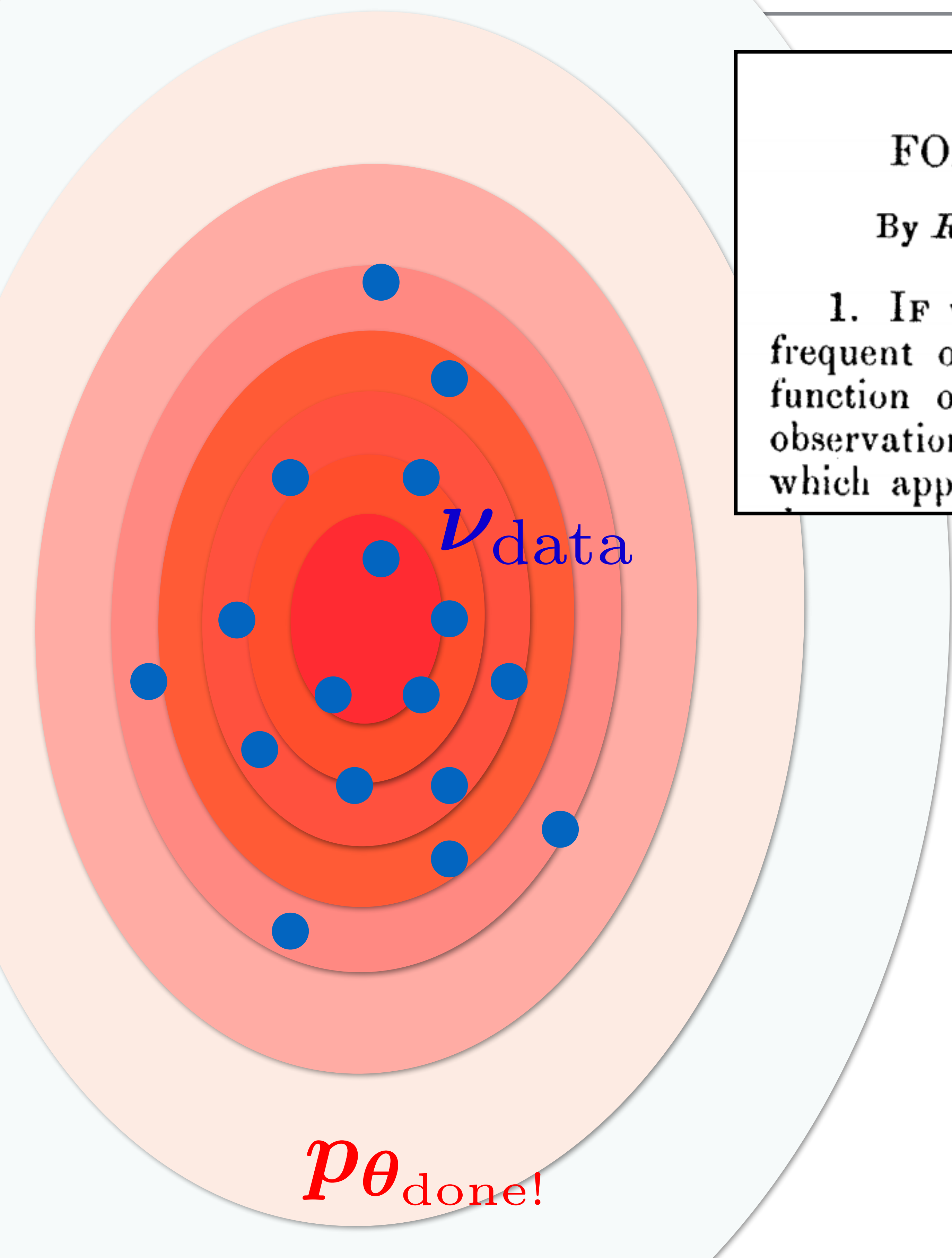


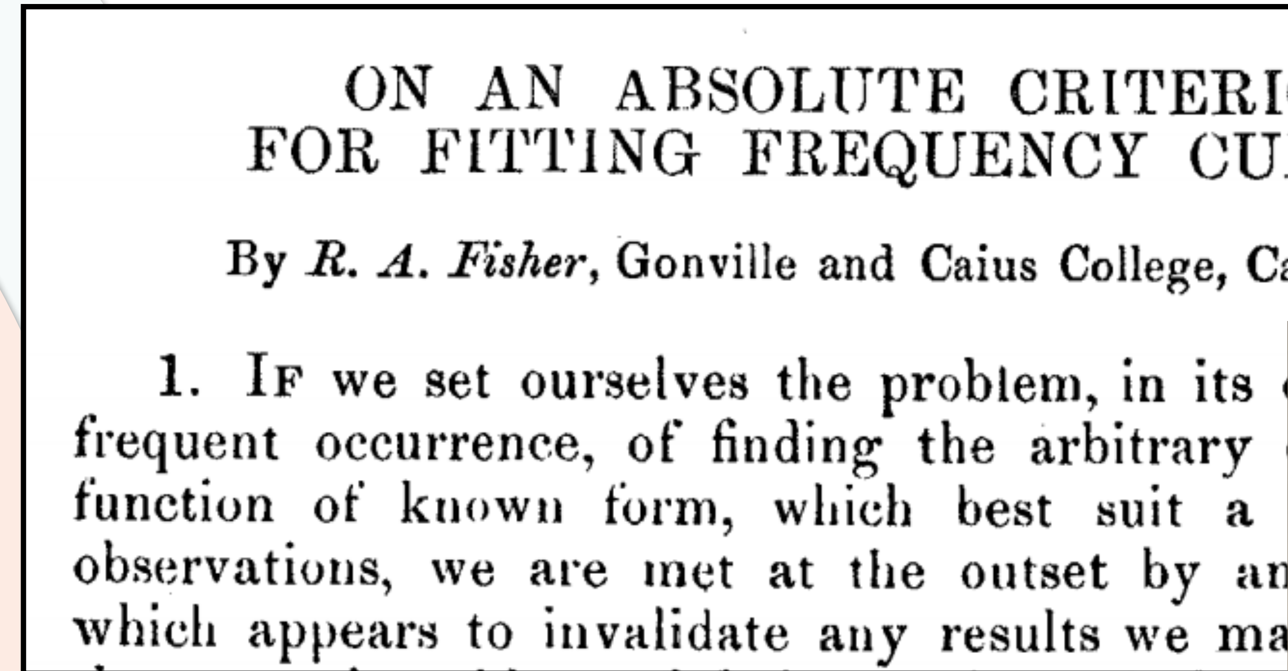
Density Fitting



We stop when there
is a good fit.

Maximum Likelihood Estimation





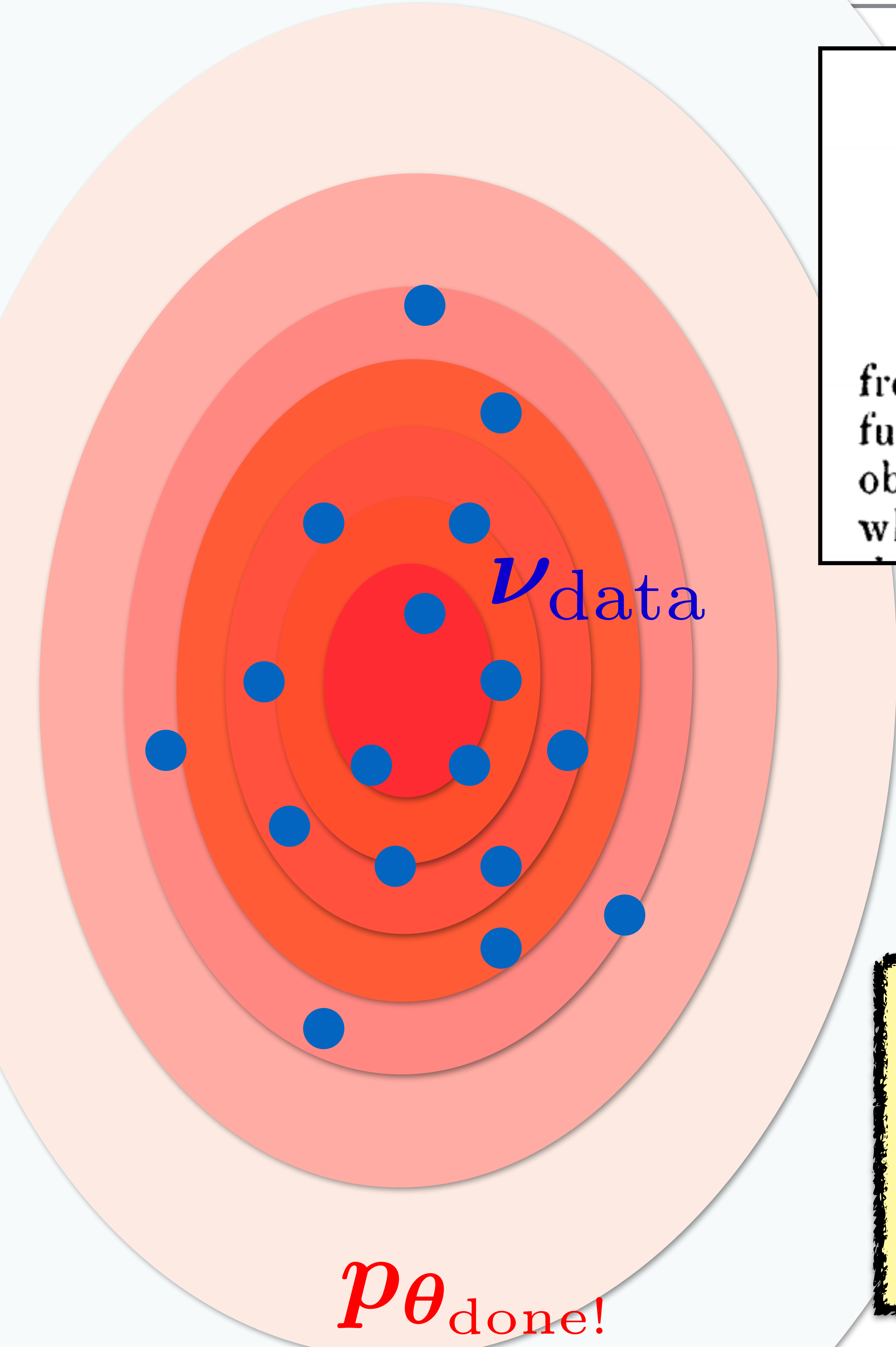
ON AN ABSOLUTE CRITERION
FOR FITTING FREQUENCY CURVES.

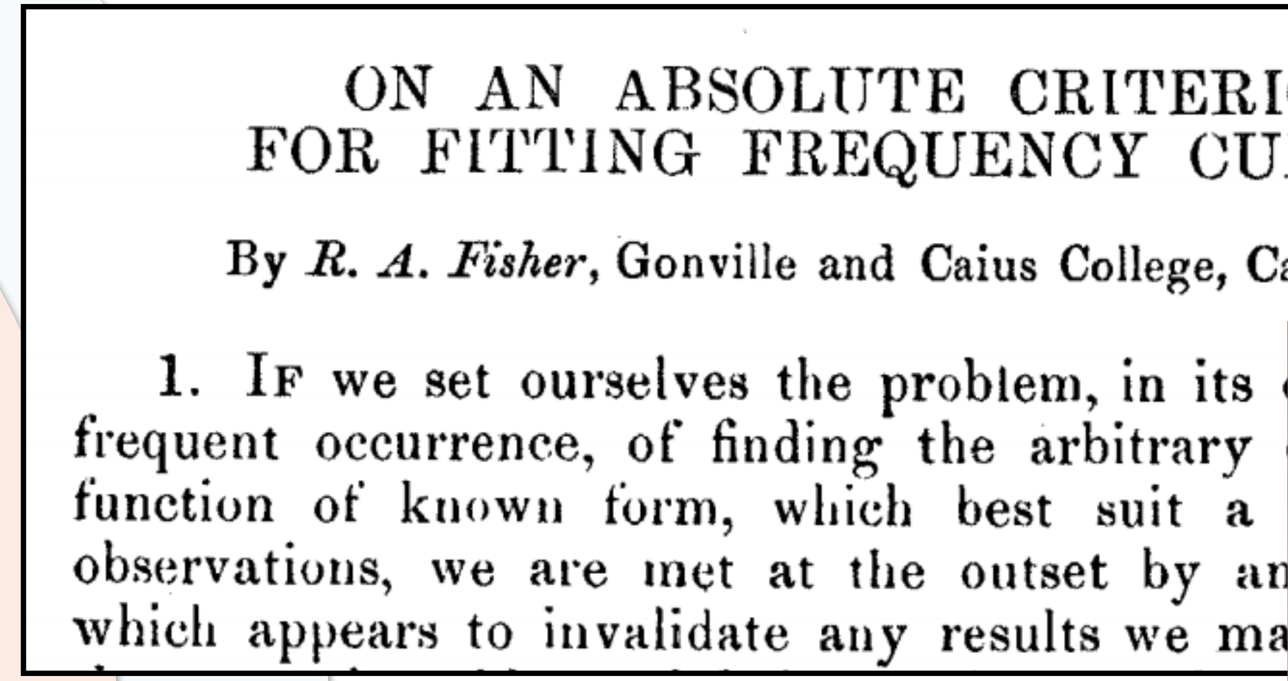
By *R. A. Fisher*, Gonville and Caius College, Cambridge.

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

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

Maximum Likelihood Estimation





ON AN ABSOLUTE CRITERION
FOR FITTING FREQUENCY CURVES.

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

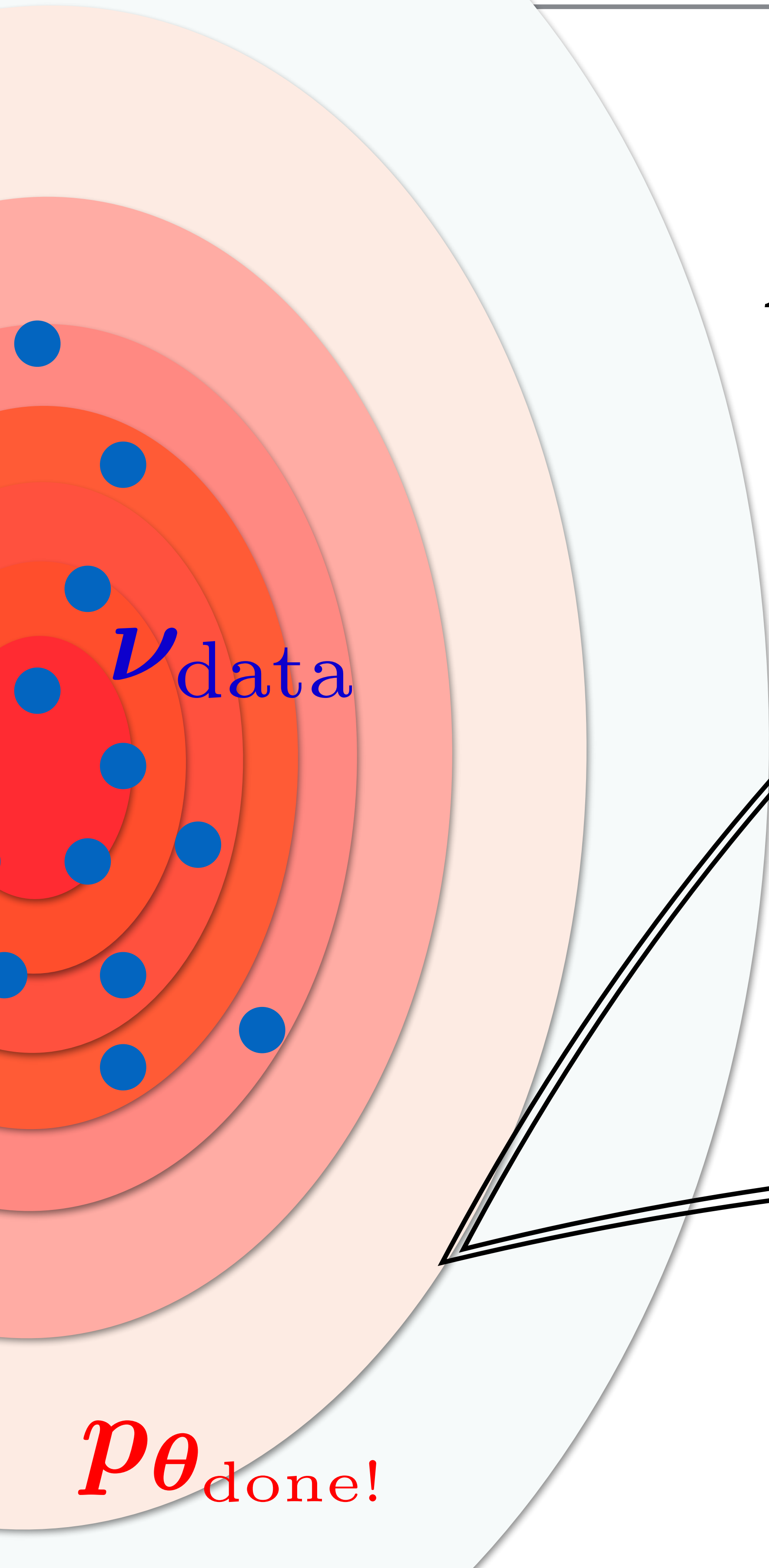
1. If we set ourselves the problem, in its
frequent occurrence, of finding the arbitrary
function of known form, which best suit a
observations, we are met at the outset by an
which appears to invalidate any results we ma

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

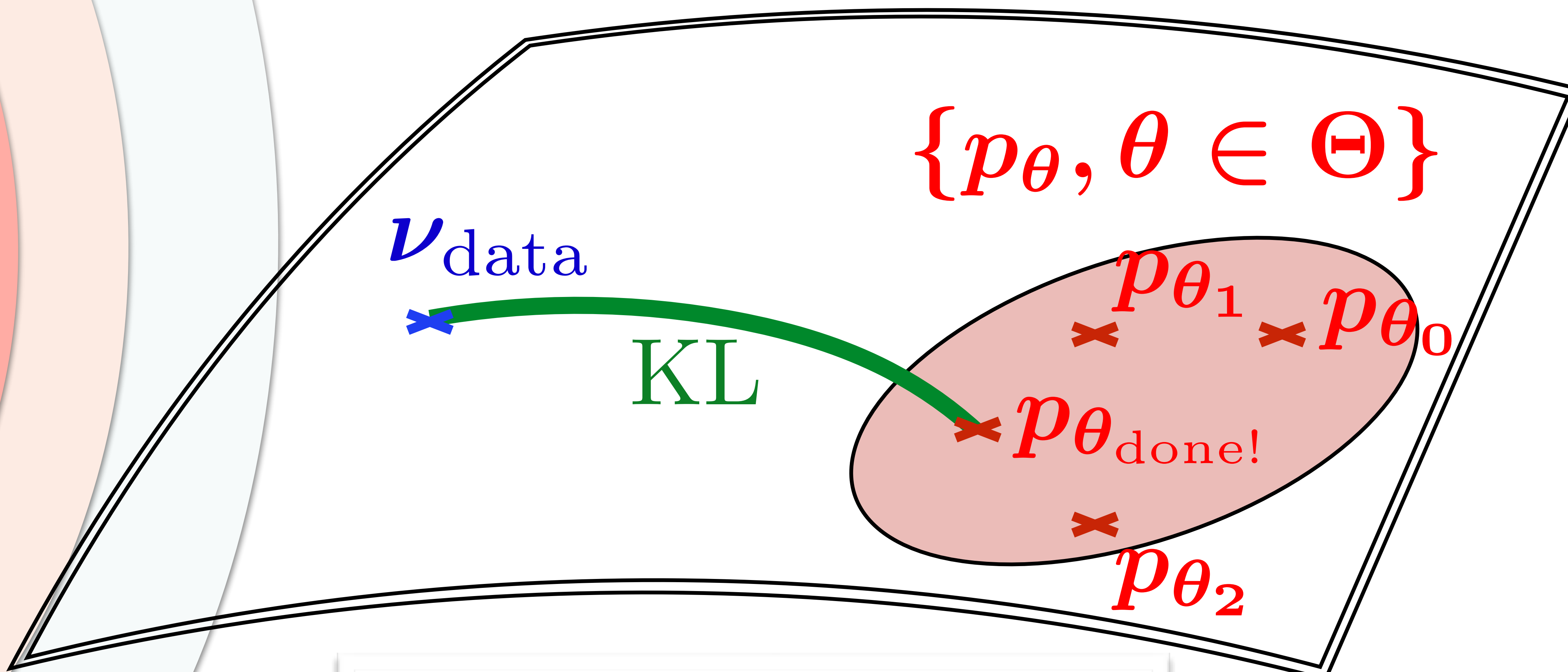
$$\log 0 = -\infty$$

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

Maximum Likelihood Estimation

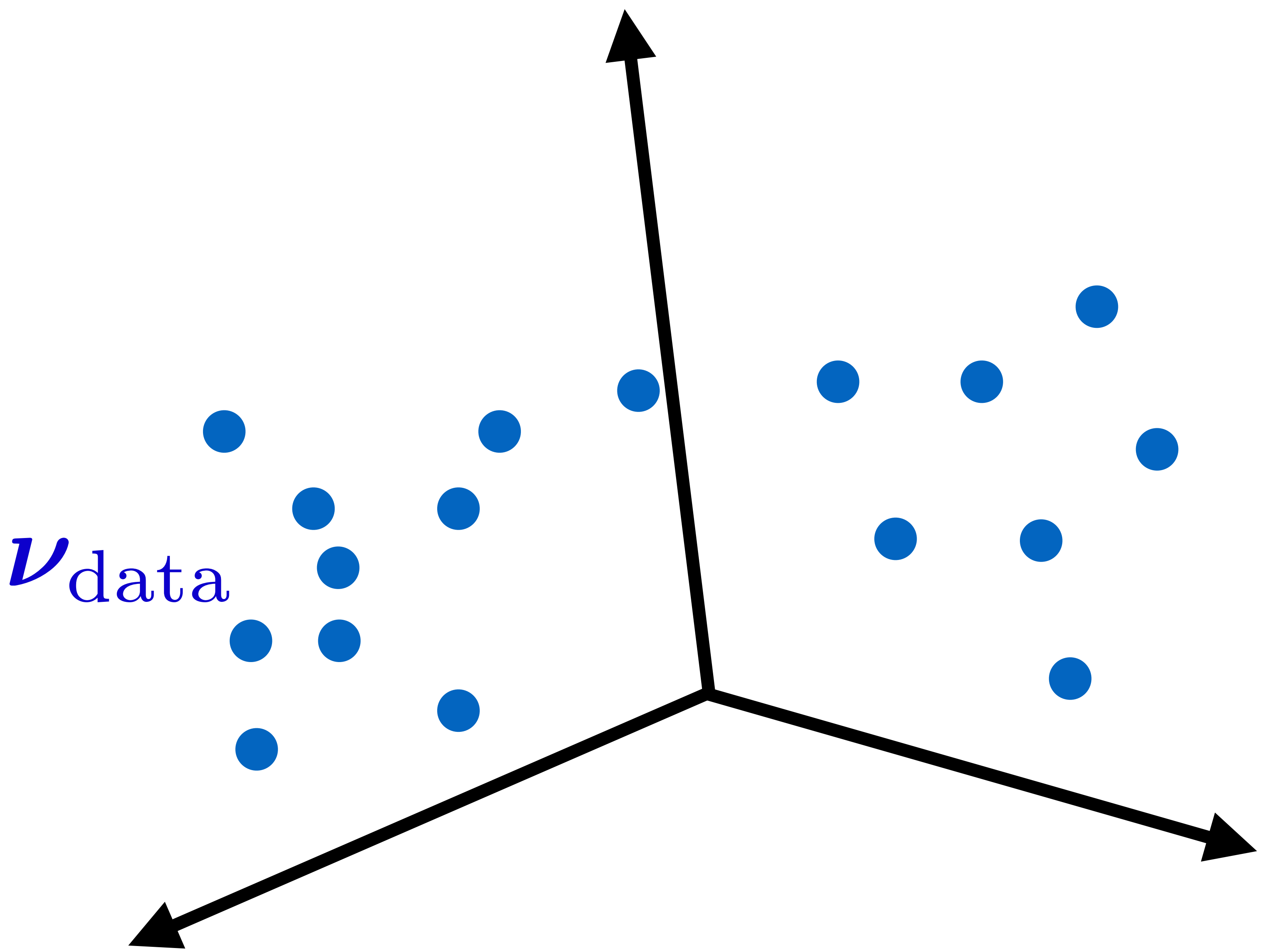


Equivalent to a KL projection in the space of probability measures



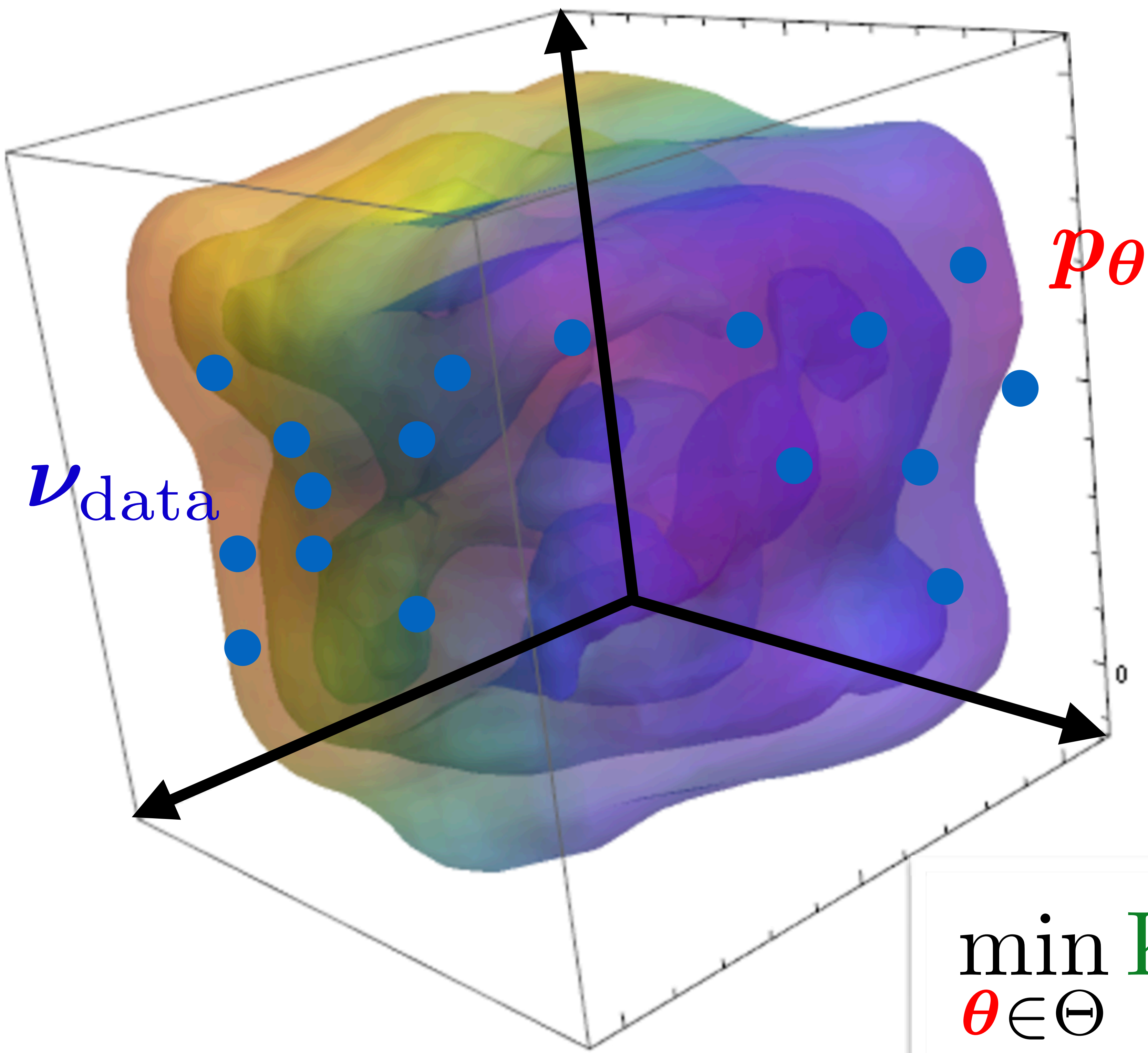
$$\min_{\boldsymbol{\theta} \in \Theta} \mathrm{KL}(\boldsymbol{\nu}_{\mathrm{data}} \| p_{\boldsymbol{\theta}})$$

In higher dimensional spaces...



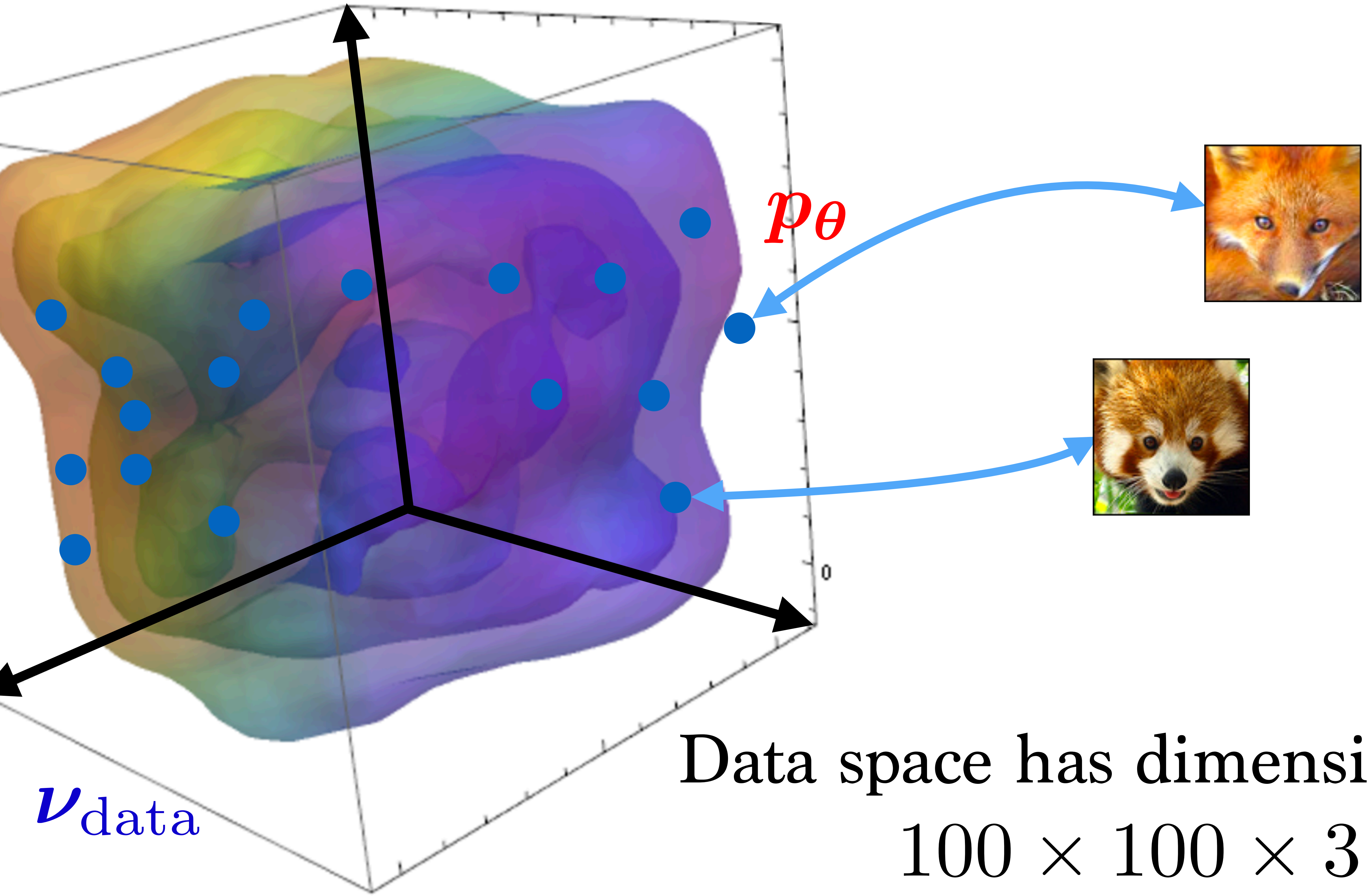
$$\min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\boldsymbol{\nu}_{\mathrm{data}} \| \boldsymbol{p}_{\boldsymbol{\theta}})$$

In higher dimensional spaces...



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

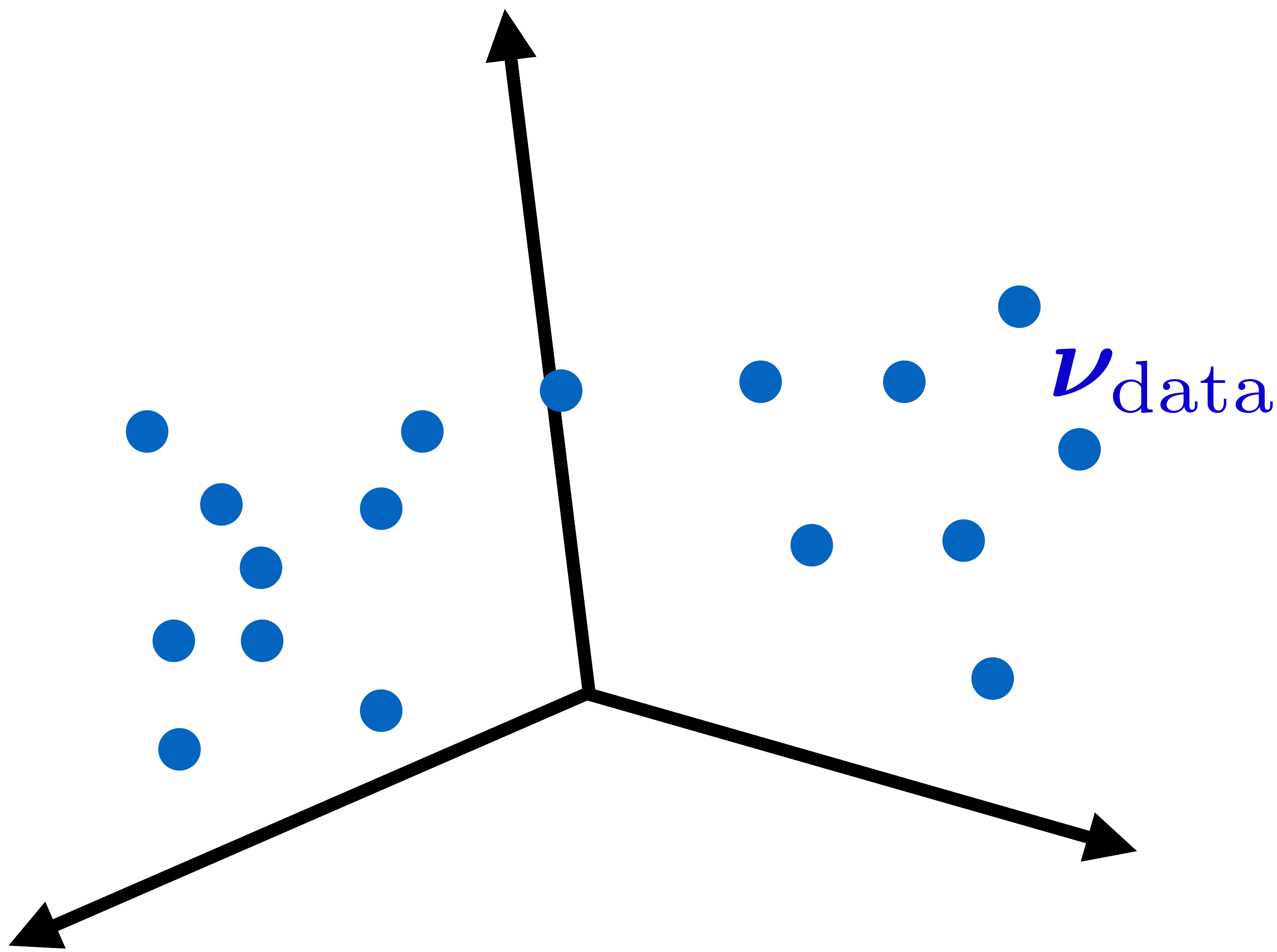
In higher dimensional spaces...



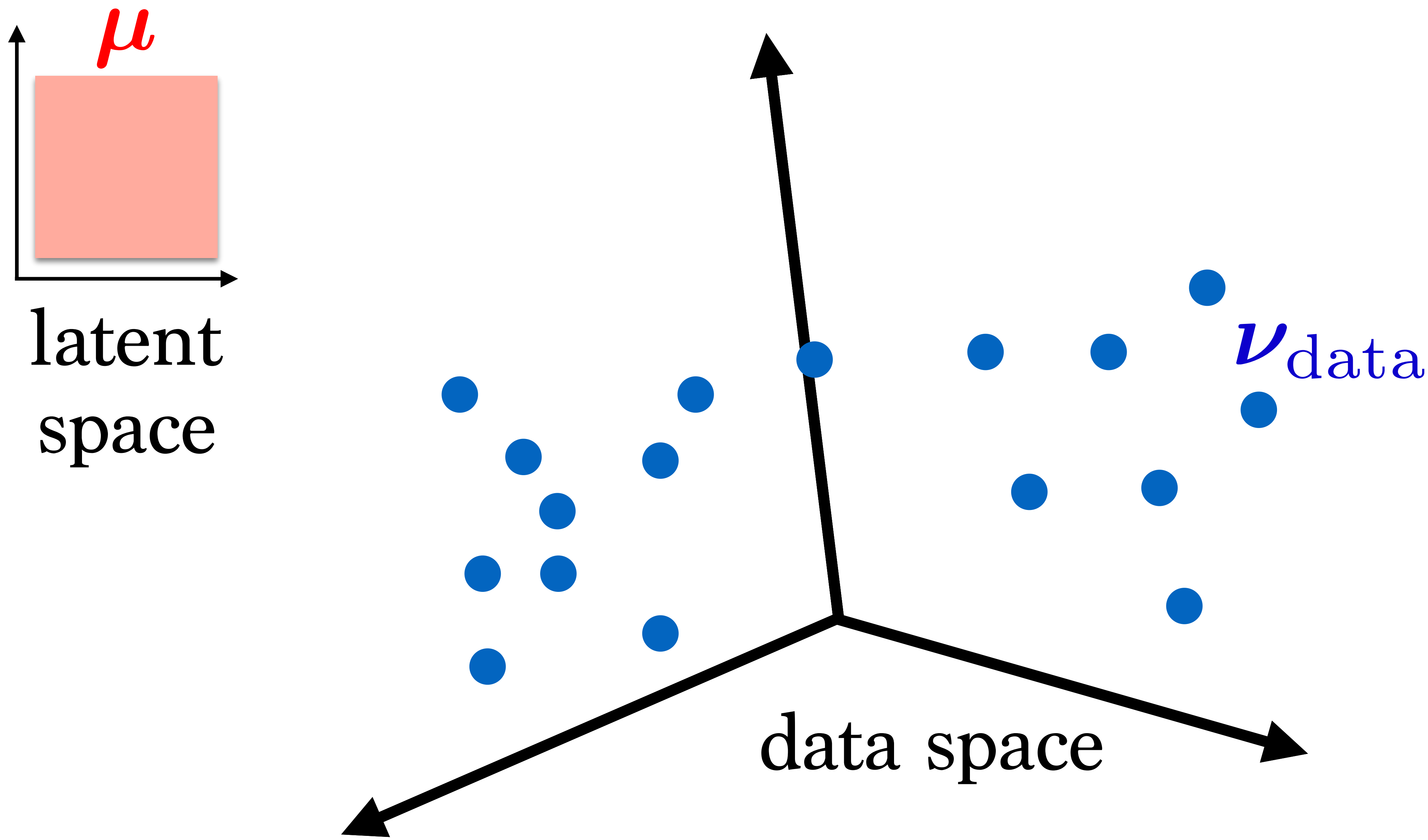
Data space has dimension

$$100 \times 100 \times 3$$

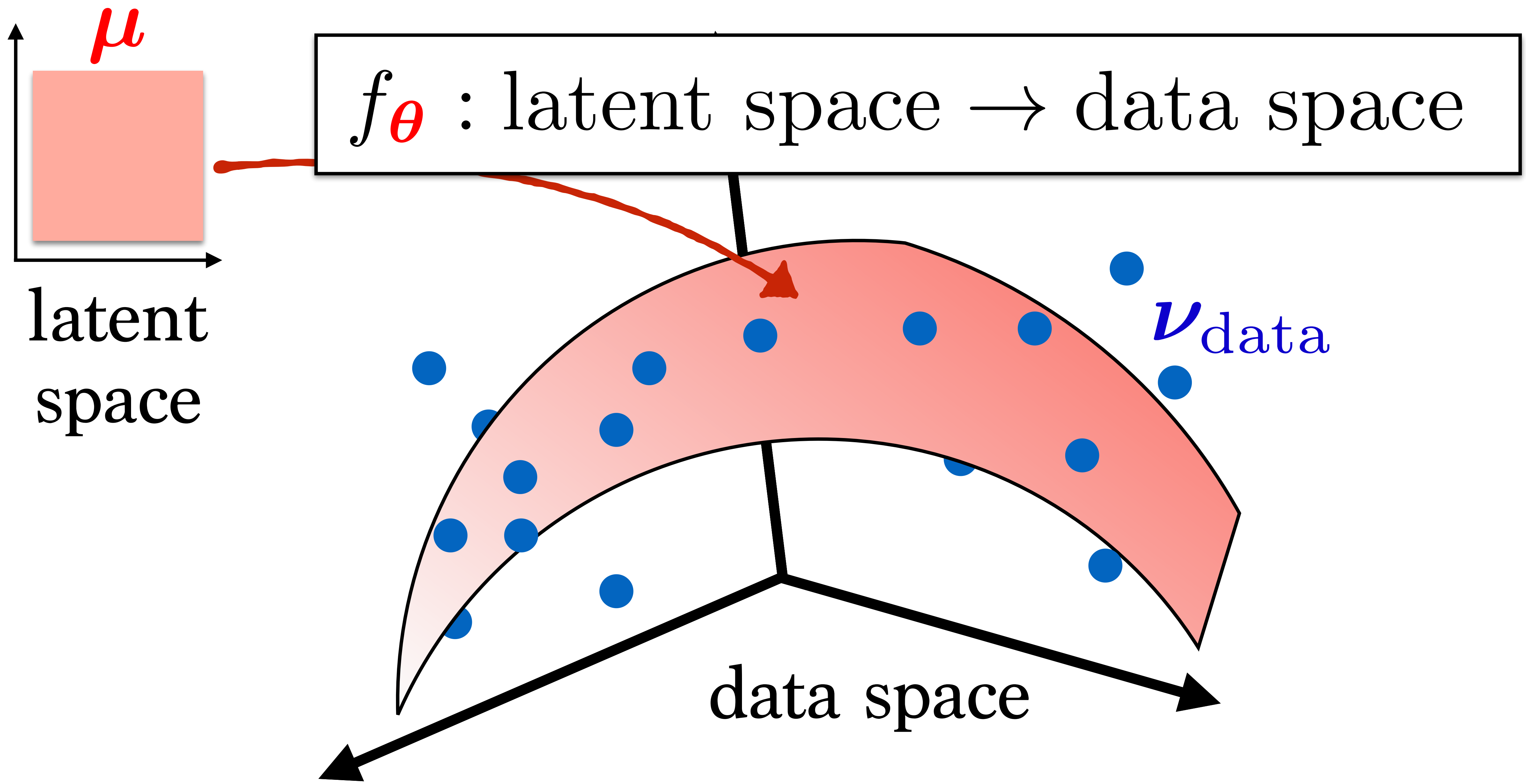
Generative Models



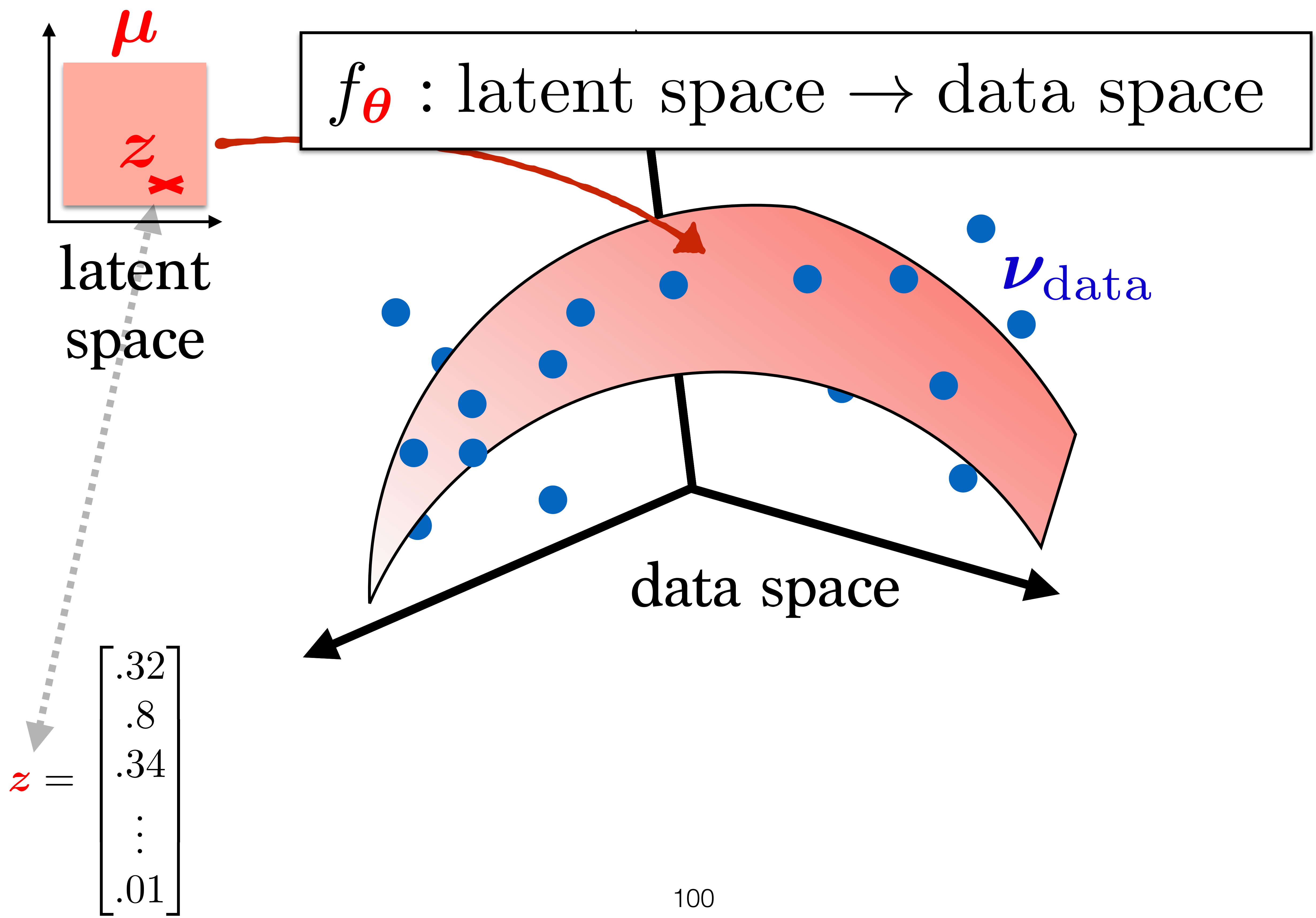
Generative Models



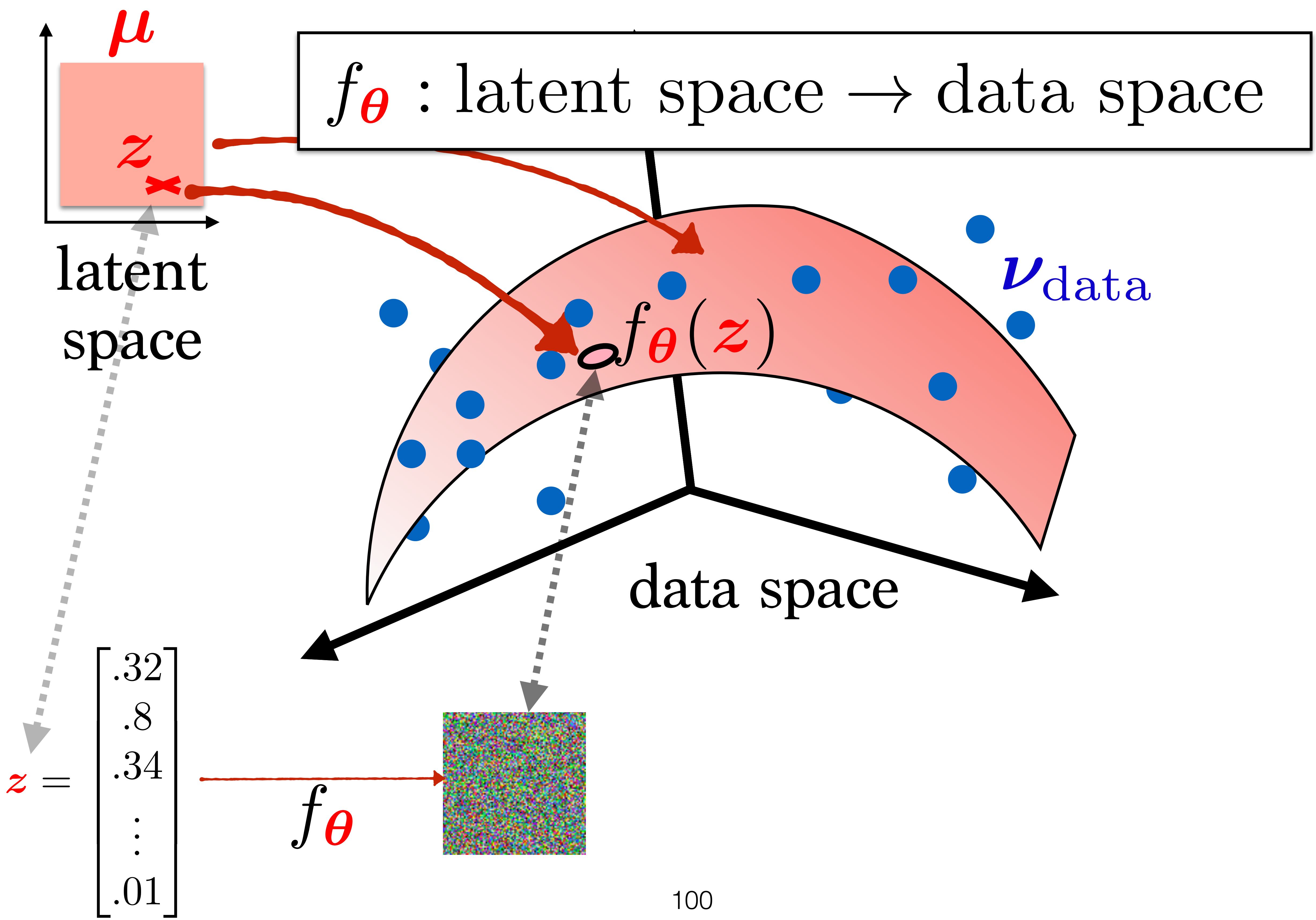
Generative Models



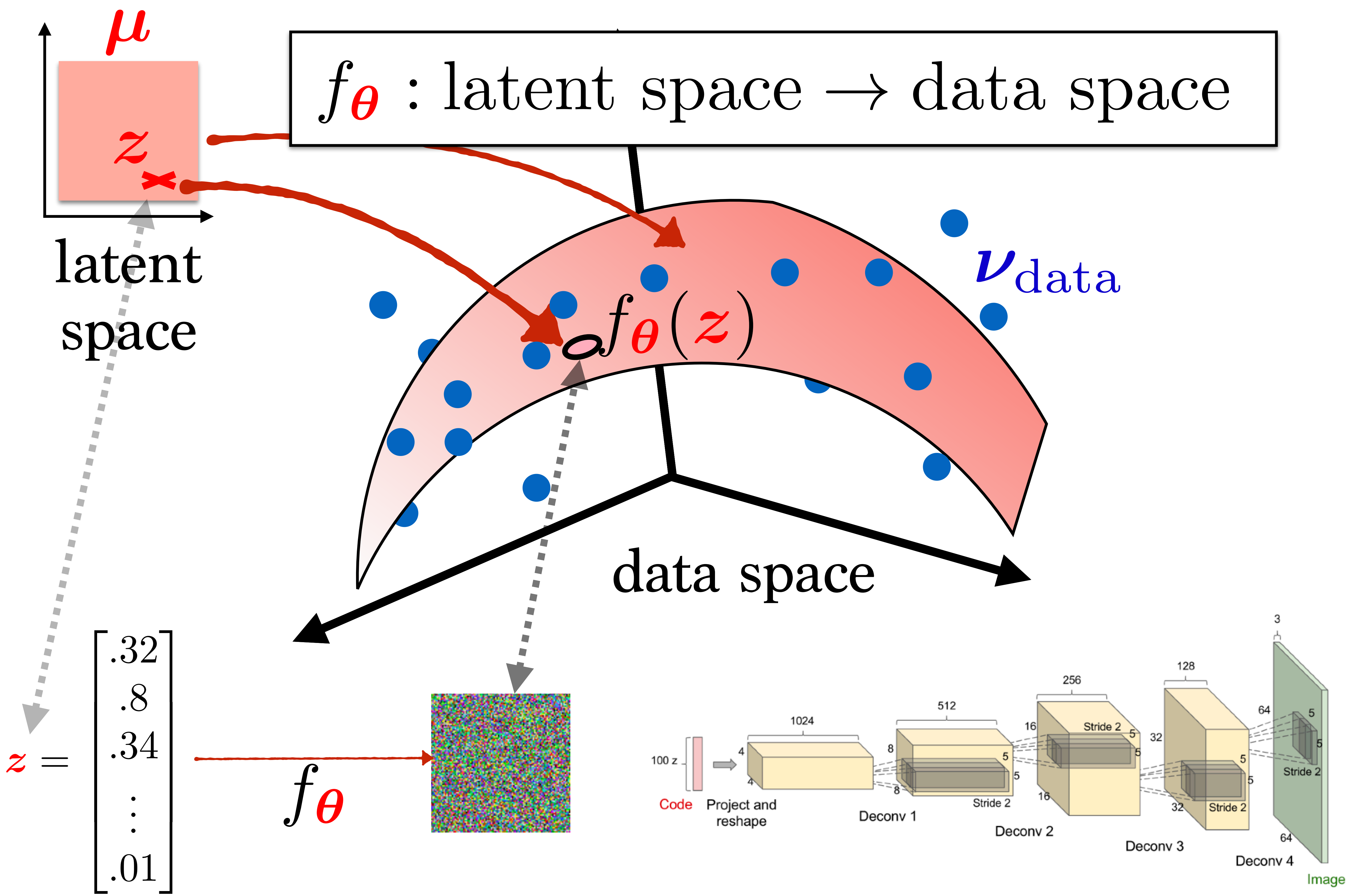
Generative Models



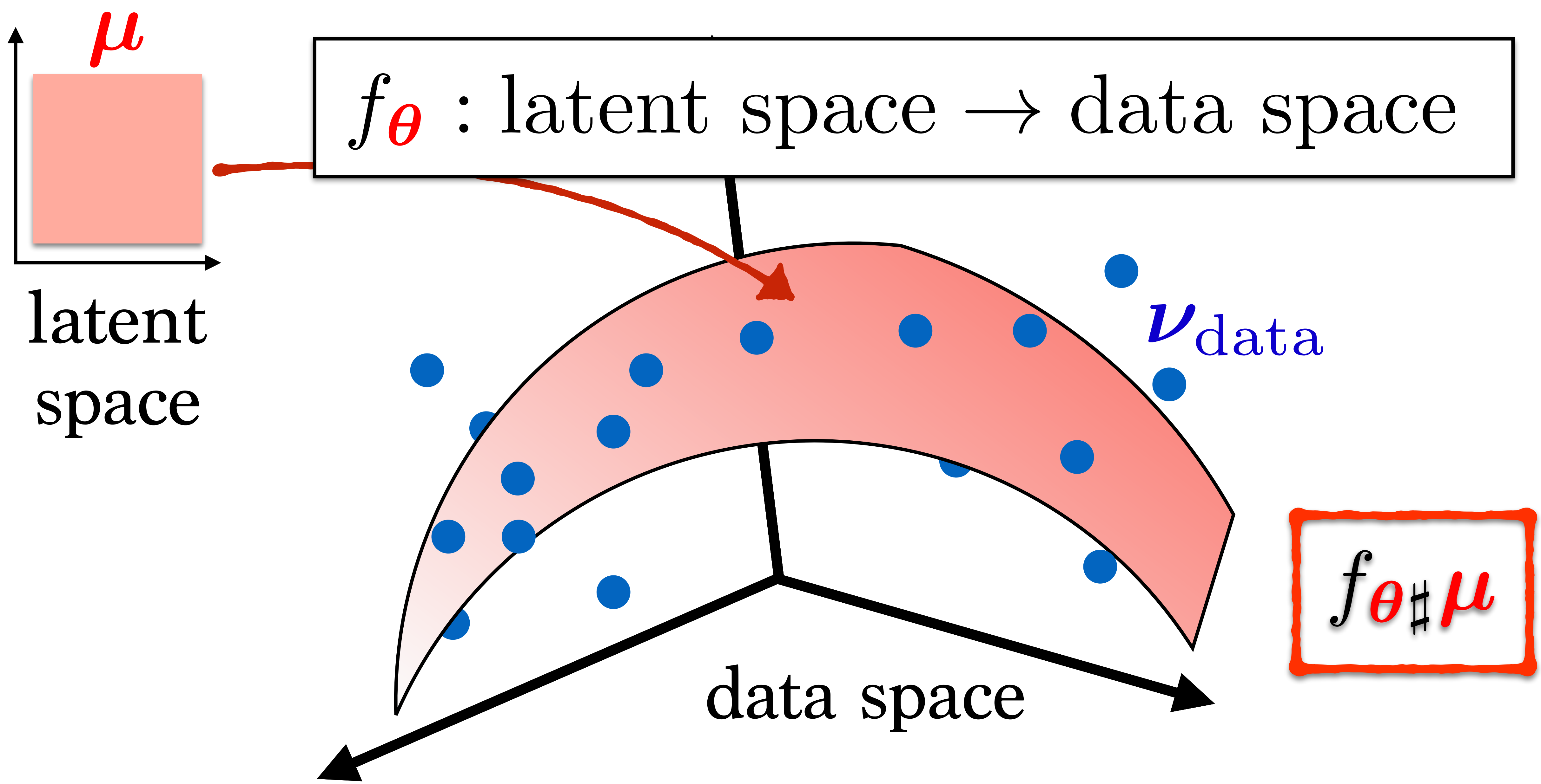
Generative Models



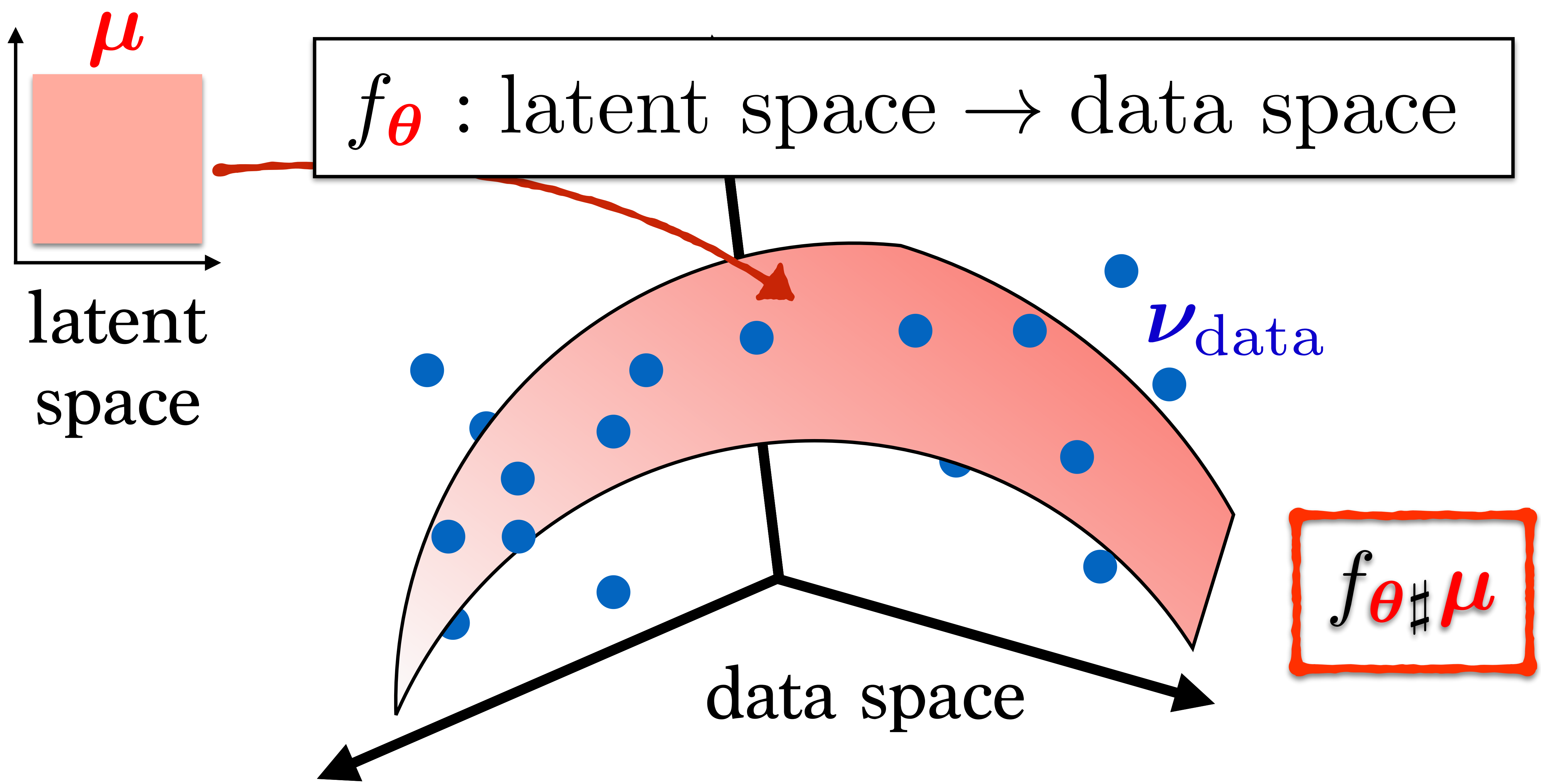
Generative Models



Generative Models

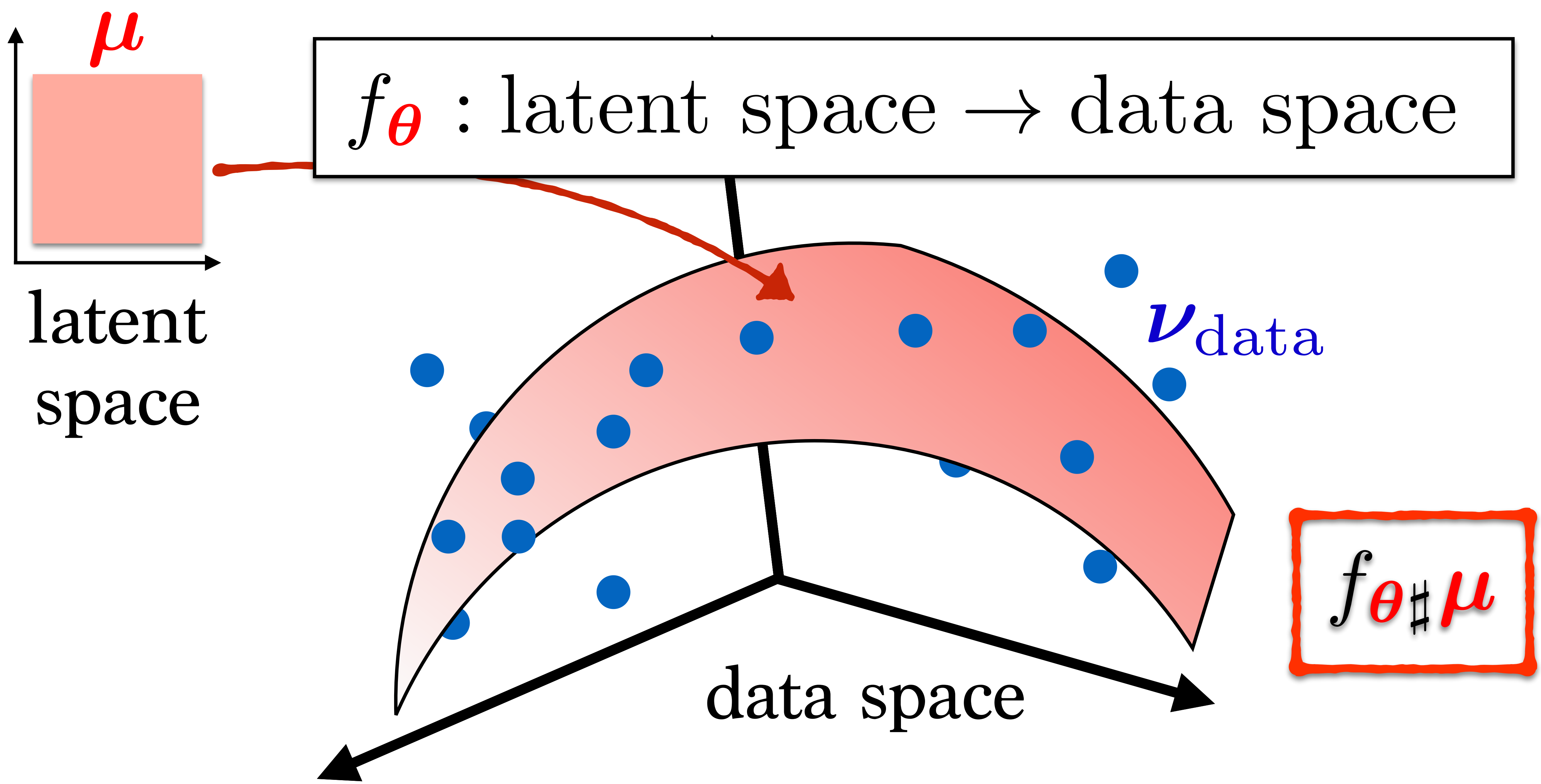


Generative Models



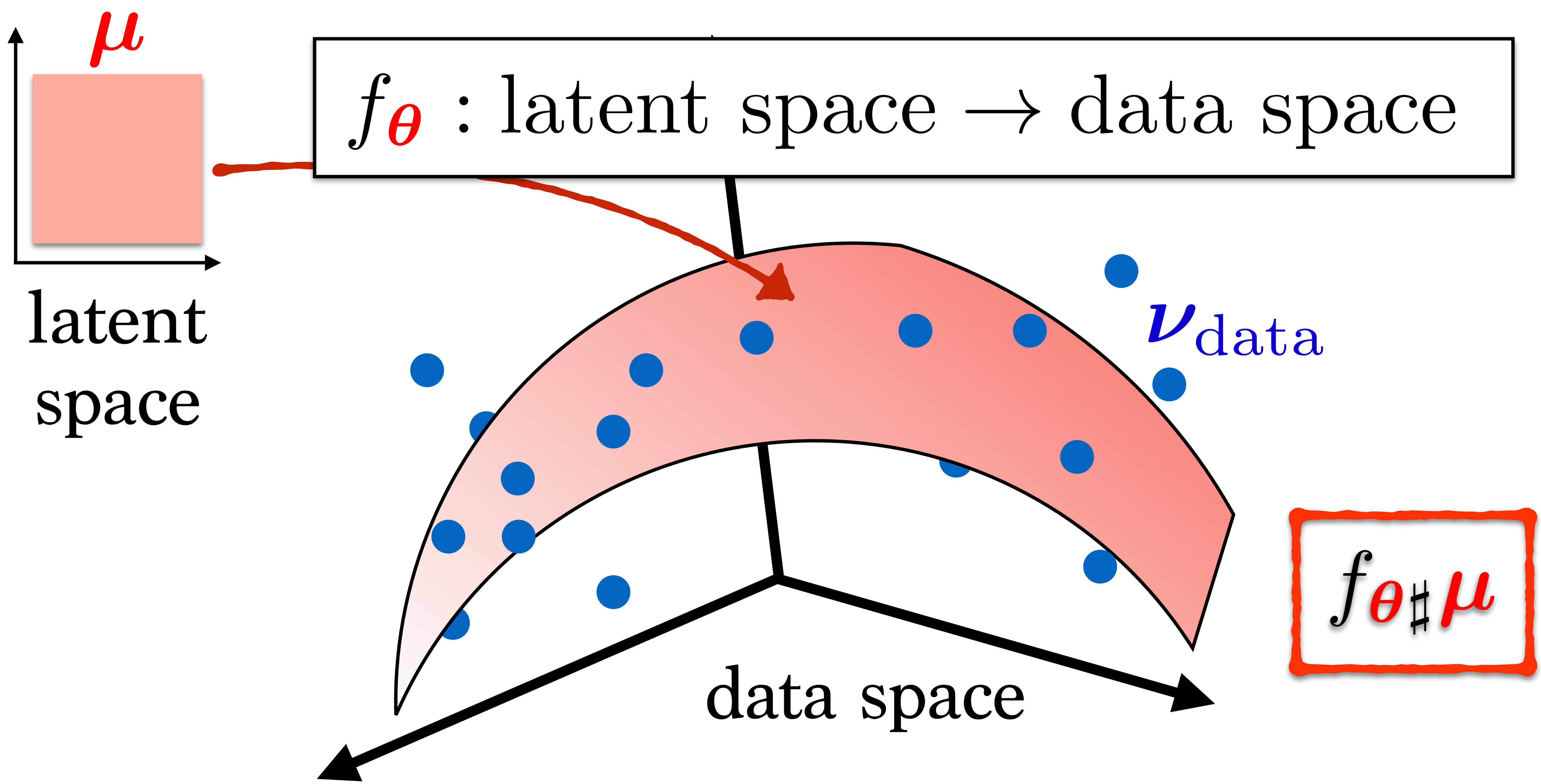
Push-forward: $\forall B \subset \Omega,\ f_\sharp\mu(B) := \mu(f^{-1}(B))$

Generative Models



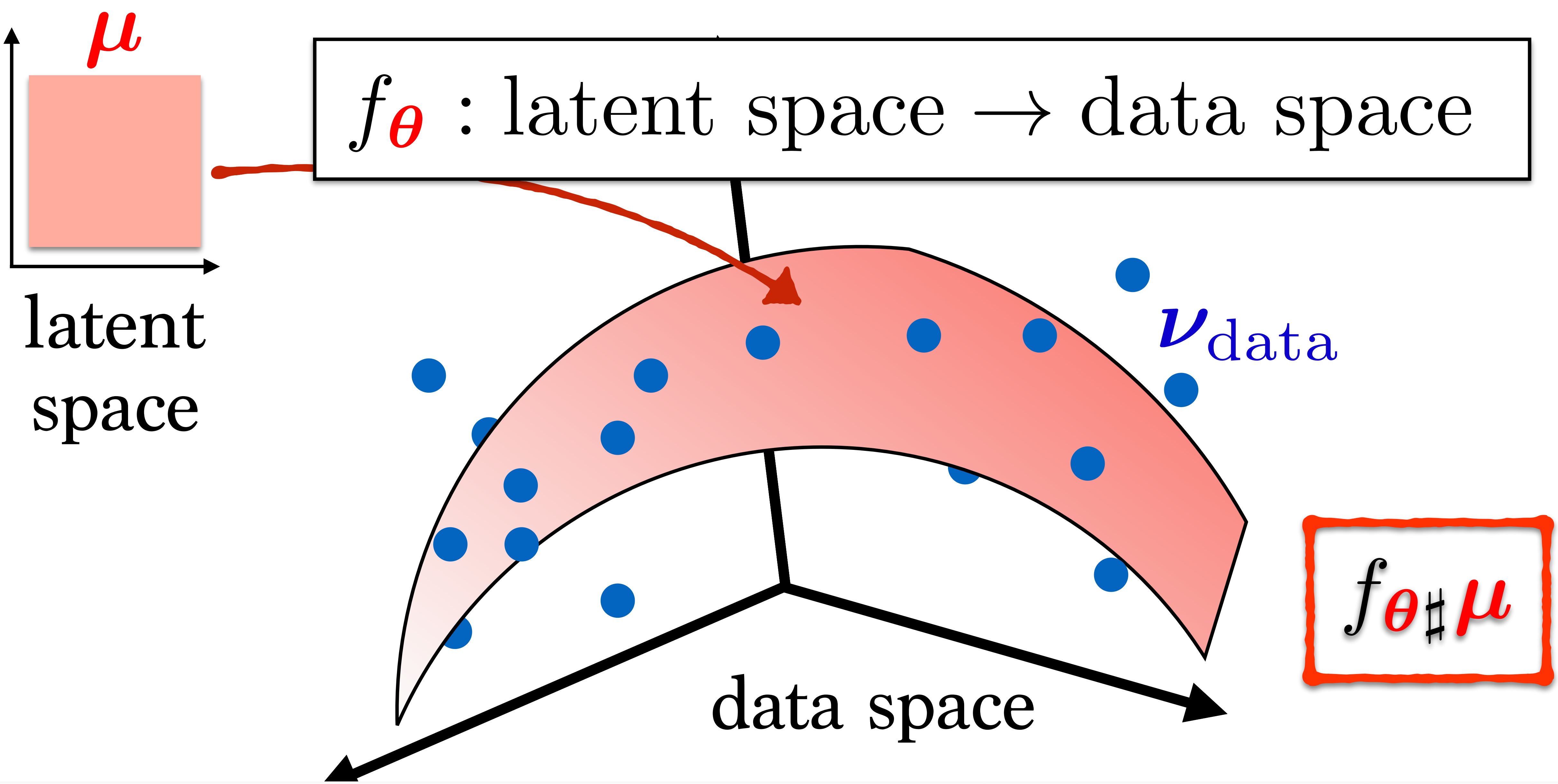
Goal: find $\boldsymbol{\theta}$ such that $f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu}$ fits $\boldsymbol{\nu}_{\text{data}}$

Generative Models



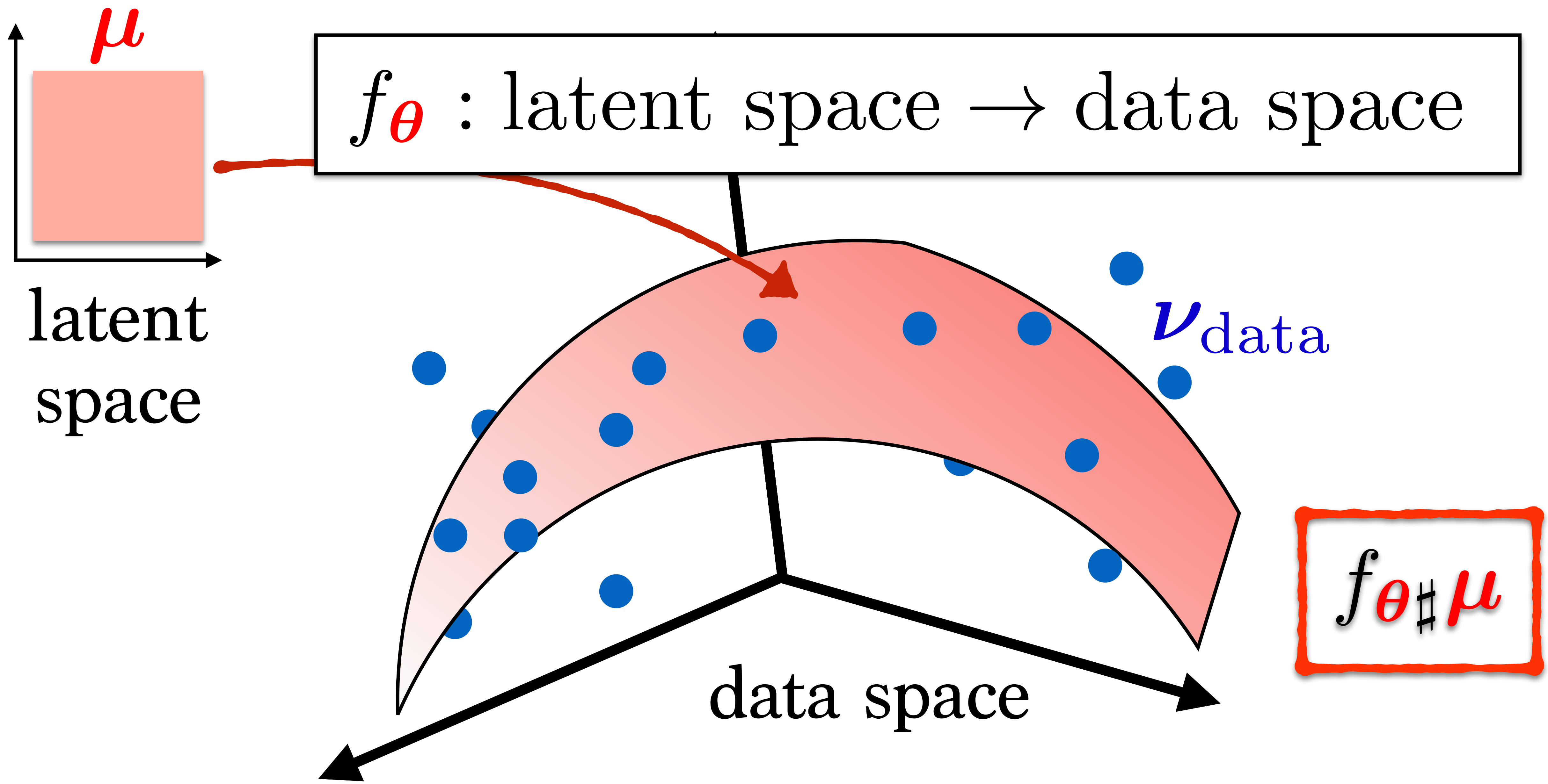
Goal: find $\boldsymbol{\theta}$ such that $f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu}$ fits $\boldsymbol{\nu}_{\mathrm{data}}$

Generative Models



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

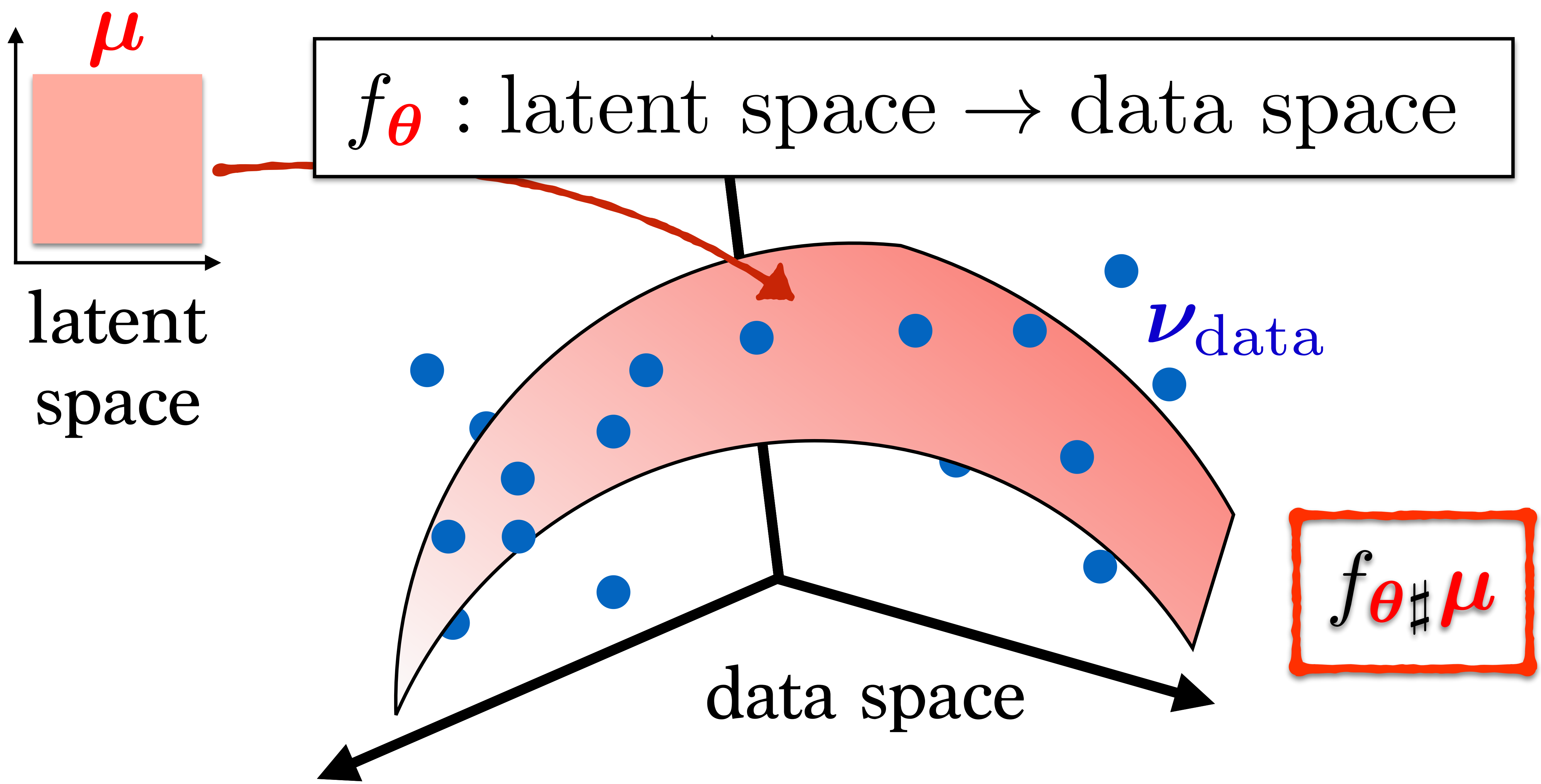
Generative Models



MLE

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

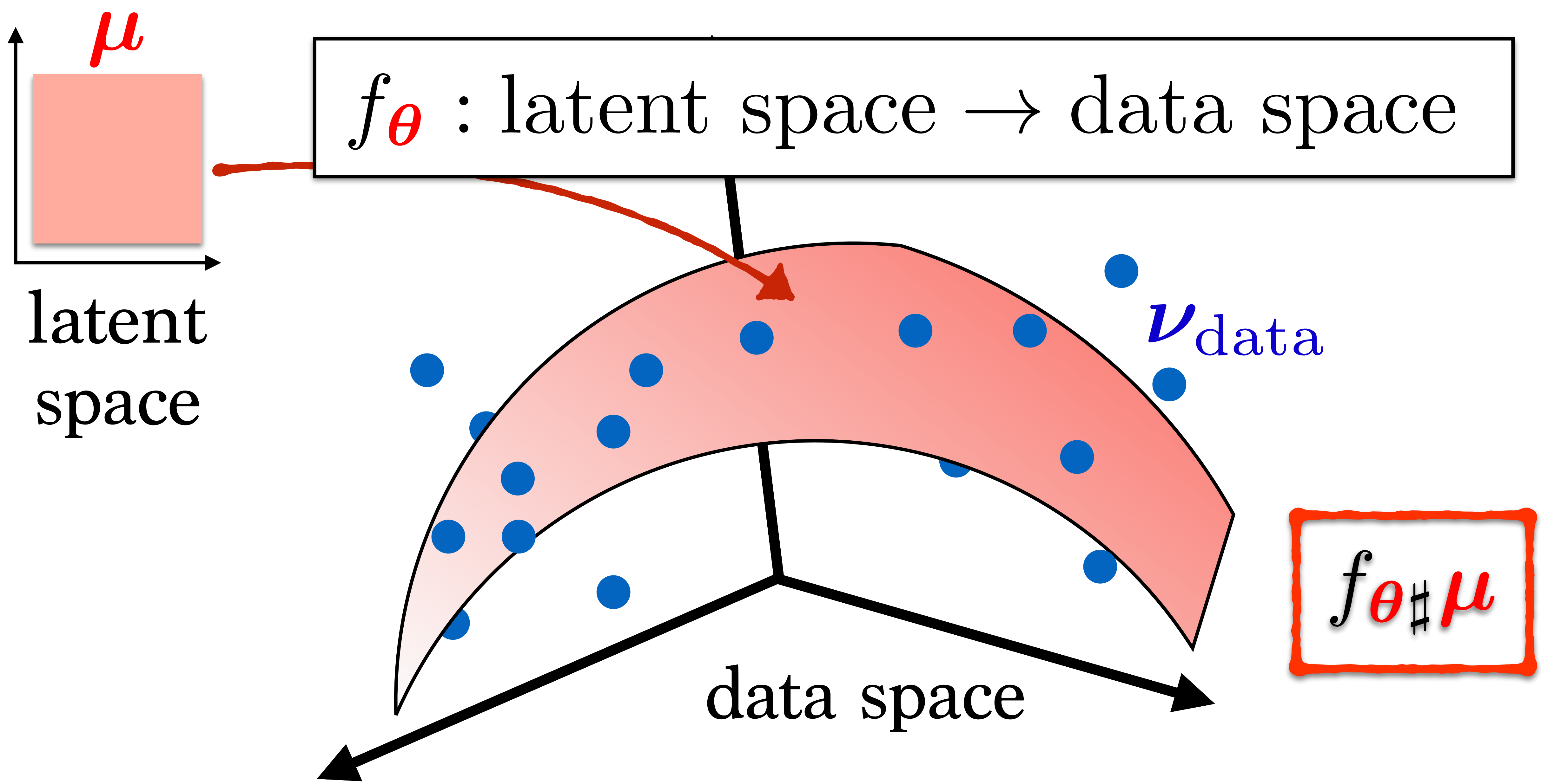
Generative Models



MLE

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

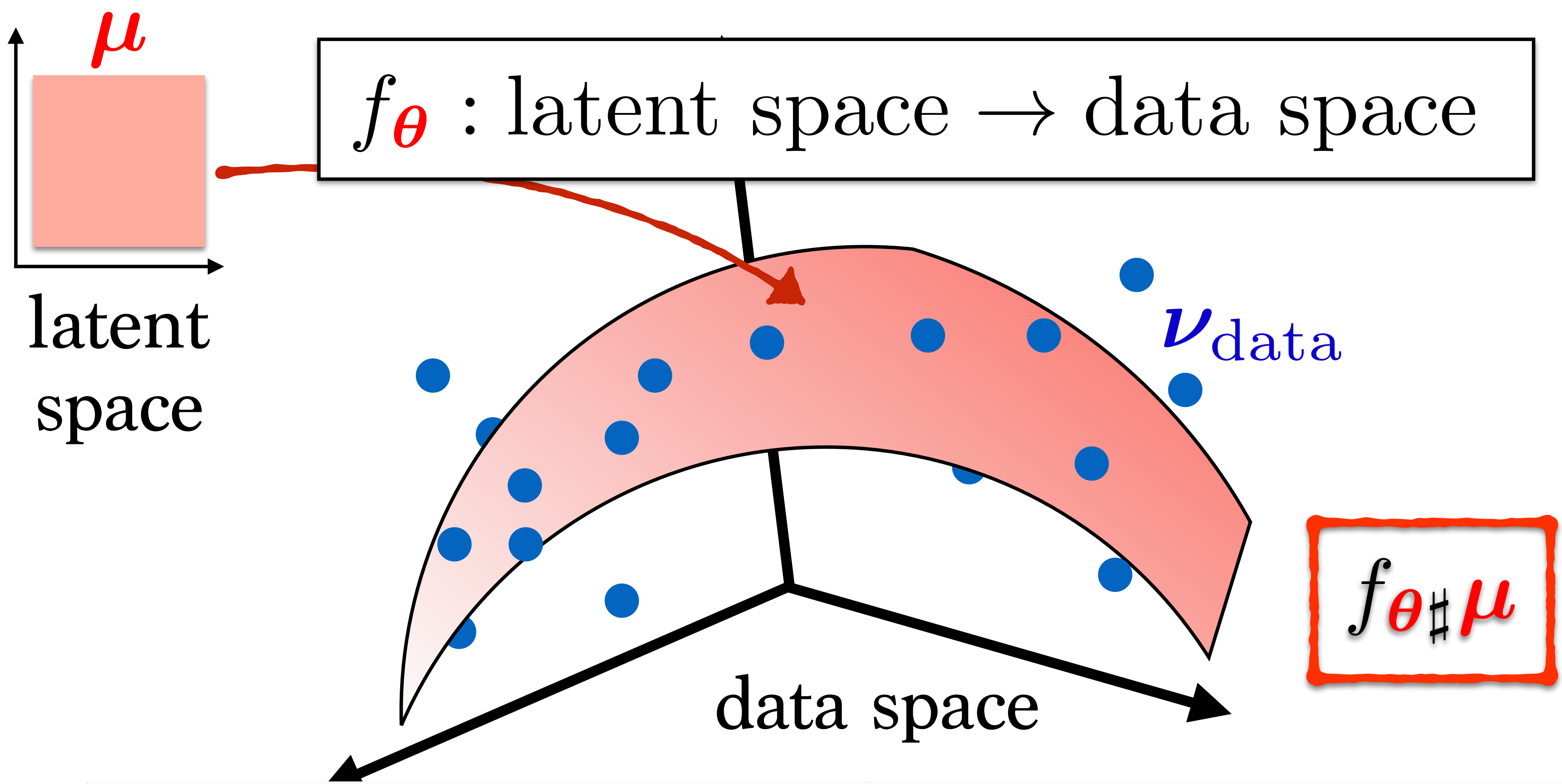
Generative Models



~~MLE~~

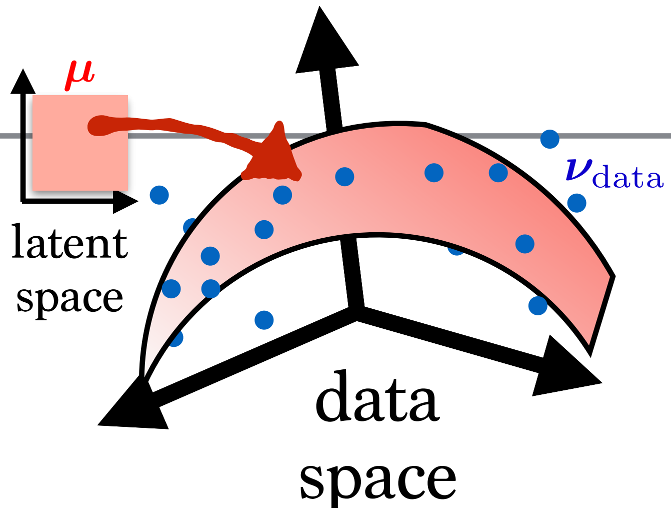
$$\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log f_{\theta\sharp}\mu(x_i) \qquad \min_{\theta\in\Theta} \mathrm{KL}(\nu_{\text{data}} \| f_{\theta\sharp}\mu)$$

Generative Models



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 measures, able to handle measures with non-overlapping supports.

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

Minimum △ Estimation

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

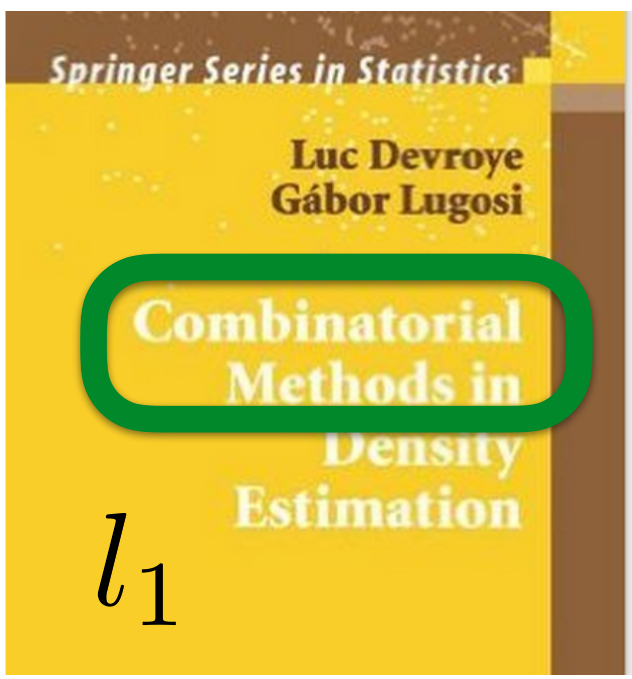
MINIMUM CHI-SQUARE, NOT MAXIMUM LIKELIHOOD!

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

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,*]

Use *Wasserstein distances* to define a loss between data and model.

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

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.

Challenge: $\nabla_{\boldsymbol{\theta}} W(\boldsymbol{\nu}_{\text{data}}, f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu})$?

[Montavon'16] use regularized OT in a finite setting.

[Arjovsky'17] *(WGAN)* uses a NN to approximate dual solutions and recover gradient w.r.t. parameter

[Bernton'17] *(Wasserstein ABC)*

[Genevay'17, Salimans'17] *(Sinkhorn approach)*

Concluding Remarks

1. Introduction to optimal transport
2. Optimal transport algorithms
3. Applications (W as a loss)
4. Applications (W for estimation)

This Saturday: **OT & ML Workshop**

7 Talks by: Jacob, Kraig, Andoni, Gangbo, Bottou, Flamary, Bach

17 posters and spotlight presentations.

Organizers: Bousquet, Cuturi, Peyré, Sha, Solomon

What we could not talk about...

- almost infinite supply of maths...
- **Statistical** challenges to compute W.
- If **linear assignment** = Wasserstein, then **quadratic assignment** = Gromov-Wasserstein.
- Wasserstein gradient flows (a.k.a. **JKO** flow).
- **Dynamical** aspects of optimal transport
- Transporting vectors and matrices

https://optimaltransport.github.io/