Optimal Transport for Signal Processing

A tutorial at MLSP 2024

Felipe Tobar[1] Laetitia Chapel[2]

[1]Initiative for Data & Artificial Intelligence, Universidad de Chile

[2]IRISA, Obelix team, Institut Agro Rennes-Angers

22 September, 2024

github.com/felipe-tobar/OT-tutorial-MLSP-2024

Overview

1. Introduction
2. Part I: The Optimal Transport Problem
3. Part II: The Wasserstein distance
4. Closing remarks

Speaker's presentation

Felipe Tobar

Associate Professor
IDIA, Universidad de Chile

Research themes: Gaussian Processes, Optimal Transport, Diffusion Models
`www.dim.uchile.cl/~ftobar`

Laetitia Chapel

Full Professor in Computer Science
IRISA Lab, France

Research themes: Optimal Transport, machine learning on structured data
`people.irisa.fr/Laetitia.Chapel`

Forenote on implementation

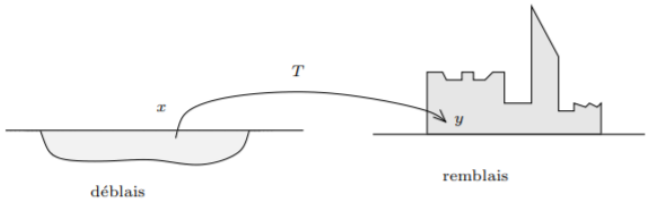
- Examples in this tutorial are based on **POT: Python Optimal Transport Toolbox**
- Available here `pythonot.github.io`

Why Optimal Transport

State main advantages and show applications in different fields. Emphasis on why OT and not other techniques

Origins of OT: Gaspard Monge (1781)

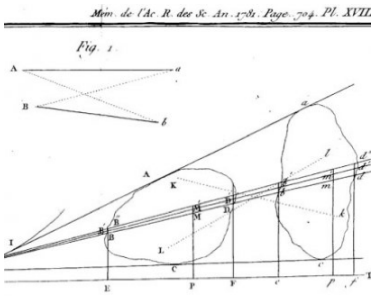
How to transport a pile of sand onto a hole in an optimal way?



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.

Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'enfuit que le déblai & le remblai étant donnés de figure & de position, il n'eſt pas indifférent que telle molécule du déblai ſoit tranſportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine diſtribution à faire des molécules du premier dans le ſecond, d'après laquelle la ſomme de ces produits ſera la moindre poſſible, & le prix du tranſport total ſera un *minimum*.



Brief history of OT

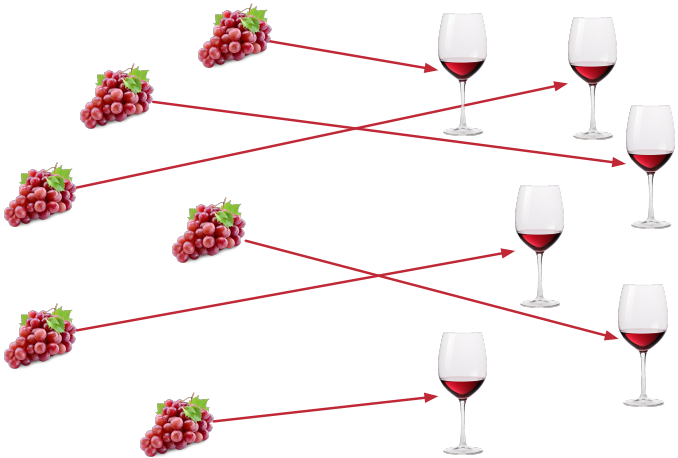
From Monge to today

Part I: The Optimal Transport Problem

Intuition

Motivation: vineyards transporting grapes from harvest site to processing plants

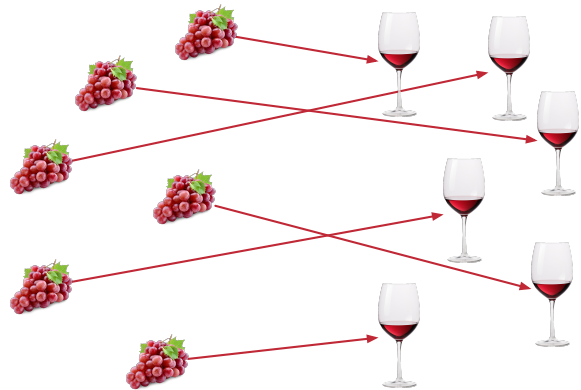
The assigment problem



Build the assignment problem from intuition. Use the above figure to explain all possible ways to assign: straight lines, what's the cost, *cheapest transport*.

The assigment problem: encoding real-world

- Weighted masses
- Different number of sources/targets
- Straight path is not possible
- New sample becomes available



Monge formulation[1]

Objective: Move a pile of mass from one location to another at a minimum effort

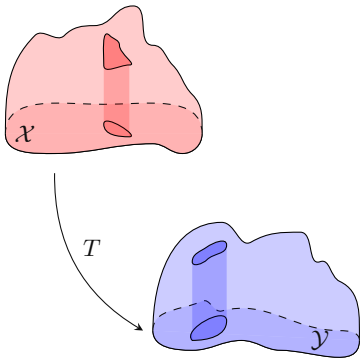
Let us first set up our notation

- **Piles of mass** are probability distributions, μ and ν, corresponding to random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$.
- **Moving procedure** is a function $T : x \in \mathcal{X} \mapsto Y \in \mathcal{Y}$.
- **Moving cost** encoded as $c : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto c(x, y) \in \mathbb{R}$.

Solve: Optimise the total transport cost

$$\sum_{x \in \mathcal{X}} c(x_i, T(x_i)) \tag{1}$$

over $M_{X,Y} = \{T : \mathcal{X} \to \mathcal{Y},\ s.t.,\ T_{\#\mu} = \nu\}$.



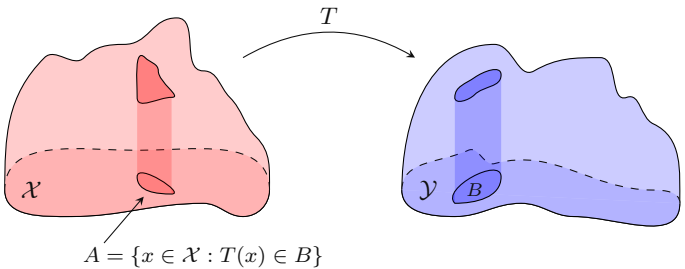
[1] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. De l'Imprimerie Royale.

The transport map (aka the *pushforward* operator $T_{\#}$)

T *transports* mass from $\mathcal{X}$ to $\mathcal{Y}$, meaning that for any subset $A \in \mathcal{Y}$, one has

$$\nu(A) = \mu(T^{-1}(A)), \tag{2}$$

where $T^{-1}(A) = \{x \in \mathcal{X}, s.t.\ T(x) \in A\}$ is the preimage of A under T.



$T\#\mu = \nu$

i.e., $\mu(A) = \nu(B)$

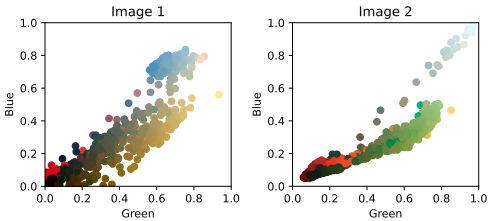
Figure adapted from Thorpe's book.[2]

[2] Infinite thanks to Elsa Cazelles (IRIT, CNRS) for kindly sharing these beautiful `tikz` figures.

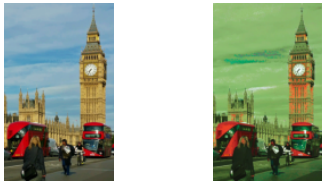
Example 1: Colour transfer

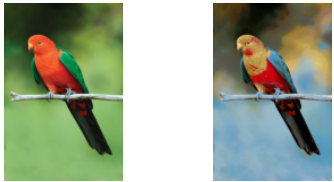
Original images

Histograms



Histograms

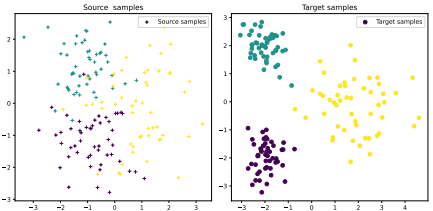




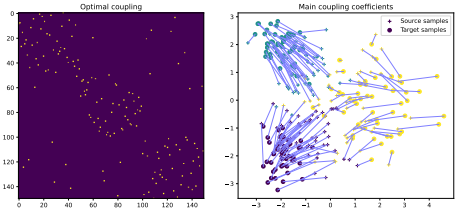
Notebook: Colour_transfer.ipynb

Example 2: Domain adaptation

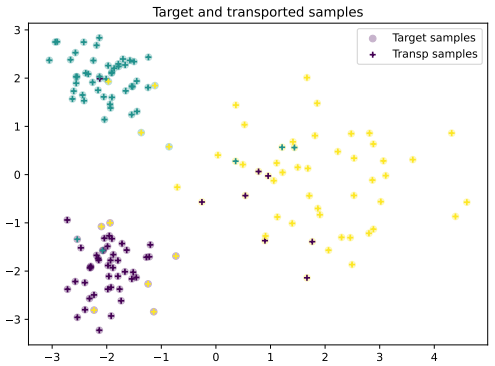
Original images



Histograms

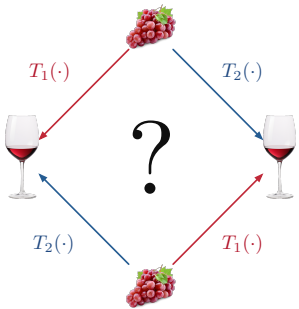


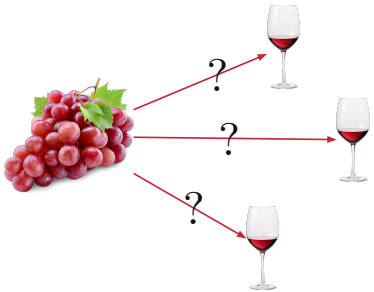
Histograms



Notebook: Domain_adaptation.ipynb

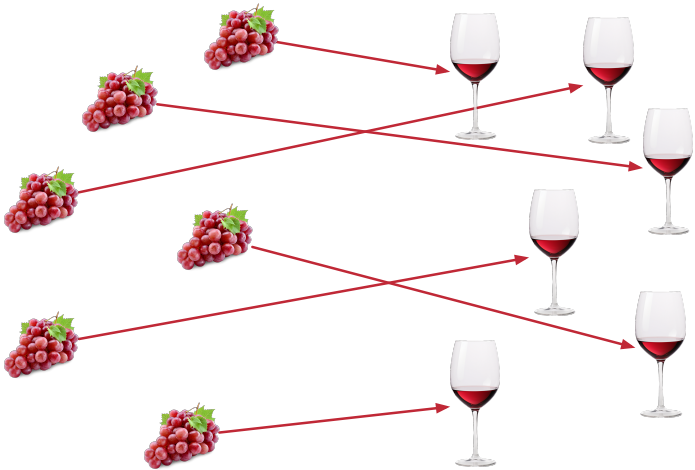
Neither existence nor uniqueness is guaranteed



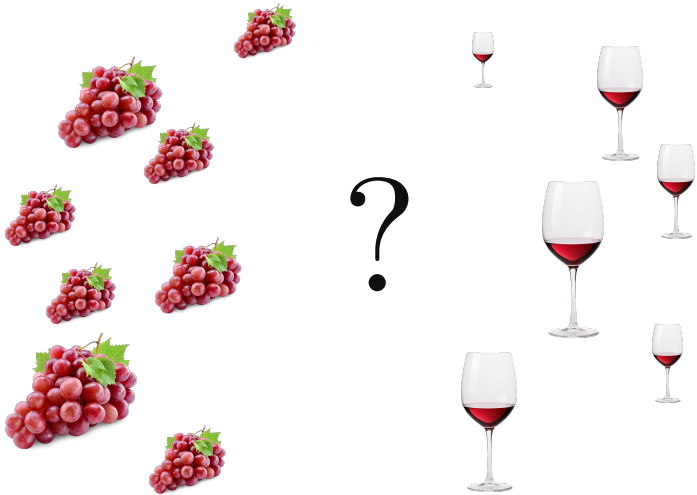


Observation: In the two examples above, each sample *weights the same*, i.e., pixels, class instances. In some cases, we might have *weighted samples*. In such cases, **Monge's map might be unable to transport the mass.**

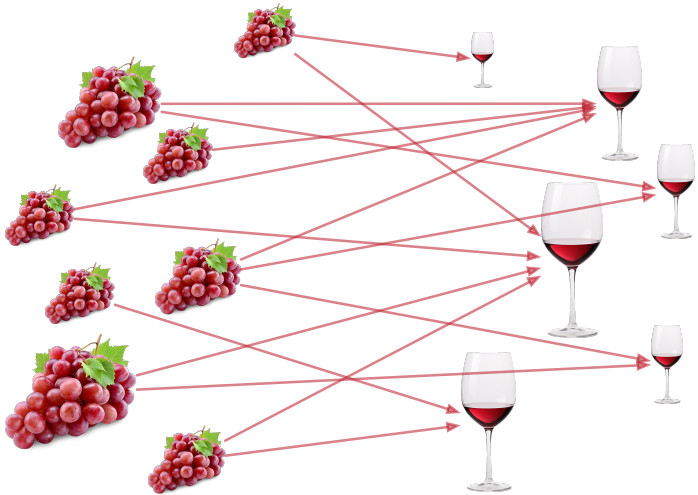
Kantorovich formulation: mass splitting



Kantorovich formulation: mass splitting



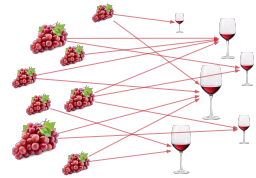
Kantorovich formulation: mass splitting

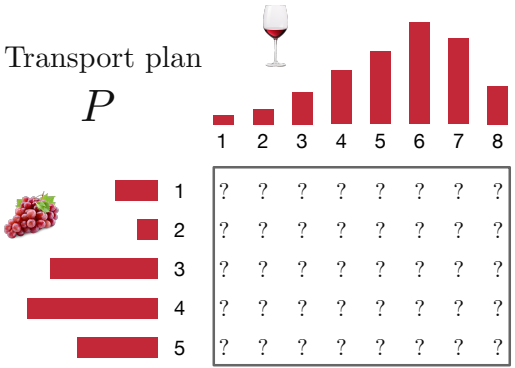


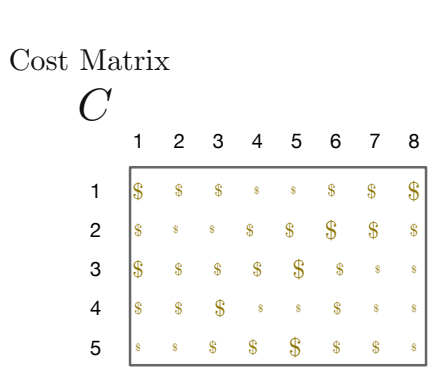
Transport plan

$$\inf_{P \in \Pi_{\mu,\nu}} \langle P, C \rangle = \sum_{i,j}^{n,m} C_{ij} P_{ij}$$

where $\Pi_{\mu,\nu} \langle P, C \rangle = \{P \in [0,1]^{m \times n} : \sum_{i=1}^{m} P_{ij} = \nu_j, \sum_{j=1}^{n} P_{ij} = \mu_i\}$

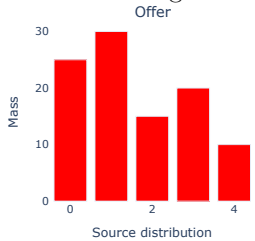


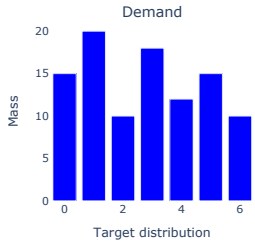


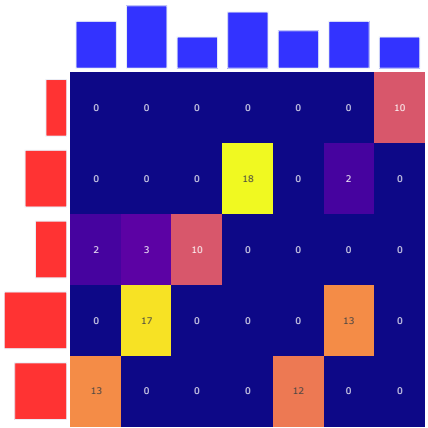


Example 3: Discrete Kantorovich plan

Let consider the following source and target distributions



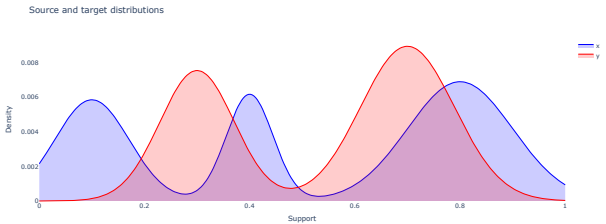


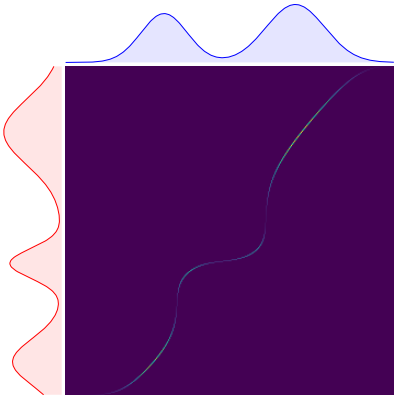


Notebook: kantorovich.ipynb

Example 4: Continuous Kantorovich plan

Let us now consider two distributions over a continuous support





Observe that the plan remained *sparse*, i.e., the mass did not spread much

This motivates the following results

Notebook: kantorovich.ipynb

Observations

- Let us consider a cost $c(x, y) = |x - y|^p, p \geq 1$. Then, if μ and ν are absolutely continuous wrt the Lebesgue measure, the Kantorovich problem as a unique solution. Furthermore, this solution is the same solution of the Monge problem.
- If $p = 2$, the optimal map is the gradient of a convex function
- In some cases the optimal plan will require to split mass (e.g., in the case of atomic measures) and thus Monge's solution may fail to exist.
- Luckily, from a (Kantorovich) transport plan we can always extract a tranport map, e.g., via the barycentric projection

Dual formulation

Recall the primal formulation: $OT(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \iint c(x,y)\mathrm{d}\pi(x,y)$

$$OT(\mu,\nu) = \sup_{(\phi,\psi)\in\Phi_c} \left(\int_{\mathcal{X}} \phi d\mu + \int_{\mathcal{X}} \psi d\nu \right)$$

where

$$\Phi_c := \{(\phi,\psi) \in L_1(\mu) \times L_1(\nu) \text{ s.t. } \phi(x) + \psi(y) \leq c(x,y)\}$$

Dual formulation

Recall the primal formulation: $OT(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \iint c(x,y)\mathrm{d}\pi(x,y)$

$$OT(\mu,\nu) = \sup_{(\phi,\psi)\in\Phi_c} \left(\int_{\mathcal{X}} \phi d\mu + \int_{\mathcal{X}} \psi d\nu \right)$$

where

$$\Phi_c := \{(\phi,\psi) \in L_1(\mu) \times L_1(\nu) \text{ s.t. } \phi(x) + \psi(y) \leq c(x,y)\}$$

- ϕ and ψ are scalar function also known as Kantorovich potentials
- Primal dual relationship : the support of $\pi \in \Pi^*(\mu,\nu)$, that is $\pi(x,y)$ is where $\phi(x) + \psi(y) = c(x,y)$.

Dual formulation

Recall the primal formulation: $OT(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \iint c(x,y)\mathrm{d}\pi(x,y)$

$$OT(\mu,\nu) = \sup_{(\phi,\psi)\in\Phi_c} \left(\int_{\mathcal{X}} \phi d\mu + \int_{\mathcal{X}} \psi d\nu \right)$$

where

$$\Phi_c := \{(\phi,\psi) \in L_1(\mu) \times L_1(\nu) \text{ s.t. } \phi(x) + \psi(y) \leq c(x,y)\}$$

- ϕ and ψ are scalar function also known as Kantorovich potentials
- Primal dual relationship : the support of $\pi \in \Pi^*(\mu,\nu)$, that is $\pi(x,y)$ is where $\phi(x) + \psi(y) = c(x,y)$.

In the discrete setting:

$$\int_{\mathcal{X}} \phi d \left(\sum_{i=1}^{n} a_i \delta_{x_i} \right) + \int_{\mathcal{X}} \psi d \left(\sum_{j=1}^{m} b_j \delta_{y_i} \right) = \sum_{i=1}^{n} a_i \underbrace{\phi(x_i)}_{\alpha_i} + \sum_{j=1}^{m} b_j \underbrace{\psi(y_j)}_{\beta_j}$$

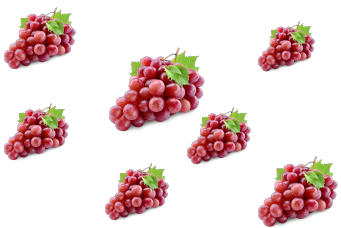
and Φ_c becomes $\{(\alpha,\beta) \in \mathbb{R}^n \times \mathbb{R}^m \text{ s.t. } \alpha_i + \beta_j \leq c(x_i,y_j)\}$

Informal way of interpreting Kantorovich duality principle for the discrete case

$$OT(\mu,\nu) = \min_{\pi \in \Pi(a,b)} \langle C, \pi \rangle = \max_{(\alpha,\beta)\in D_c} \langle \alpha, a \rangle + \langle \beta, b \rangle$$

with

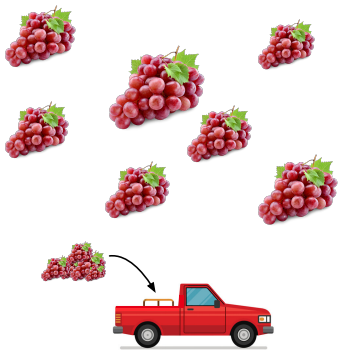
$$D_c := \{(\alpha,\beta) \in \mathbb{R}^n \times \mathbb{R}^m \text{ such that } \forall (i,j) \in \{1,\dots,n\} \times \{1,\dots,m\}, \alpha_i + \beta_j \leq C_{ij}\}$$



Intuition from Caffarelli : the shipper's problem

One vendor comes and sets her principle:

- α_i = price for **loading** a unit of coal at place x_i (no matter where it goes)
- β_j = price for **unloading** a unit of coal at place y_j (no matter from which mines it comes from)



Intuition from Caffarelli : the shipper's problem

- There are exactly a_i units at mine x_i and b_j needed at factory y_j; the vendor asks the price (that she wants to maximize!)

$$\langle \alpha, a \rangle + \langle \beta, b \rangle$$

Intuition from Caffarelli : the shipper's problem

- There are exactly a_i units at mine x_i and b_j needed at factory y_j; the vendor asks the price (that she wants to maximize!)

$$\langle \alpha, a \rangle + \langle \beta, b \rangle$$

- Negative price are allowed !

Intuition from Caffarelli : the shipper's problem

- There are exactly a_i units at mine x_i and b_j needed at factory y_j; the vendor asks the price (that she wants to maximize!)

$$\langle \alpha, a \rangle + \langle \beta, b \rangle$$

- Negative price are allowed !
- Does the vendor have a competitive offer? Her pricing scheme implies that transferring one unit of coal from mine x_i to factory y_j costs exactly $\alpha_i + \beta_j$.

Intuition from Caffarelli : the shipper's problem

- There are exactly a_i units at mine x_i and b_j needed at factory y_j; the vendor asks the price (that she wants to maximize!)

$$\langle \alpha, a \rangle + \langle \beta, b \rangle$$

- Negative price are allowed !
- Does the vendor have a competitive offer? Her pricing scheme implies that transferring one unit of coal from mine x_i to factory y_j costs exactly $\alpha_i + \beta_j$.
- Recall the primal problem : the cost of shipping one unit from x_i to y_j is $C_{i,j}$.

Intuition from Caffarelli : the shipper's problem

- There are exactly a_i units at mine x_i and b_j needed at factory y_j; the vendor asks the price (that she wants to maximize!)

$$\langle \alpha, a \rangle + \langle \beta, b \rangle$$

- Negative price are allowed !
- Does the vendor have a competitive offer? Her pricing scheme implies that transferring one unit of coal from mine x_i to factory y_j costs exactly $\alpha_i + \beta_j$.
- Recall the primal problem : the cost of shipping one unit from x_i to y_j is $C_{i,j}$.
- Good deal for the driver implies $\alpha_i + \beta_j \leq C_{ij}$.

Intuition from Caffarelli : the shipper's problem

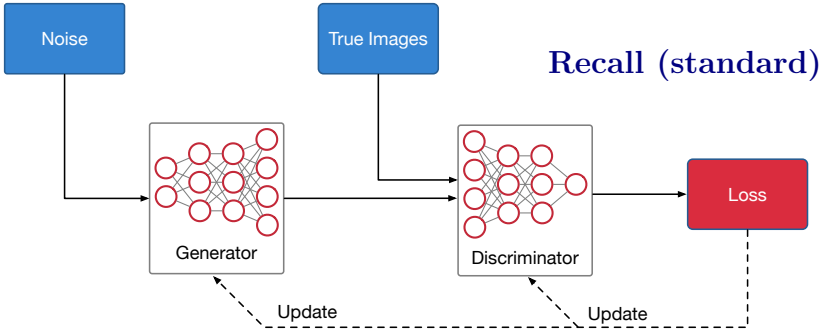
- There are exactly a_i units at mine x_i and b_j needed at factory y_j; the vendor asks the price (that she wants to maximize!)

$$\langle \alpha, a \rangle + \langle \beta, b \rangle$$

- Negative price are allowed !
- Does the vendor have a competitive offer? Her pricing scheme implies that transferring one unit of coal from mine x_i to factory y_j costs exactly $\alpha_i + \beta_j$.
- Recall the primal problem : the cost of shipping one unit from x_i to y_j is $C_{i,j}$.
- Good deal for the driver implies $\alpha_i + \beta_j \leq C_{ij}$.
- the driver checks that the vendor's proposition is a better deal:

$$\begin{aligned}\sum_{i,j} \pi_{ij} C_{ij} \geq \sum_{i,j} \pi_{ij}(\alpha_j + \beta_j) &= \left(\sum_i \alpha_i \sum_j \pi_{ij} \right) + \left(\sum_j \beta_j \sum_i \pi_{ij} \right) \\ &= \langle \alpha, a \rangle + \langle \beta, b \rangle\end{aligned}$$

Example 5: Wasserstein GANs



$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z))]$$

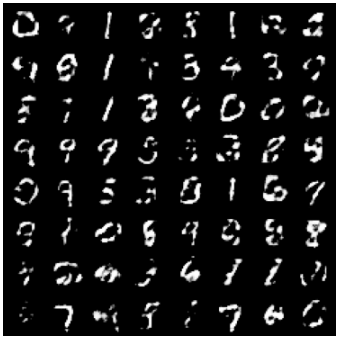
Notice the remarkable similarity between the objectives of the (dual) OT formulation and GANs

Example 5: Wasserstein GANs

GANs vs WGANs: Implementation details

- Discriminator loss no longer a likelihood fn
- Optimised with RMSProp
- Loss for D and G have the same form (Kantorovich potential, $p = 1$)
- Discriminator's inner loop training n_{critic} no longer equal to 1
- Learned parameters are clipped to ensure $||f||_L = 1$

GAN



WGAN

Notebooks: gan.ipynb & wgan.ipynb

Part II: The Wasserstein distance

Motivation

OT defines a family of distances between measures

The **Kantorovitch problem**

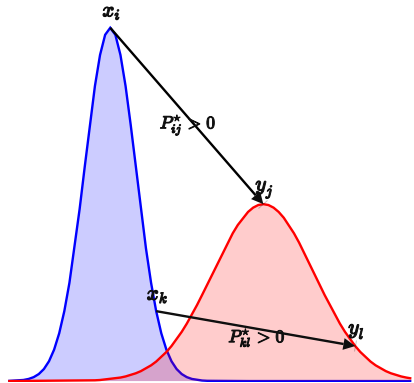
$$P^\star \in \inf_{P \in \Pi_{\mu,\nu}} \langle P, C \rangle = \sum_{i,j}^{n,m} C_{ij} P_{ij}$$

allows defining the **Wasserstein distance** of order p

$$W_p^p(\mu, \nu) = \langle P^\star, C \rangle$$

where the moving cost $c(x, y) = d(x, y)^p = \|x - y\|^p$.
It is often depicted as an "horizontal" distance

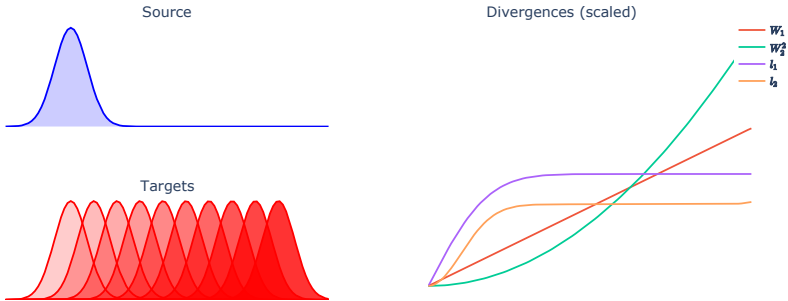
- ✓ **symmetry**
- ✓ **identity of indiscernibles**
- ✓ **triangular inequality**



Notebook: Horizontal distance.ipynb

The Wasserstein distance

- Does not need overlapping support (as KL)
- Determines the *degree of dissimilarity* between distributions



Notebook: Wasserstein_distance.ipynb

On the suitability of W_p for learning

- Thus far we have referred to *spaces of probability functions*, but we are interested in applying W_p on spaces of **generative models**.
- Learning is such a space requires, more than a distance, a notion of **convergence**
- Consider μ_{data} to be the true data distribution. We want to find a model $(P_\theta)_{\theta \in \Theta}$ such that $P_\theta \to \mu_{\text{data}}$, or equivalently, $D(\mu_{\text{data}}, P_\theta) \to 0$ — for a **reasonable** distance D.

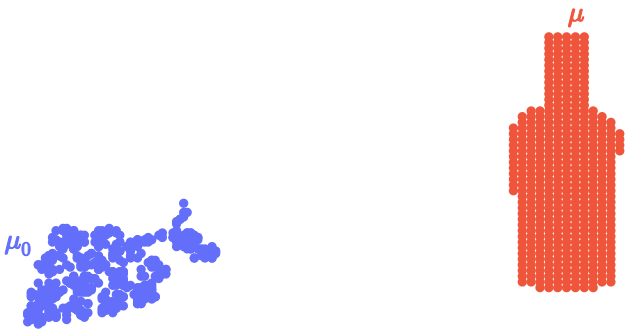
Discussion: Consider δ_{x_0} and $\delta_{x_i}, x_i \to x_0$

Example 6: Gradient flows on Wasserstein space

Wasserstein space $\mathbb{W}_p$: space endowed with the distance W_p

- In the space $\mathbb{W}_p(\mathbb{R}^d)$, we have $W_p(\mu_n, \mu) \to 0$ iff $\mu_n \to \mu$ (weak topology)

Consider the loss $W_2^2(\mu_t, \mu)$. The figure below shows how a distribution μ_0 evolves under de application of gradient flow of this loss.



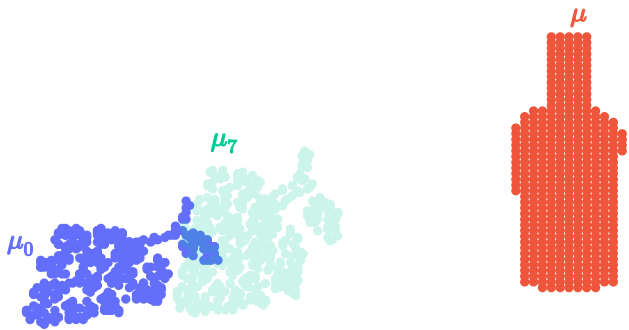
Notebook: Wasserstein Gradient Flows.ipynb

Example 6: Gradient flows on Wasserstein space

Wasserstein space $\mathbb{W}_p$: space endowed with the distance W_p

- In the space $\mathbb{W}_p(\mathbb{R}^d)$, we have $W_p(\mu_n, \mu) \to 0$ iff $\mu_n \to \mu$ (weak topology)

Consider the loss $W_2^2(\mu_t, \mu)$. The figure below shows how a distribution μ_0 evolves under de application of gradient flow of this loss.



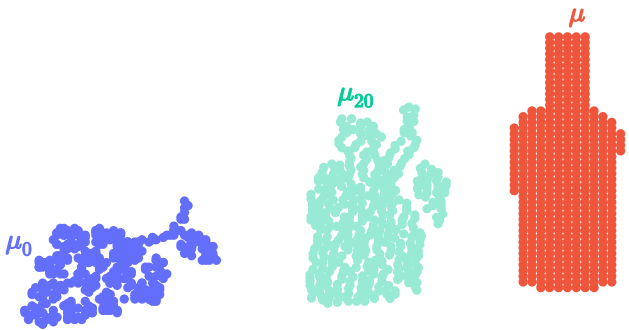
Notebook: Wasserstein Gradient Flows.ipynb

Example 6: Gradient flows on Wasserstein space

Wasserstein space $\mathbb{W}_p$: space endowed with the distance W_p

- In the space $\mathbb{W}_p(\mathbb{R}^d)$, we have $W_p(\mu_n, \mu) \to 0$ iff $\mu_n \to \mu$ (weak topology)

Consider the loss $W_2^2(\mu_t, \mu)$. The figure below shows how a distribution μ_0 evolves under de application of gradient flow of this loss.



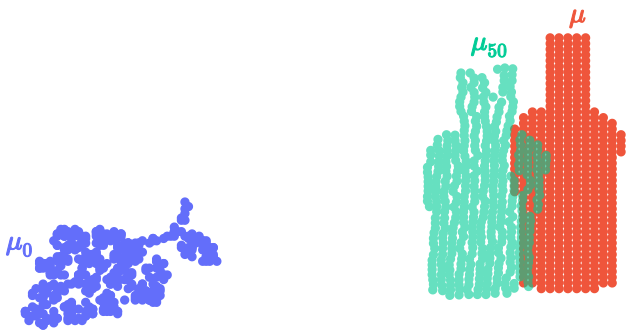
Notebook: Wasserstein Gradient Flows.ipynb

Example 6: Gradient flows on Wasserstein space

Wasserstein space $\mathbb{W}_p$: space endowed with the distance W_p

- In the space $\mathbb{W}_p(\mathbb{R}^d)$, we have $W_p(\mu_n, \mu) \to 0$ iff $\mu_n \to \mu$ (weak topology)

Consider the loss $W_2^2(\mu_t, \mu)$. The figure below shows how a distribution μ_0 evolves under de application of gradient flow of this loss.



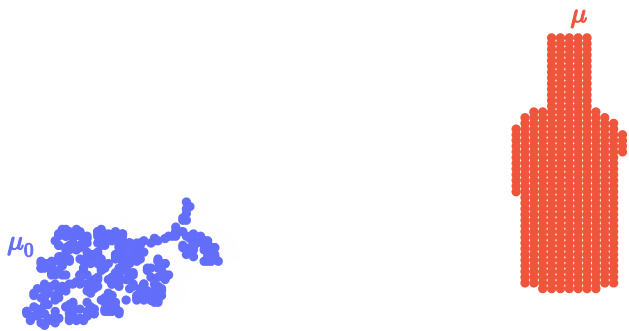
Notebook: Wasserstein Gradient Flows.ipynb

Example 6: Gradient flows on Wasserstein space

Wasserstein space $\mathbb{W}_p$: space endowed with the distance W_p

- In the space $\mathbb{W}_p(\mathbb{R}^d)$, we have $W_p(\mu_n, \mu) \to 0$ iff $\mu_n \to \mu$ (weak topology)

Consider the loss $W_2^2(\mu_t, \mu)$. The figure below shows how a distribution μ_0 evolves under de application of gradient flow of this loss.



Notebook: Wasserstein Gradient Flows.ipynb

Geodesic paths between distributions

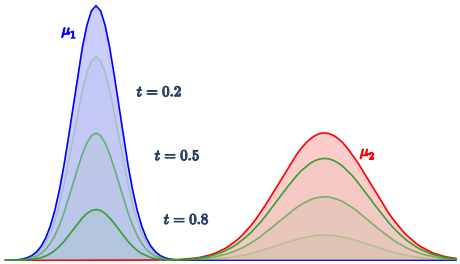
A geodesic generalizes the concept of a straight line between two points

It is a curve that represents the shortest path between two manifolds

Euclidean space with a l_2 distance is a **geodesic space**

$$\forall t \in [0,1], \quad \mu^{1\to 2}(t) = t\mu_2 + (1-t)\mu_1$$

Allows "vertical" interpolation between the distributions



Notebook: Wasserstein Geodesics.ipynb

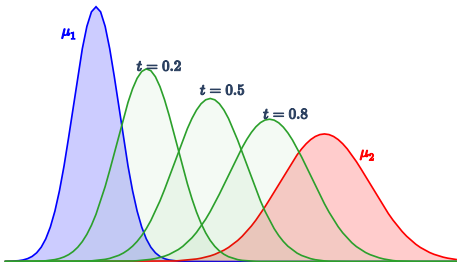
Geodesic properties of the Wasserstein space

$\mathbb{W}_p$ is a **geodesic space**

- Given a Monge map T between μ_1 and μ_2 such that $T_{\#}\mu_1 = \mu_2$, a geodesic curve $\mu^{1\to 2}$ is

$$\forall t \in [0,1], \quad \mu^{1\to 2}(t) = (tT + (1-t)\mathrm{Id})_{\#}\mu_1$$

- It represents the shortest path (on the Wasserstein space $\mathbb{W}_p$) between μ_1 and μ_2
- Allows "horizontal" interpolation between the distributions



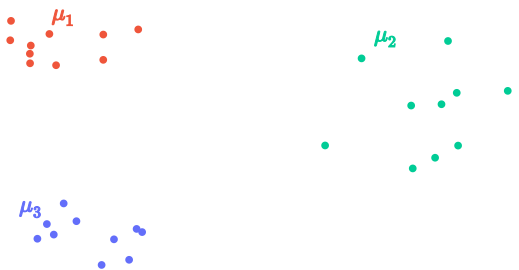
Notebook: Wasserstein Geodesics.ipynb

The Wasserstein barycenter

Given a set of distributions μ_s, compute a Wasserstein Frechet mean

$$\overline{\mu} = \arg\min_{\mu} \sum_{i=1}^{s} \lambda_i W_p^p(\mu, \mu_i)$$

where $\lambda_i > 0$ and $\sum_{i=1}^{s} \lambda_i = 1$.
Generalizes the interpolation between more than 2 measures.
For discrete measures $\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \Rightarrow$ we can fix the weights a_i and/or the support x_i.

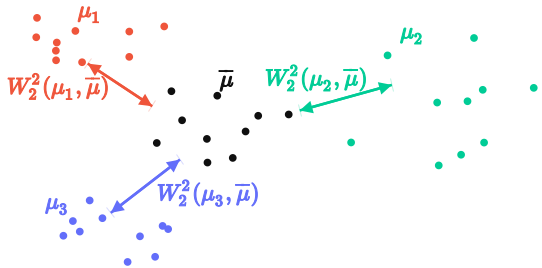


The Wasserstein barycenter

Given a set of distributions μ_s, compute a Wasserstein Frechet mean

$$\overline{\mu} = \arg\min_{\mu} \sum_{i=1}^{s} \lambda_i W_p^p(\mu, \mu_i)$$

where $\lambda_i > 0$ and $\sum_{i=1}^{s} \lambda_i = 1$.
Generalizes the interpolation between more than 2 measures.
For discrete measures $\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \Rightarrow$ we can fix the weights a_i and/or the support x_i.



The Wasserstein barycenter

Example on averaging over images

Figure 1: In the **Euclidean** space

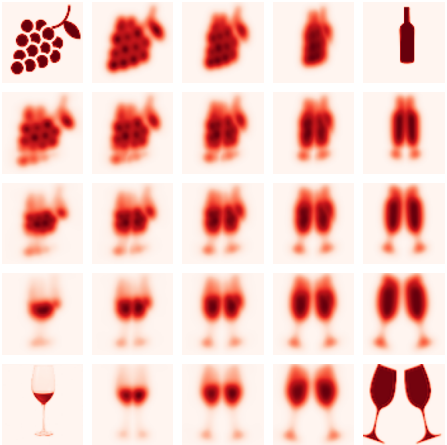


Figure 2: In the **Wasserstein** space

Notebook: Wass bary 4 distribs.ipynb

Examples

See which ones make sense here: OT spectral transport, Wasserstein Bays, VAEs.

What we did not see

Computational OT
multimarginal, unbalanced OT, partial OT, Gromov-Wasserstein,

Conclusions & the future

- OT is now on the toolkit for many fields such as signal processing, machine learning etc.
- Defines a meaningful distances between distribution, with the extra information on how the particles should be moved
- Some open challenges: computational complexity, curse of dimensionality (number of samples to approximate the solutions depends is exponential with the dimension), robustify the solution with statistical guarantees (noise? outliers?), OT on different spaces than Euclidean ones, adding some extra constraints (like a temporal consistency)

Optimal Transport for Signal Processing

A tutorial at MLSP 2024

Felipe Tobar[1] Laetitia Chapel[2]

[1]Initiative for Data & Artificial Intelligence, Universidad de Chile

[2]IRISA, Obelix team, Institut Agro Rennes-Angers

22 September, 2024