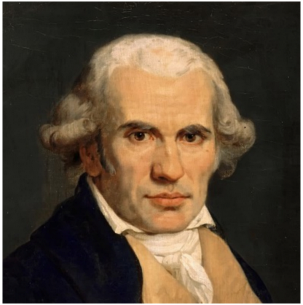
A few words about Optimal Transport (OT) and its generalizations

Xavier Bacon

Exposé à SAFRAN Paris-Saclay

7/12/2023

1. Static OT
   - Formulation and examples
   - Numerical aspects
2. Generalizations and applications
   - Multi-species OT
   - A spatial exchange economy problem
   - Dynamical, unbalanced and matrix-valued OT



(a) G. Monge

(b) L. Kantorovich

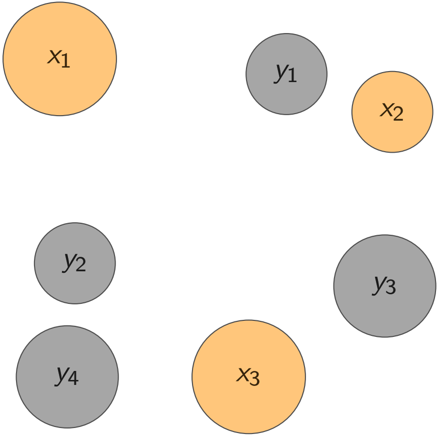
Table of Contents

What is Optimal Transport? 1/2

How can we describe the transportation of mass ?

What is Optimal Transport? 1/2

How can we describe the transportation of mass ?

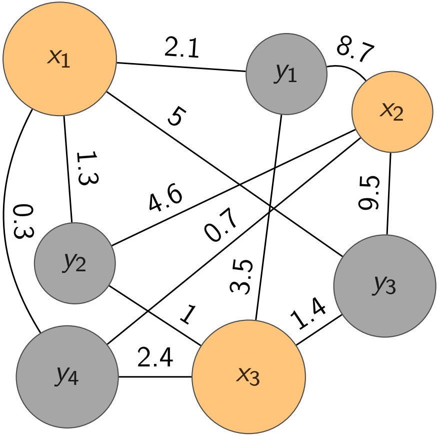


$X = \{x_1, x_2, x_3\}, \ Y = \{y_1, y_2, y_3, y_4\}$

μ and ν **two weighted point clouds**

What is Optimal Transport? 1/2

How can we describe the transportation of mass ?



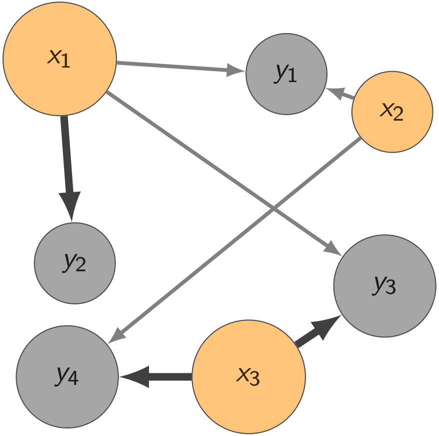
$X = \{x_1, x_2, x_3\}, \ Y = \{y_1, y_2, y_3, y_4\}$

μ and ν **two weighted point clouds**

$c = c(x, y)$ **transport cost**

What is Optimal Transport? 1/2

How can we describe the transportation of mass ?



$X = \{x_1, x_2, x_3\},\ Y = \{y_1, y_2, y_3, y_4\}$

μ and ν **two weighted point clouds**

$c = c(x, y)$ **transport cost**

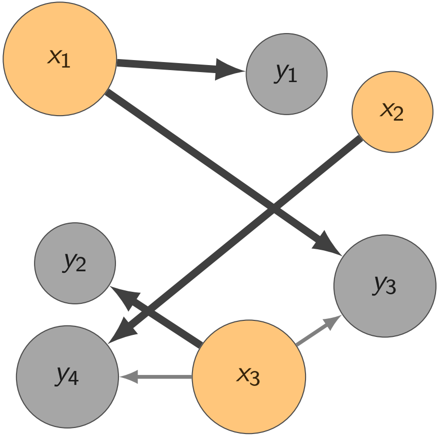
$\gamma =$ **transference plan**

0.1	0.2	0.1	0
0.1	0	0	0.1
0	0	0.2	0.2

$\mathbf{Cost}_c(\gamma) = \langle c | \gamma \rangle$

What is Optimal Transport? 1/2

How can we describe the transportation of mass ?



$X = \{x_1, x_2, x_3\}, \; Y = \{y_1, y_2, y_3, y_4\}$

μ and ν **two weighted point clouds**

$c = c(x,y)$ **transport cost**

$\gamma =$

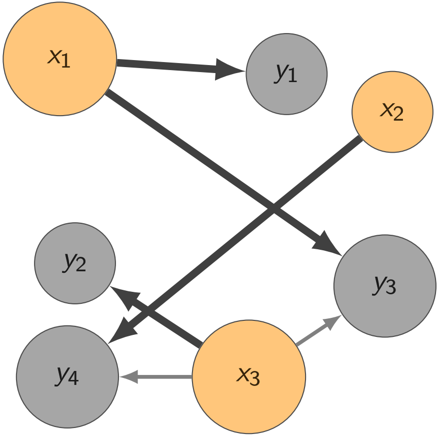
0.2	0	0.2	0
0	0	0	0.2
0	0.2	0.1	0.1

transference plan

$\mathbf{Cost}_c(\gamma) = \langle c | \gamma \rangle$

What is Optimal Transport? 1/2

Among all transference, which is the best?



$X = \{x_1, x_2, x_3\}, \ Y = \{y_1, y_2, y_3, y_4\}$

μ and ν **two weighted point clouds**

$c = c(x,y)$ **transport cost**

$$\gamma = \begin{pmatrix} 0.2 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 \\ 0 & 0.2 & 0.1 & 0.1 \end{pmatrix}$$ **transference plan**

$\mathbf{Cost}_c(\gamma) = \langle c \mid \gamma \rangle$

What is Optimal Transport? 2/2

- $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$ are respectively the source and the target distributions.
- $c = c(x,y)$ is the **transport cost**.

The (Static) Optimal Transport problem is defined as follows

$$\mathrm{OT}_c(\mu,\nu) = \min\left\{\int_{X\times Y} c(x,y)\,d\gamma(x,y) \text{ s.t. } \gamma \in \Pi(\mu,\nu)\right\},$$

where $\Pi(\mu,\nu)$ denotes the set of all $\gamma \in \mathscr{P}(X\times Y)$ having μ and ν as marginals, called **transference plans**.

(Discrete setting) $\mu = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$, $\nu = \sum_{j=1}^{m} \beta_j \delta_{y_j}$, with $\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 1$, $\alpha_i, \beta_j \geqslant 0$,

$$\mathrm{OT}_c(\mu,\nu) = \mathrm{OT}_c(\alpha,\beta) = \min\left\{\langle c|\gamma\rangle \text{ s.t. } \gamma\cdot\mathbb{1} = a, \gamma^T\cdot\mathbb{1} = b\right\}.$$

Wasserstein distance

p-Wasserstein distance

If (X, d) is a metric space and $c = d^p$, then

$$\mathrm{OT}_c \text{ is a distance on } \mathscr{P}(X) \text{ if } 0 < p \leqslant 1,$$
$$\mathrm{OT}_c^{1/p} \text{ is a distance on } \mathscr{P}(X) \text{ if } p > 1.$$

- OT_c metrizes the weak-* convergence of probabilities (convergence in Law).
- If $p = 2$, OT_2 is a geodesical distance: any pair of probabilities μ and ν can be connected by a continuous path of length $\mathrm{OT}_2(\mu, \nu)$.

Solving OT problem

HOW DO WE SOLVE OT_c ?

Solving OT problem

HOW DO WE SOLVE OT_c ?

💡 It depends on the cost c.

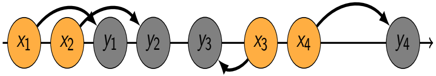
Solving OT problem

HOW DO WE SOLVE OT_c ?

- 💡 It depends on the cost c.
- ⚠ Computing $\mathrm{OT}_c(\mu, \nu)$ means solving an (linear) optimization problem, except for few cases...

1 dimensional OT

- $X = Y = \mathbb{R}$, $c(x,y) = |x-y|^p$, $p \geqslant 1$ (convex case).



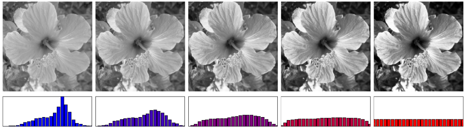


Figure: Histogram equalization

1 dimensional Wasserstein

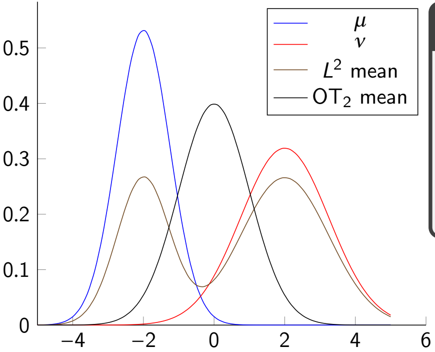
Denoting F^{-1} the (generalized) quantile function

$$W_p(\mu,\nu)^p = \left\| F_\mu^{-1} - F_\nu^{-1} \right\|_{L^p},$$

that is $(\mathbb{R}, W_p)$ is isometric to $L^p(\mathbb{R})$ through the map $\mu \mapsto F_\mu^{-1}$.

Computational aspect : $O(n \log n + m \log m)$

Gaussian OT



Gaussian quadratic OT

$X = Y = \mathbb{R}^d$, $c(x,y) = \frac{1}{2}|x-y|^2$,
$\mu \sim \mathcal{N}(m_\mu, \Sigma_\mu)$, $\nu \sim \mathcal{N}(m_\nu, \Sigma_\nu)$.

$$\mathrm{OT}_2(\mu,\nu)^2 = |m_\mu - m_\nu|^2 + \mathrm{Bures}(\Sigma_\mu, \Sigma_\nu)^2$$

- OT_2 between Gaussian densities is the L^2 distance between parameters.

$$\mathrm{Bures}(\Sigma_\mu,\Sigma_\nu)^2 = \mathrm{trace}\left[\Sigma_\mu + \Sigma_\nu 2\left(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2}\right)^{1/2}\right].$$

Table of Contents

Computational aspects : motivations

- $X = \{x_1, \cdots, x_n\}$, $Y = \{y_1, \cdots, y_m\}$ and the cost $c = c(x, y)$ is arbitrary.
- $\mu = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$, $\nu = \sum_{j=1}^{m} \beta_j \delta_{y_j}$ (weighted point clouds).

As a linear problem, OT_c can be solved using Hungarian algorithm but...

Computing OT_c using Hungarian algorithm

$$\triangle\!\!\!! \; O((n+m)nm\log(n+m)) \; \triangle\!\!\!!$$

Computational aspects : motivations

- $X = \{x_1, \cdots, x_n\}$, $Y = \{y_1, \cdots, y_m\}$ and the cost $c = c(x, y)$ is arbitrary.
- $\mu = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$, $\nu = \sum_{j=1}^{m} \beta_j \delta_{y_j}$ (weighted point clouds).

As a linear problem, OT_c can be solved using Hungarian algorithm but...

Computing OT_c using Hungarian algorithm

$$\triangle O((n+m)nm\log(n+m)) \triangle$$

WE MUST REGULARIZE

Regularized OT

X, Y finite and $\alpha \in \mathbb{R}^X, \beta \in \mathbb{R}^Y$ denotes the source and the target distributions. We introduce the discrete entropy of γ as

$$Entropy(\gamma) = -\sum_{i,j} \gamma_{ij} [\log(\gamma_{ij}) - 1]$$

Entropic OT (Cuturi 2013)

For a fixed ε 0, the regularized OT problem is given by

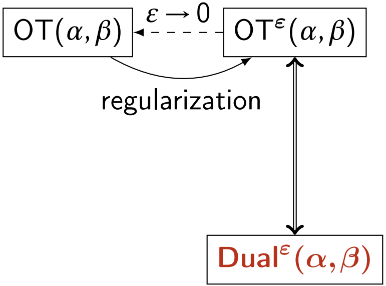
$$\mathrm{OT}^{\varepsilon}(\alpha, \beta) = \min \left\{ \langle c | \gamma \rangle - \varepsilon \, \boldsymbol{Entropy}(\gamma) \;\; s.t. \;\; \gamma \in \Pi(\alpha, \beta) \right\},$$

whose unique solution converges to the solution of $\mathscr{T}(\alpha, \beta)$ with maximal entropy.

- The original linear optimization problem ($\varepsilon = 0$) is replaced by a strongly convex problem.

Strategy

In instead of solving $\mathrm{OT}^{\varepsilon}$, we solve the **dual problem of $\mathrm{OT}^{\varepsilon}$**. Once solved, we construct the solution to our initial problem, using the optimal *primal-dual* conditions:



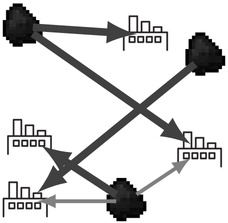
$$\gamma^{\varepsilon} = \mathrm{diag}\left[\exp(\varphi^{\varepsilon}/\varepsilon)\right] \cdot K^{\varepsilon} \cdot \mathrm{diag}\left[\exp(\psi^{\varepsilon}/\varepsilon)\right],$$

where

$$\begin{aligned} \gamma^{\varepsilon} &= \operatorname{argmin} \mathrm{OT}^{\varepsilon}(\alpha,\beta) \\ (\varphi^{\varepsilon},\psi^{\varepsilon}) &= \operatorname{argmax} \mathbf{Dual}^{\varepsilon}(\alpha,\beta) \\ K^{\varepsilon} &= \exp(-c/\varepsilon) \ \text{(heat kernel)} \end{aligned}$$

A short digression about duality 1/2

An external provider (say Alice) offers to transport the coal with her own trucks according to the following contract.



Alice chooses the price $\varphi(x)$ for loading in x.
Alice chooses the price $\psi(y)$ for unloading in y.
Alice assures me that it will cost less than doing it myself

$$\varphi(x) + \psi(y) \leqslant c(x,y).$$

Alice charges me $\langle \varphi | \mu \rangle + \langle \psi | \nu \rangle$.

$$\mathrm{Dual}(\mu,\nu) = \max\{\langle \varphi | \mu \rangle + \langle \psi | \nu \rangle : \varphi(x) + \psi(y) \leqslant c(x,y)\}. \qquad \text{(Alice problem)}$$

A short digression about duality 2/2

$$\mathrm{Dual}(\mu,\nu) = \max\{\langle\varphi|\mu\rangle + \langle\psi|\nu\rangle : \varphi(x)+\psi(y) \leqslant c(x,y)\}$$

What is the link between $\mathrm{Dual}(\mu,\nu)$ and $\mathrm{OT}(\mu,\nu)$?

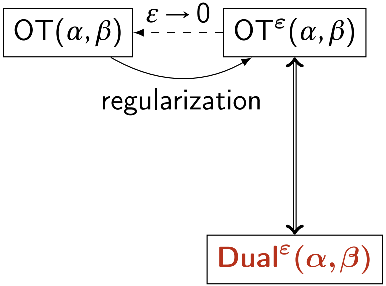
Theorem (strong duality)

$\mathrm{Dual}(\mu,\nu) = \mathrm{OT}(\mu,\nu)$ and any solution of $\mathrm{Dual}(\mu,\nu)$ leads to a solution of $\mathrm{OT}(\mu,\nu)$.

$\mathrm{OT}(\mu,\nu)$ is a optimization problem on $\mathbb{R}^{n\times m}$ while $\mathrm{Dual}(\mu,\nu)$ is on $\mathbb{R}^{n+m}$.

Strategy

In instead of solving $\mathrm{OT}^{\varepsilon}$, we solve the **dual problem of $\mathrm{OT}^{\varepsilon}$**. Once solved, we construct the solution to our initial problem, using the optimal *primal-dual* conditions:



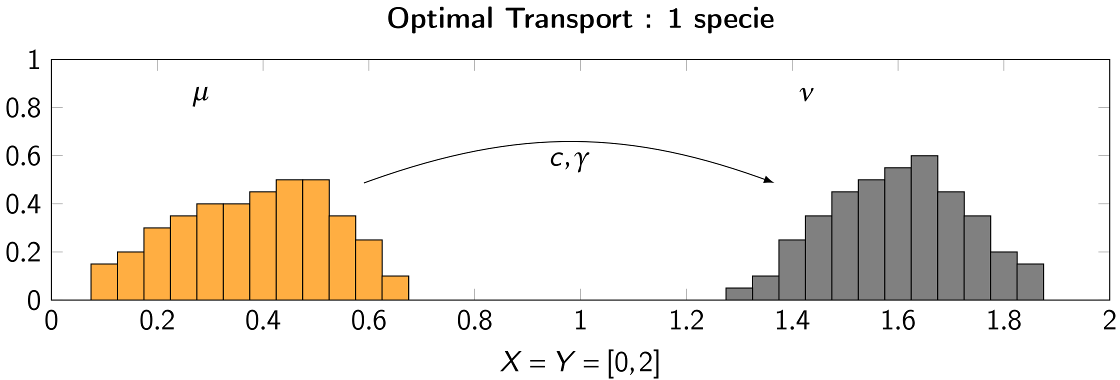
Solving $\mathrm{Dual}^{\varepsilon}(\alpha,\beta)$ using Sinkhorn's algorithm (Cuturi, 2013) : setting $u^{\varepsilon} = \exp(\varphi^{\varepsilon}/\varepsilon)$ and $v^{\varepsilon} = \exp(\psi^{\varepsilon}/\varepsilon)$, starting with $u^0 = \mathbb{1}_X, v^0 = \mathbb{1}_Y$, we compute alternatively

$$u^{k+1} = \frac{\alpha}{K \cdot v^k} \text{ and } v^{k+1} = \frac{\beta}{K^T \cdot u^{k+1}}$$

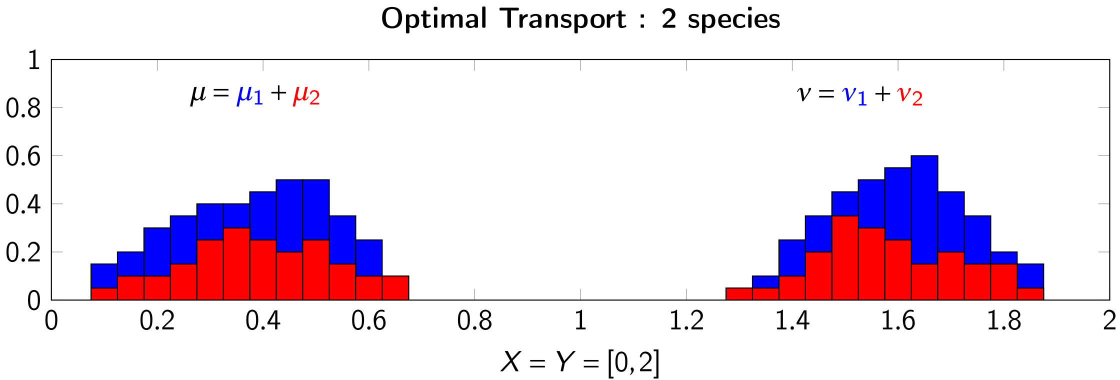
whose convergence is linear.

Table of Contents

From scalar OT to multi-species OT

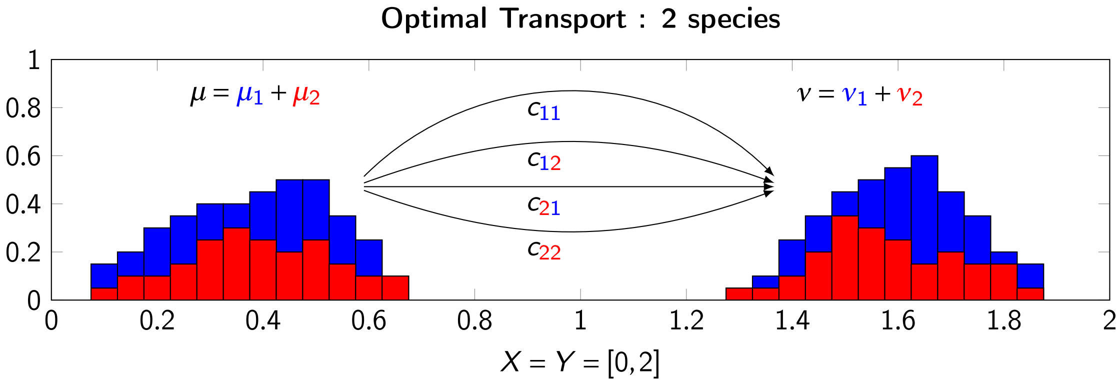


From scalar OT to multi-species OT



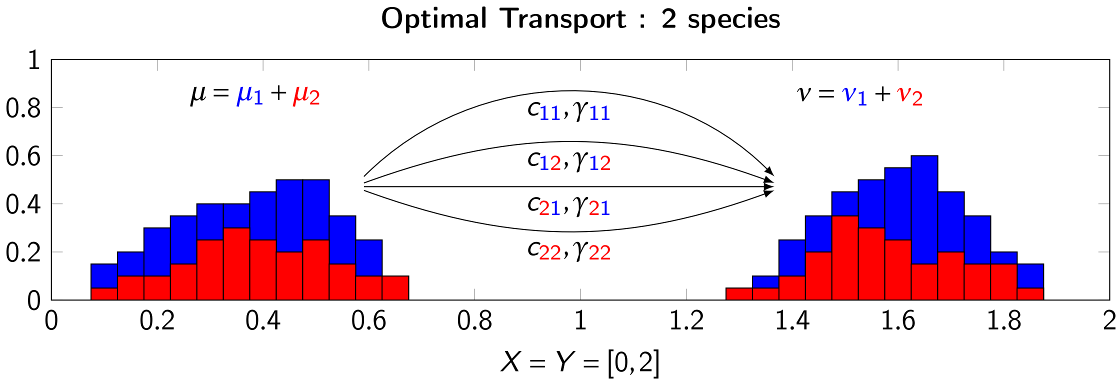
- $\mu = \mu_1 + \mu_2 \in \mathscr{P}(X)$, $\nu = \nu_1 + \nu_2 \in \mathscr{P}(Y)$.

From scalar OT to multi-species OT



- $\mu = \mu_1 + \mu_2 \in \mathscr{P}(X)$, $\nu = \nu_1 + \nu_2 \in \mathscr{P}(Y)$.
- Four costs c_{11}, c_{12}, c_{21}, c_{22}.

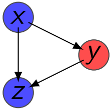
From scalar OT to multi-species OT



- $\mu = \mu_1 + \mu_2 \in \mathscr{P}(X)$, $\nu = \nu_1 + \nu_2 \in \mathscr{P}(Y)$.
- Four costs c_{11}, c_{12}, c_{21}, c_{22}.
- ...And as much transference plans γ_{11}, γ_{12}, γ_{21}, γ_{22}.

Distance on $\mathbb{R}_+^n$ measures

- **Classical OT:** if (X,d) is a Polish space, for $p \in [1,\infty)$ and $c = d^p$, $\mathscr{T}^{1/p}$ is a distance on $\mathscr{P}_p(X)$.



- **Multi-species OT:** setting $c_{ij} = d_{ij}^p$, we make the following assumption

$$d_{ik}(x,z) \leqslant d_{ij}(x,y) + d_{jk}(y,z) \quad \textit{\textbf{(Mixed Triangles Inequalities)}}$$

Theorem (B. 2020)

Given n^2 functions (d_{ij}) defined on $X \times X$ and $\mathbb{R}_+$-valued such that
[1] $\forall (i,j) \in [\![1,n]\!]^2, d_{ij}$ is symmetric.
[2] (MTI) is satisfied for all $(i,j,k) \in [\![1,n]\!]^3$ et $(x,y,z) \in X^3$.
[3] $\forall i \in [\![1,n]\!], \forall x \in X, d_{ii}(x,x) = 0$.
[4] $\forall (i,j) \in [\![1,n]\!]^2, i \neq j, \forall (x,y) \in X \times X, d_{ij}(x,y) \neq 0$.
Then MultiOT is a distance between multi-valued probabilities.

Table of Contents

Primal problem

- X is a compact metric space.

Primal problem

$$(\mathscr{P}) = \max \left\{ \qquad \right\}.$$

Primal problem

- X is a compact metric space.
- $\mu = (\mu_1, \ldots, \mu_N) \in \mathscr{M}_+(X)^N$ are the source distributions of goods in region X.

Primal problem

$$(\mathscr{P}) = \max \left\{ \qquad \qquad \qquad \qquad \right\}.$$

Primal problem

- X is a compact metric space.
- $\mu = (\mu_1, \ldots, \mu_N) \in \mathscr{M}_+(X)^N$ are the source distributions of goods in region X.
- $\nu = (\nu_1, \ldots, \nu_N) \in \mathscr{M}_+(X)^N$ are the target distributions of goods in region X.

Primal problem

$$(\mathscr{P}) = \max_{\boldsymbol{\nu}} \left\{ \qquad \text{s.t. } \mu_i(X) = \boldsymbol{\nu_i}(X) \right\}.$$

Primal problem

- X is a compact metric space.
- $\mu = (\mu_1, \ldots, \mu_N) \in \mathscr{M}_+(X)^N$ are the source distributions of goods in region X.
- $\nu = (\nu_1, \ldots, \nu_N) \in \mathscr{M}_+(X)^N$ are the target distributions of goods in region X.
- $\mathscr{U} = \mathscr{U}(\boldsymbol{\nu})$ is the average utility.

Primal problem

$$(\mathscr{P}) = \max_{\boldsymbol{\nu}} \left\{ \mathscr{U}(\boldsymbol{\nu}) \qquad \text{s.t. } \mu_i(X) = \boldsymbol{\nu_i}(X) \right\}.$$

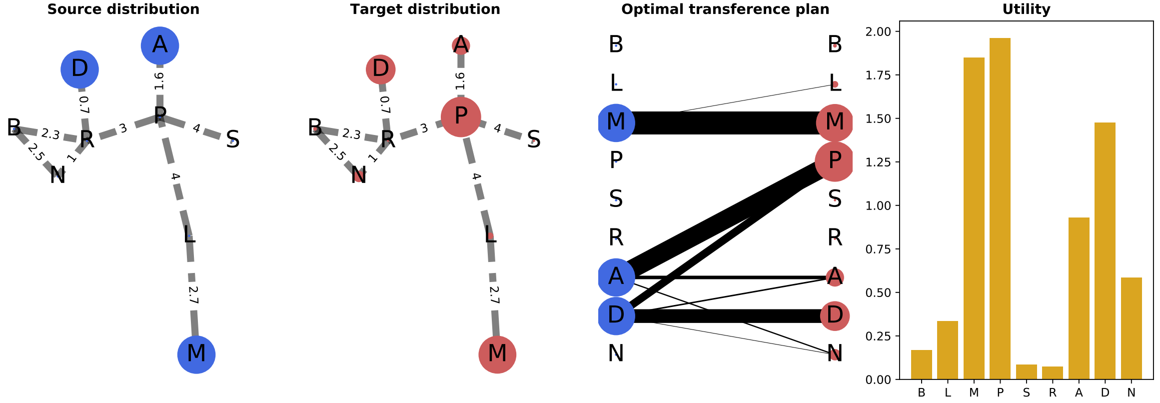
Primal problem

- X is a compact metric space.
- $\mu = (\mu_1, \ldots, \mu_N) \in \mathcal{M}_+(X)^N$ are the source distributions of goods in region X.
- $\nu = (\nu_1, \ldots, \nu_N) \in \mathcal{M}_+(X)^N$ are the target distributions of goods in region X.
- $\mathscr{U} = \mathscr{U}(\boldsymbol{\nu})$ is the average utility.
- $\mathscr{T}(\mu, \boldsymbol{\nu}) = \sum_{i=1}^{N} \mathrm{OT}_{c_i}(\mu_i, \boldsymbol{\nu_i})$ is the transport cost between μ and $\boldsymbol{\nu}$.

Primal problem

$$(\mathscr{P}) = \max_{\boldsymbol{\nu}} \left\{ \mathscr{U}(\boldsymbol{\nu}) - \mathscr{T}(\mu, \boldsymbol{\nu}) \text{ s.t. } \mu_i(X) = \boldsymbol{\nu_i}(X) \right\}.$$

An example


Source distribution
D
A
0.7
1.6
B
2.3
R
3
P
4
S
2.5
1
N
4
L
2.7
M
Target distribution
A
D
0.7
1.6
P
B
2.3
R
3
4
S
2.5
1
N
4
L
2.7
M
Optimal transference plan
B
L
M
P
S
R
A
D
N
B
L
M
P
S
R
A
D
N
Utility
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
B
L
M
P
S
R
A
D
N

Economic interpretation

Let (β, φ) be optimal in $(\mathscr{P})$ and $(\mathscr{D})$. Then we have an **equilibrium** for the initial monetary endowment $\boldsymbol{w}\ \langle \boldsymbol{\varphi^c} \,|\, \boldsymbol{\alpha} \rangle\ \langle \boldsymbol{\varphi} \,|\, \boldsymbol{\beta} \rangle (= \mathscr{T}_c(\alpha, \beta))$ in the sense that:

Sellers at x maximize their profits by exporting their goods $\alpha(x)$:

$$\text{profits}_i(x) = \max_y \ \varphi_i(y) - c_i(x,y) \ \ (= -\varphi_i^{c_i}(x))$$
$$\text{total profits}(x) = \langle \text{profits}(x) \,|\, \alpha(x) \rangle$$

Consumers in y have an initial endowment $w(y)$ and buy $\beta_i(y)$ in such a way:

$$\beta_i(y) = \operatorname*{argmax}_\beta \left\{ \boldsymbol{U}(\boldsymbol{y}, \beta) \text{ s.t. } \langle \varphi \,|\, \beta \rangle \leqslant \underbrace{\overbrace{\langle -\varphi^c(y) \,|\, \alpha(y) \rangle}^{\text{export profit}} + w(y)}_{\text{total revenue}} \right\}$$

Markets are clear : it exists an optimal transport plan for all target endowments.

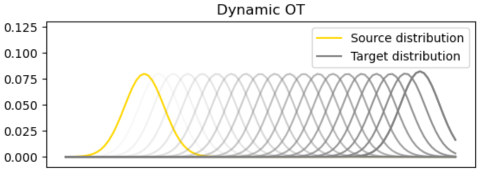
Table of Contents

Dynamical formulation of OT 1/2

- $X = Y =$ a domain of $\mathbb{R}^d$, $(\mu,\nu) \in \mathscr{P}(X)^2$ and $c(x,y) = \frac{1}{2}|x-y|^2$.
- We introduce the **time variable** t and consider all the kinematics $\rho = \rho_t$ which linked μ (initial time, $t=0$) to ν (final time, $t=1$).

Dynamical formulation of OT 1/2

- $X = Y =$ a domain of $\mathbb{R}^d$, $(\mu, \nu) \in \mathscr{P}(X)^2$ and $c(x,y) = \frac{1}{2}|x-y|^2$.
- We introduce the **time variable** t and consider all the kinematics $\rho = \rho_t$ which linked μ (initial time, $t = 0$) to ν (final time, $t = 1$).

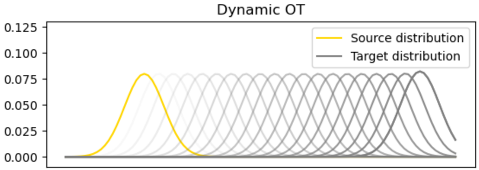


Constraints = **conservation of mass**

Action(ρ, v) $= \frac{1}{2}\int_0^1 \int_X |v_t(x)|^2 \rho_t(x)\, dx\, dt$

Dynamical formulation of OT 1/2

- $X = Y =$ a domain of $\mathbb{R}^d$, $(\mu,\nu) \in \mathscr{P}(X)^2$ and $c(x,y) = \frac{1}{2}|x-y|^2$.
- We introduce the **time variable** t and consider all the kinematics $\rho = \rho_t$ which linked μ (initial time, $t=0$) to ν (final time, $t=1$).



Constraints = **conservation of mass**

$$\textbf{Action}(\rho,v) = \frac{1}{2}\int_0^1 \int_X |v_t(x)|^2 \rho_t(x)\,dx\,dt$$

$$\text{DynamicOT}(\mu,\nu) = \min_{\rho:[0,1]\to P(X)} \left\{ Action(\rho,v) \text{ s.t. } \partial_t \rho + \text{div}_x\, \rho v = 0 \text{ and } \rho_0 = \mu, \rho_1 = \nu \right\}$$

Dynamic formulation of OT 2/2

Is there a link between DynamicOT(μ,ν) and OT$_2(\mu,\nu)$?

Theorem

If X is convex, the static and dynamic formulations are equivalent.

- If γ is an optimal transference plan, then $\rho_t = \pi_t \# \gamma$ is an optimal path, where $\pi_t(x,y) = (1-t)x + ty$.
- As a consequence, the set of all probabilities endowed with the 2-Wasserstein distance is a geodesic space.

From dynamical OT to unbalanced OT

- μ and ν **two positive finites measures** : they do not share the same total mass.

From dynamical OT to unbalanced OT

- μ and ν **two positive finites measures** : they do not share the same total mass.

Constraints = ~~conservation of mass~~ **destruction/creation of mass**

$$\textbf{Action}(\rho, v, r) = \frac{1}{2}\int_0^1 \int_X \left(|v_t(x)|^2 + |r_t(x)|^2\right) \rho_t(x)\, dx\, dt$$

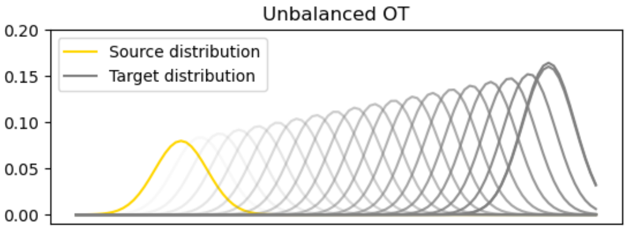
From dynamical OT to unbalanced OT

- μ and ν **two positive finites measures** : they do not share the same total mass.

Constraints = ~~conservation of mass~~ destruction/creation of mass

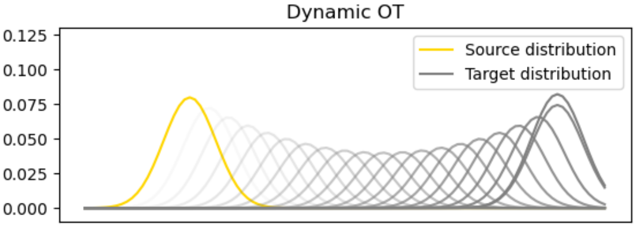
$$\text{Action}(\rho, v, r) = \frac{1}{2}\int_0^1 \int_X \left(|v_t(x)|^2 + |r_t(x)|^2\right)\rho_t(x)\,dx\,dt$$

$$\text{UOT}(\mu,\nu) = \min_{\rho:[0,1]\to M_+(X)} \left\{ Action(\rho, v, r) \text{ s.t. } \partial_t \rho + \text{div}_x\, \rho v = \rho r, \rho_0 = \mu, \rho_1 = \nu \right\}$$



Further comments

- Unbalanced OT is a distance between measures which do not share the same mass. In particular between any two (integrable) functions.
- Notice that even between two probabilities, Unbalanced OT and Static OT may be different.



To go further...matrix-valued OT

Denoting by $\mathbb{S}_+$ the set of $n \times n$ positive semi-definite matrices, previous constraints and action can be replaced by PSD-valued measures (Li & Zou, 2020) :

- S_0, S_1 two PSD-valued measures on a domain of $\mathbb{R}^d$.

Minimizing among all continuous path $\mathscr{S} : [0,1] \mapsto \mathbb{S}_+$ the previous (matricial) action subject to the previous (matricial) constraints, leads us to define a distance between **matrix-valued distributions**:

$$\mathrm{MUOT}(S_0, S_1) = \min_{\mathscr{S}} \{ \mathit{Action}(\mathscr{S}, \mathscr{V}, \mathscr{R}) \text{ s.t. } \text{Creation/Destruction Equation} \}$$

💡 In particular, we obtain a distance between covariance distributions.

Bibliography

OT monographs :

G. Peyré, M. Cuturi, **Computational Optimal Transport**.

C. Villani, **Optimal transport, old and new**.

F. Santambrogio, **Optimal Transport for Applied Mathematicians.**

Bibliography of the author :

B., G. Carlier, B. Nazaret, Applied Mathematics and Optimization Journal, 2023.

B., . Multi-species optimal transportation. Journal of Optimization Theory and Applications, 2020.

Python Libraries :

POT: Python Optimal Transport.

OTT-JAX.

The end

Thank you for your attention.