Regularized Optimal Transport is Ground Cost Adversarial

F-P. PATY M. CUTURI

ENSAE IP PARIS

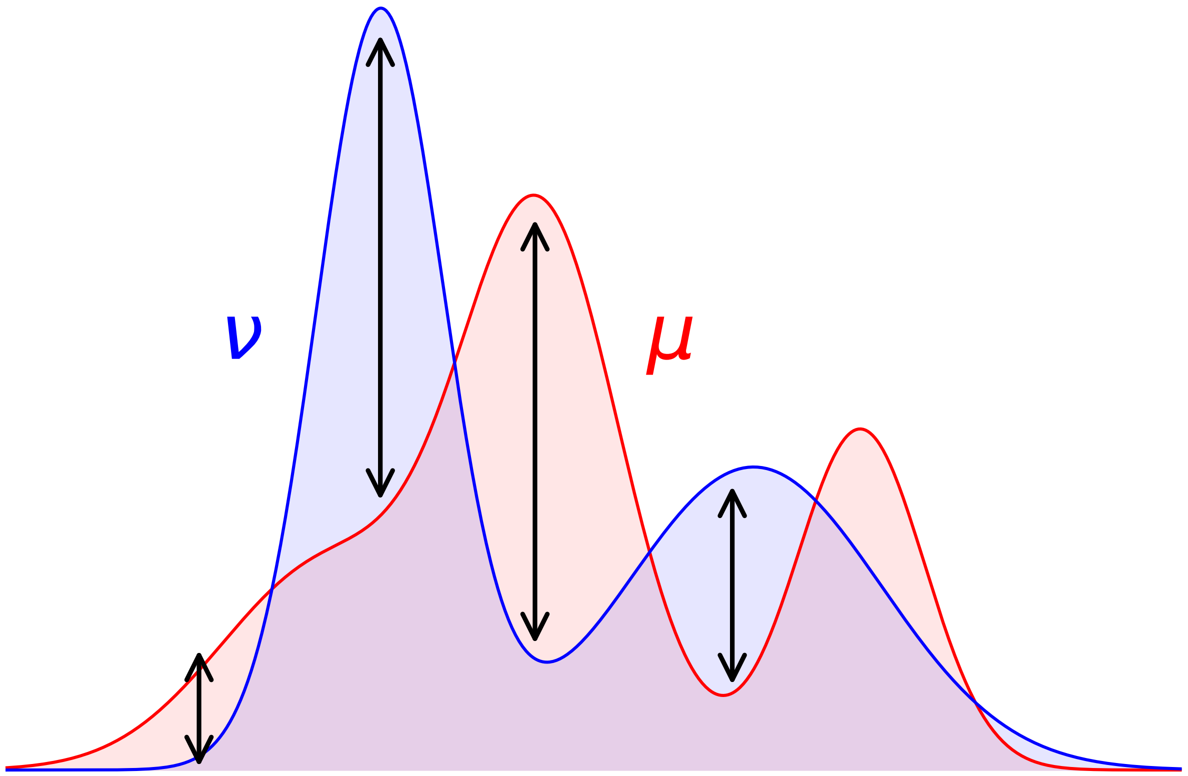
Google AI Brain Team

COMPARING DISTRIBUTIONS

1. Vertical comparison

Look at the difference, or the ratio of the densities

e.g. Total Variation distance, Kullback Leibler divergence, etc.

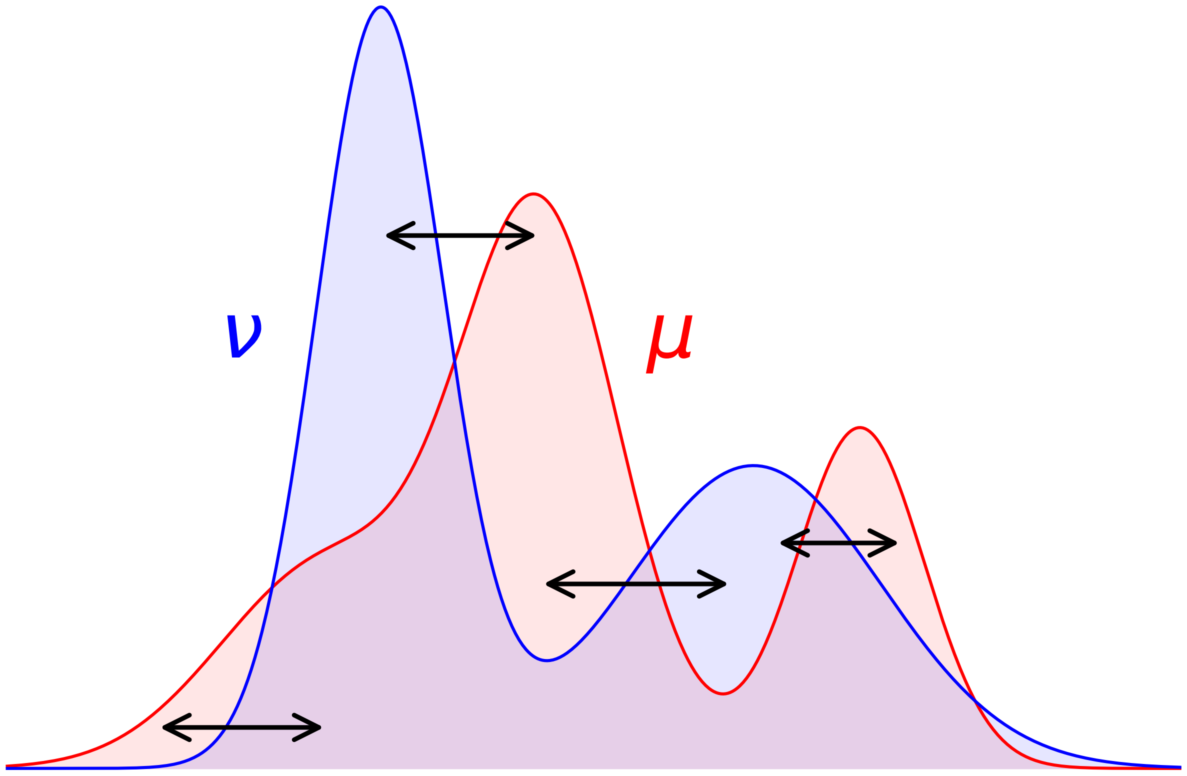


Comparing Distributions

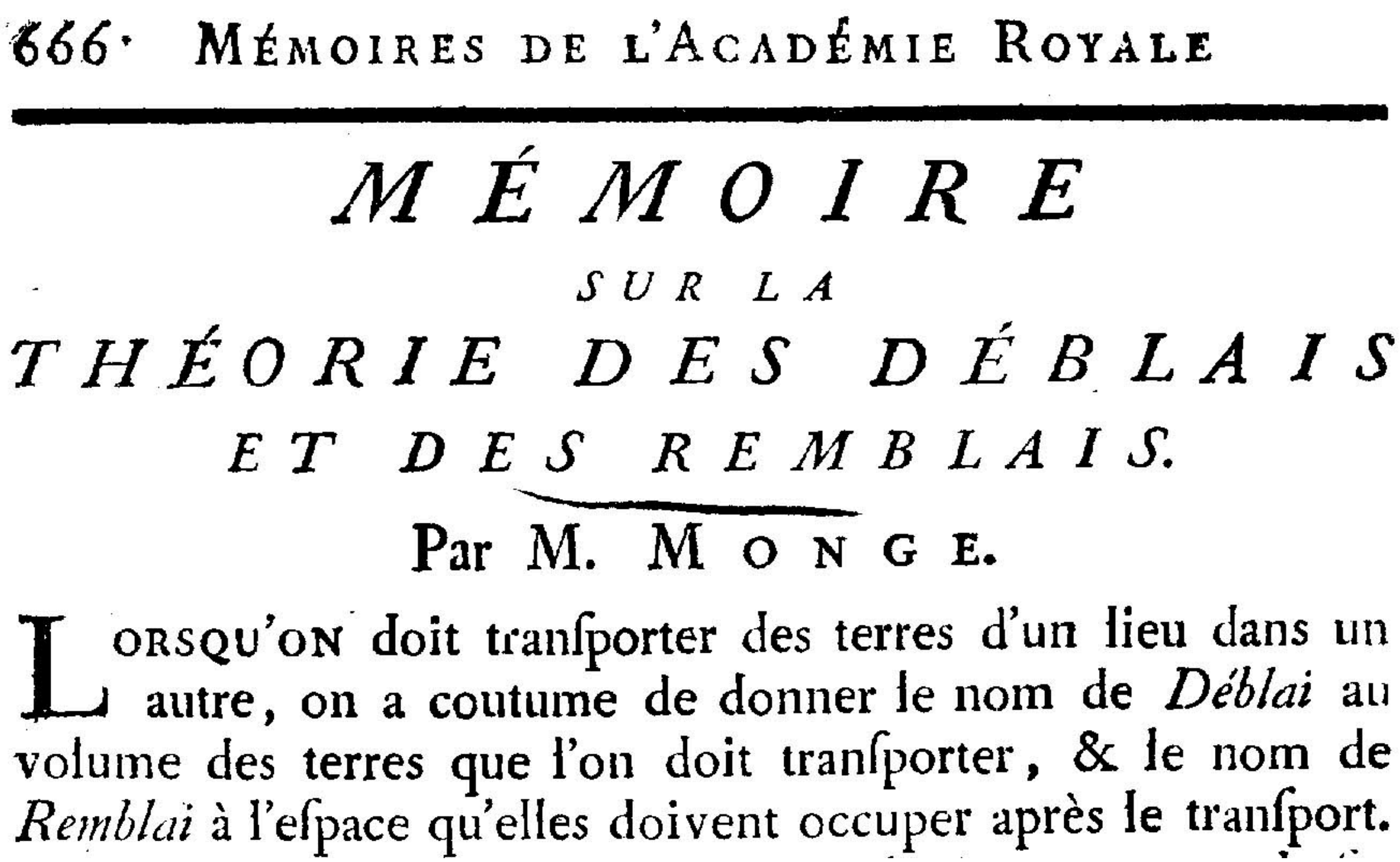
2. Horizontal comparison aka Optimal Transport

Move the mass across the ground space

⚠️ *Need for a notion of **displacement cost** on the ground space*



Some History



666· Mémoires de l'Académie Royale

MÉMOIRE

SUR LA

THÉORIE DES DÉBLAIS ET DES REMBLAIS.

Par M. Monge.

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

1781

1939

Some History

Brenier

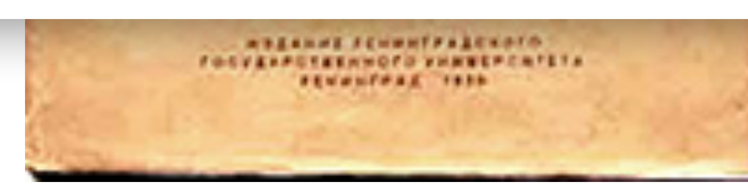
Otto

McCann

Villani

Figalli

Remblai à l'eſpace qu'elles doivent occuper après le tranſport.



1781

1939

Optimal Transport

Leonid Kantorovich

Optimal Transport

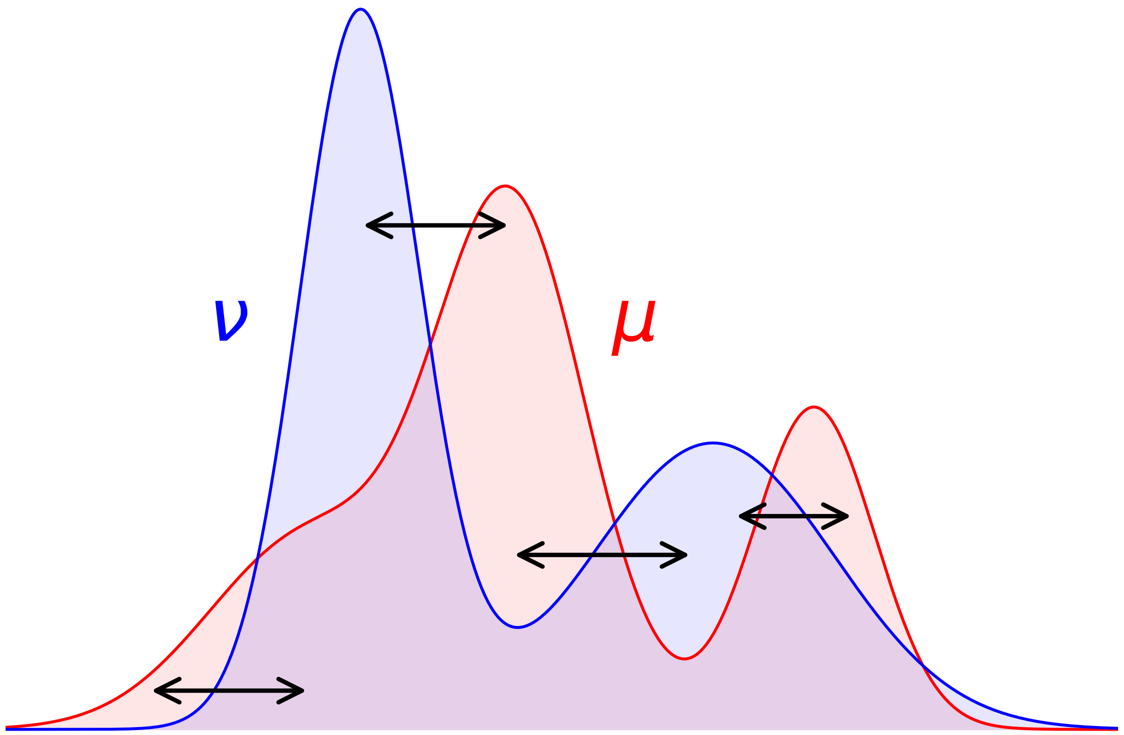
Data:

Two distributions μ and ν over $\mathbb{R}^d$

Parameter:

A (countinuous) cost function

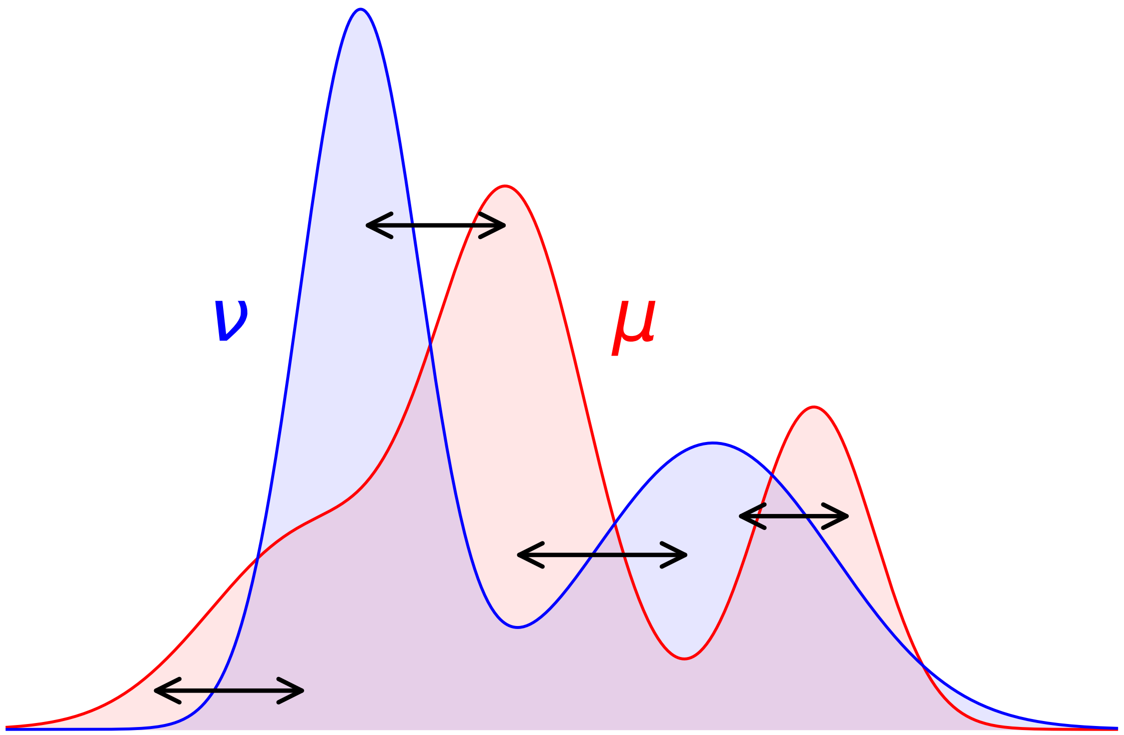
$$c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$



Optimal Transport

Data:

Two distributions μ and ν over $\mathbb{R}^d$



Parameter:

A (countinuous) cost function

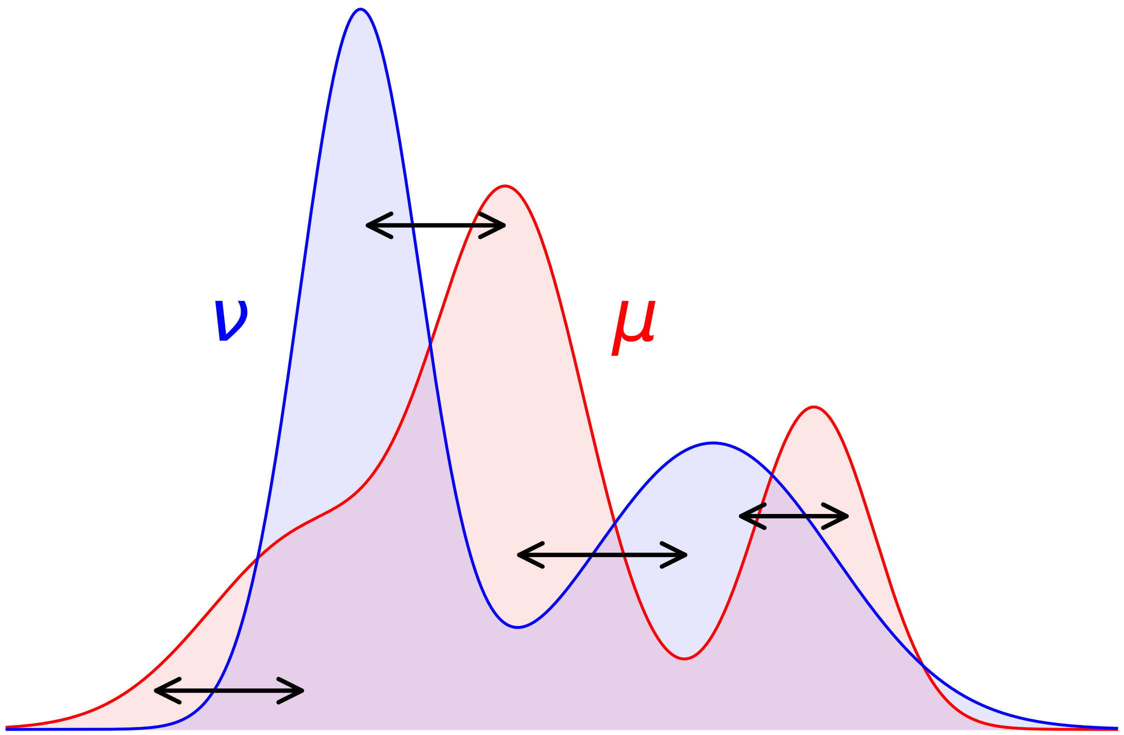
$$c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

Definition of Optimal Transport (OT):

Optimal Transport

Data:

Two distributions μ and ν over $\mathbb{R}^d$



Parameter:

A (countinuous) cost function

$$c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

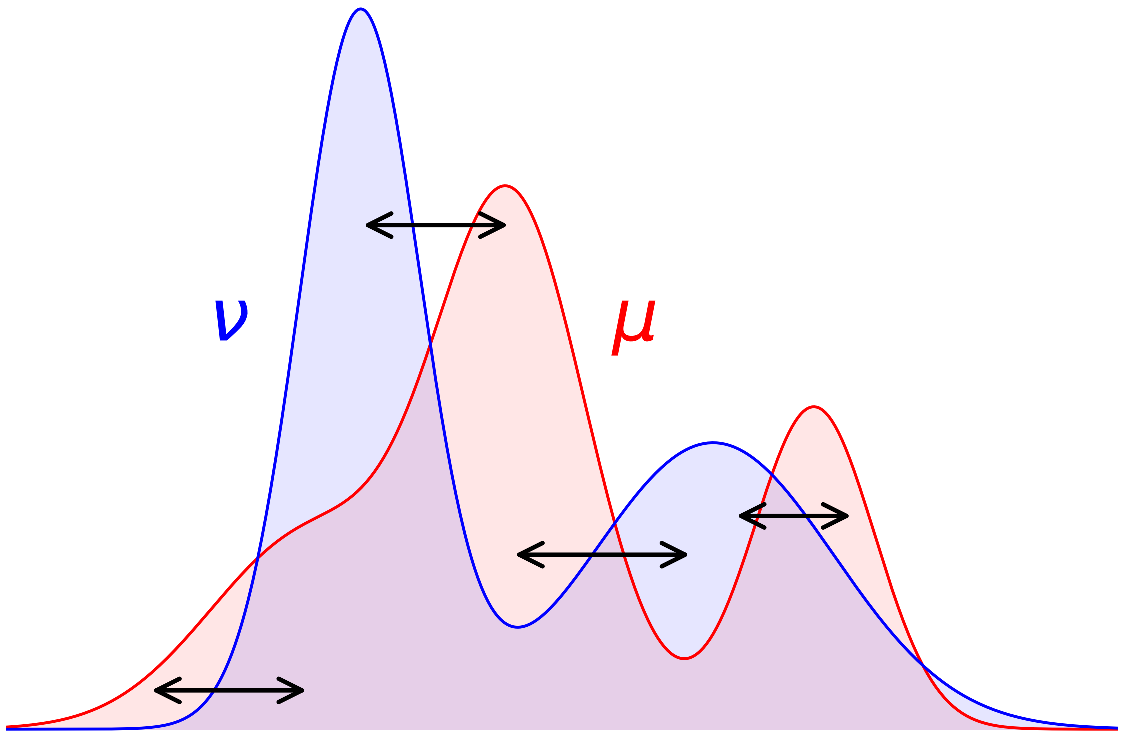
Definition of Optimal Transport (OT):

$$c(x, y)\, d\pi(x, y)$$

Optimal Transport

Data:

Two distributions μ and ν over $\mathbb{R}^d$



Parameter:

A (countinuous) cost function

$$c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

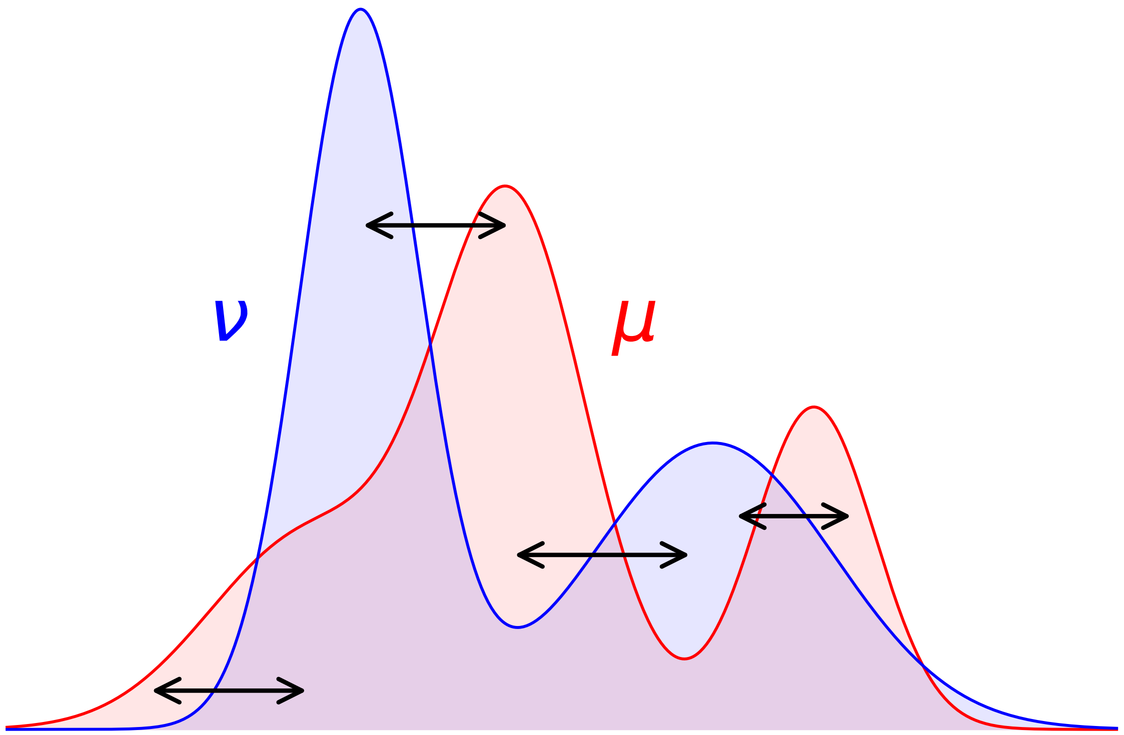
Definition of Optimal Transport (OT):

$$\iint c(x, y)\, d\pi(x, y)$$

Optimal Transport

Data:

Two distributions μ and ν over $\mathbb{R}^d$



Parameter:

A (countinuous) cost function

$$c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

Definition of Optimal Transport (OT):

$$\inf_{\pi} \iint c(x, y)\, d\pi(x, y)$$

Optimal Transport

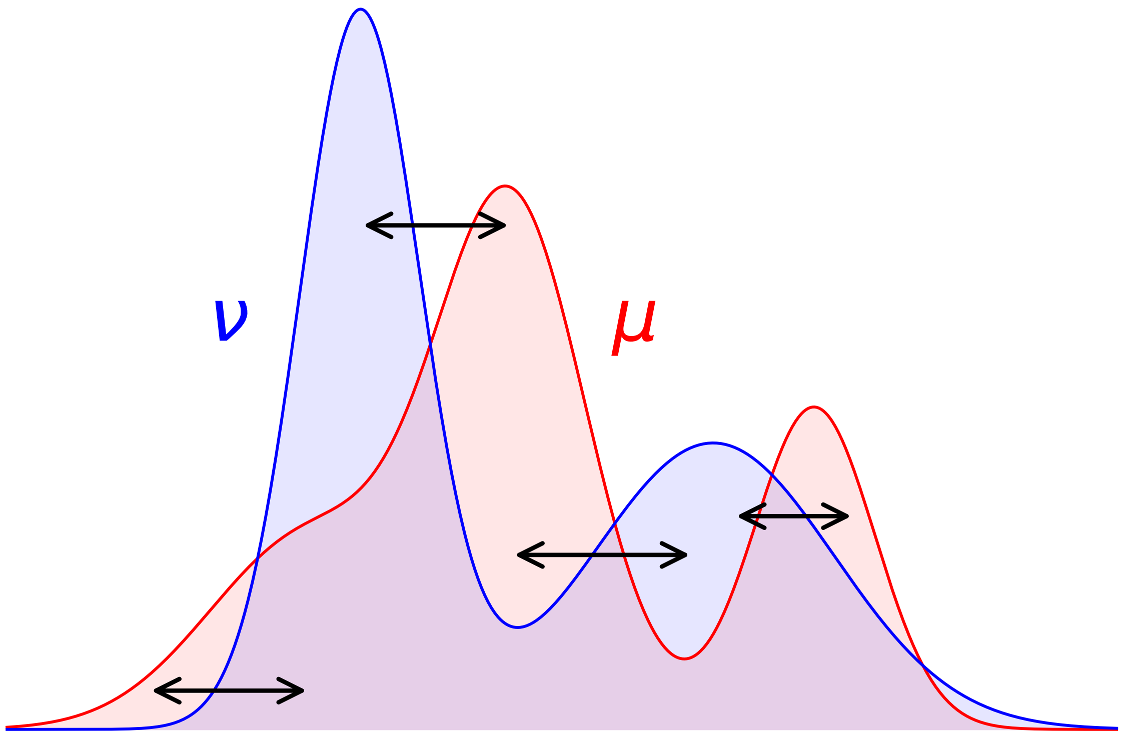
Data:

Two distributions μ and ν over $\mathbb{R}^d$

Parameter:

A (countinuous) cost function

$$c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$



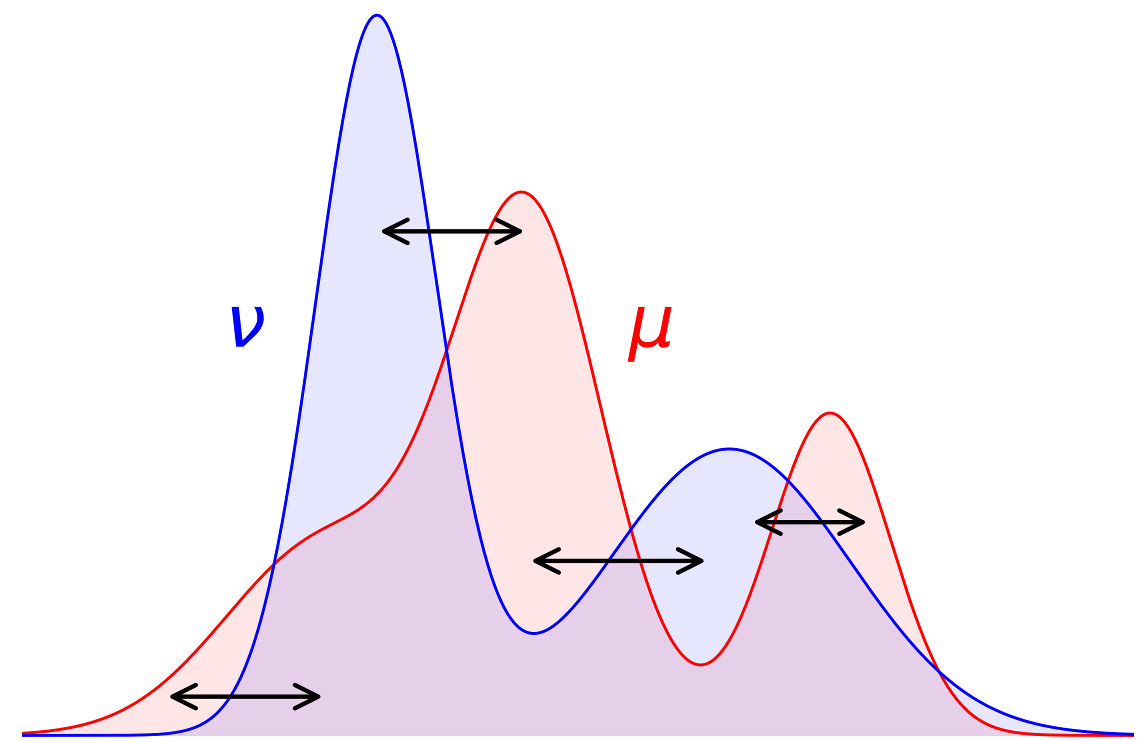
Definition of Optimal Transport (OT):

$$\mathscr{T}_c(\mu, \nu) = \inf_{\pi} \iint c(x, y)\, d\pi(x, y)$$

Optimal Transport

Data:

Two distributions μ and ν over $\mathbb{R}^d$



Parameter:

A (countinuous) cost function

$$c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

Definition of Optimal Transport (OT):

$$\mathscr{T}_c(\mu, \nu) = \inf_{\pi} \iint c(x, y)\, d\pi(x, y)$$

over all π *such that*

$$\begin{cases} \int d\pi(x, y) = d\mu(x) \quad \forall x \\ \int d\pi(x, y) = d\nu(y) \quad \forall y \end{cases}$$

Optimal Transport

Two main questions in practice

OPTIMAL TRANSPORT

Two main questions in practice

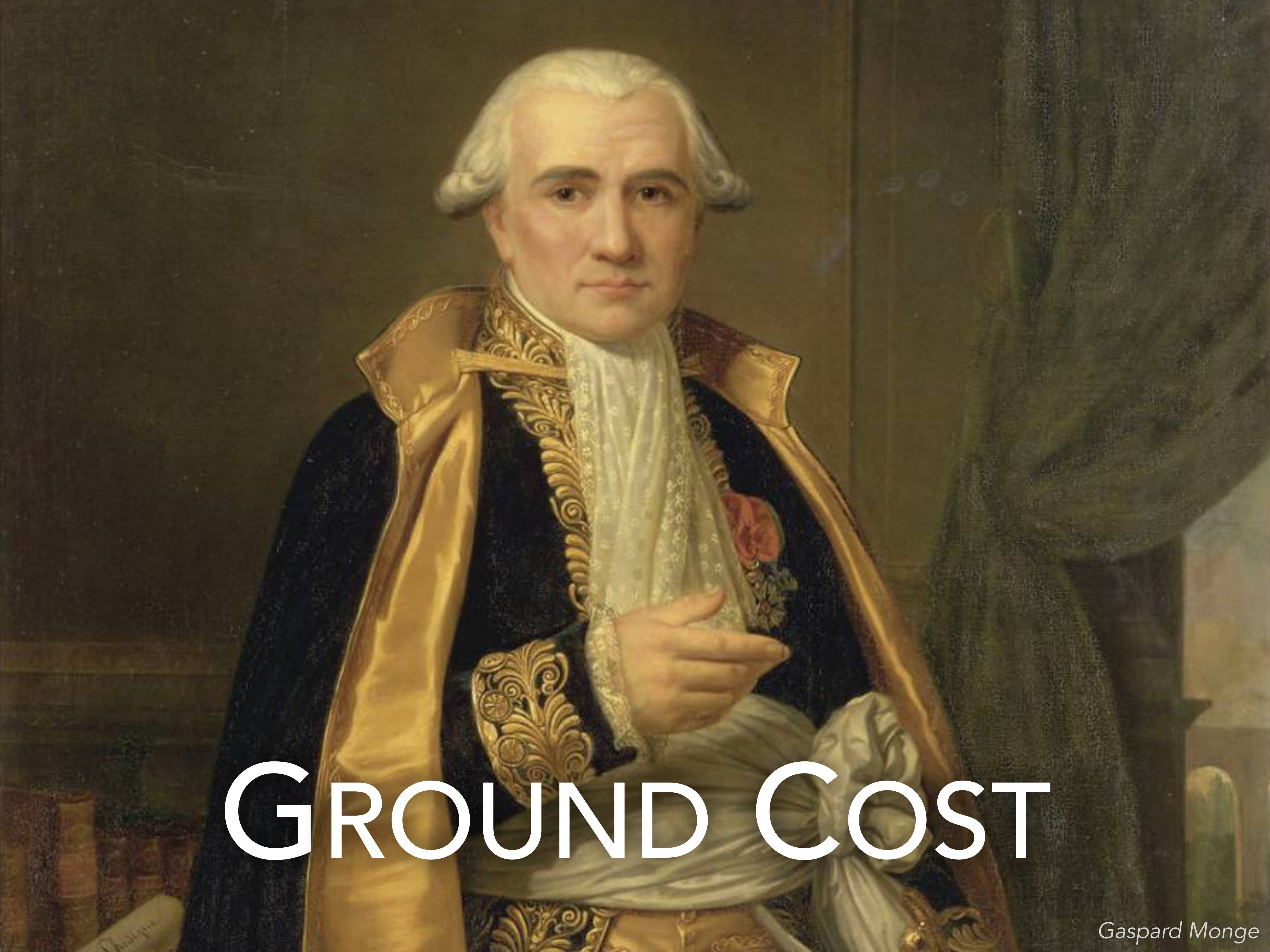
1. How to choose the ground cost c in a way that makes sense for the data distributions μ and ν ?

Optimal Transport

Two main questions in practice

1. How to choose the ground cost c in a way that makes sense for the data distributions μ and ν ?

2. How to compute/approximate the OT cost $\mathscr{T}_c(\mu, \nu)$, at least when the measures are discrete *(i.e. are finite sums of Dirac masses) in a scalable way*?


Ground Cost
Gaspard Monge

Ground Cost

1. How to choose the ground cost c in a way that makes sense for the data distributions μ and ν ?

Ground Cost

1. How to choose the ground cost c in a way that makes sense for the data distributions μ and ν?

1. Monge initially proposed $c(x, y) = \|x - y\|$

Ground Cost

1. How to choose the ground cost c in a way that makes sense for the data distributions μ and ν?

1. Monge initially proposed $c(x, y) = \|x - y\|$
2. This was generalized to cost functions of the form

$$c(x, y) = \|x - y\|^p \quad \text{where} \quad p \geq 1$$

(in this case, we say that $\mathcal{T}_c$ is the p-Wasserstein distance)

GROUND COST

1. How to choose the ground cost c in a way that makes sense for the data distributions μ and ν ?

1. Monge initially proposed $c(x, y) = \|x - y\|$
2. This was generalized to cost functions of the form

$$c(x, y) = \|x - y\|^p \quad \text{where} \quad p \geq 1$$

(in this case, we say that $\mathscr{T}_c$ is the p-Wasserstein distance)

But does it make sense when the ground space is high-dimensional ?

Ground Cost

1. How to choose the ground cost c in a way that makes sense for the data distributions μ and ν ?

1. Monge initially proposed $c(x,y) = \|x - y\|$
2. This was generalized to cost functions of the form

$$c(x,y) = \|x - y\|^p \quad \text{where} \quad p \geq 1$$

(in this case, we say that $\mathscr{T}_c$ is the p-Wasserstein distance)

But does it make sense when the ground space is high-dimensional ?

But does it make sense when the data lives on a low-dimensional manifold ?

GROUND COST

Idea: Find a ground cost c that is **adversarial**, *i.e.* that best separates the two distributions by maximizing the OT cost

$$\max_{c \in \mathscr{C}} \mathscr{T}_c(\mu, \nu) \qquad \text{where } \mathscr{C} \text{ is a class of functions}$$

GROUND COST

Idea: Find a ground cost c that is **adversarial**, *i.e.* that best separates the two distributions by maximizing the OT cost

$$\max_{c \in \mathscr{C}} \mathscr{T}_c(\mu, \nu)$$ where $\mathscr{C}$ is a class of functions

$$\max_{c} \mathscr{T}_c(\mu, \nu) - f(c)$$ for some convex f

$$f(c) = \begin{cases} 0 & \text{if } c \in \mathscr{C} \\ +\infty & \text{if } c \notin \mathscr{C} \end{cases}$$

Ground Cost

Idea: Find a ground cost c that is **adversarial**, *i.e.* that best separates the two distributions by maximizing the OT cost

$$\max_{c} \mathscr{T}_c(\mu, \nu) - f(c) \quad \text{for some convex } f$$

- Links with the Robust Optimization literature
- Links with the matchings literature in Economics
- Initially proposed by Genevay *et al.* in 2017 to learn generative models
- When $\mathscr{C}$ is the set of Mahalanobis distances, it defines the Subspace Robust Wasserstein distances (ICML 2019)


REGULARIZATION
Alessio Figalli

REGULARIZATION

2. How to compute/approximate the OT cost $\mathscr{T}_c(\mu, \nu)$?

1. This is a Linear Program $\longrightarrow$ $\mathcal{O}(n^3)$ complexity

2. Entropic regularization $\longrightarrow$ $\mathcal{O}(n^2)$ Sinkhorn algorithm, GPU-friendly, differentiable…

$$\inf_{\pi} \iint c(x, y)\, d\pi(x, y) + \varepsilon R(\pi)$$

where $R(\pi) = \mathrm{KL}(\pi || \mu \otimes \nu)$

Other regularizations have been proposed: *e.g.* quadratic, group-lasso, capacity constraints, with different algorithms and effects on the OT plan / value

REGULARIZATION

2. How to compute/approximate the OT cost $\mathscr{T}_c(\mu, \nu)$?

1. This is a Linear Program $\longrightarrow$ $\mathcal{O}(n^3)$ complexity

2. Entropic regularization $\longrightarrow$ $\mathcal{O}(n^2)$ Sinkhorn algorithm, GPU-friendly, differentiable…

$$\inf_{\pi} \iint c(x,y)\, d\pi(x,y) + \varepsilon R(\pi)$$

How can we interpret the effect of the regularization ?

Two views of the same phenomenon

Ground Cost Robustness ⇔ Regularization

Theorem: Regularized OT is ground cost adversarial in the following sense

$$\inf_{\pi} \iint c_0(x,y)\, d\pi(x,y) + \varepsilon R(\pi)$$

$$= \sup_{c} \mathscr{T}_c(\mu,\nu) - \varepsilon R^* \left(\frac{c - c_0}{\epsilon} \right)$$

where R is a convex regularizer

and R^* is the convex conjugate of R:

$$R^*(c) = \sup_{\pi} \int c\, d\pi - R(\pi)$$

Ground Cost Robustness ⇔ Regularization

Theorem: Regularized OT is ground cost adversarial in the following sense

$$
\inf_{\pi} \iint c_0(x, y)\, d\pi(x, y) + \varepsilon R(\pi)
$$

$$
= \sup_{c} \mathscr{T}_c(\mu, \nu) - \varepsilon R^* \left(\frac{c - c_0}{\epsilon} \right)
$$

Is the adversarial cost $c_\star$ an interesting dissimilarity measure on the ground space ?

Ground Cost Robustness ⇔ Regularization

Is the adversarial cost $c_\star$ an interesting dissimilarity measure on the ground space ?

Short answer: In a sense, no.

GROUND COST ROBUSTNESS ⇔ REGULARIZATION

Is the adversarial cost $c_\star$ an interesting dissimilarity measure on the ground space ?

Short answer: In a sense, no.

Theorem: Under some technical assumption on R *(verified for the entropic or quadratic regularizations)*, there exists functions ϕ and ψ such that

$$c : (x, y) \mapsto \phi(x) + \psi(y)$$

is an optimal adversarial cost, *i.e.* is solution to

$$\sup_{c} \mathscr{T}_c(\mu, \nu) - \varepsilon R^* \left(\frac{c - c_0}{\epsilon} \right)$$

WHAT I COULD NOT TALK ABOUT

- Restriction to nonnegative adversarial costs $\sup_{c \geq 0} \ldots$
- General duality result for regularized OT
- Extension to several measures

Thank you

francoispierrepaty.github.io