Lesson 3: Numerical aspects of optimal transport

I - Reminders / Crash course

1/ ⚠ * This section is intended as a quick reminder of some elements from the class of Max Fathi.

* In order to simplify exposition, we discard all growth / integrability hypothesis.

Let $\Omega, \Omega' \subset \mathbb{R}^d$ be measurable subsets

$$\mathcal{M}_1(\Omega) := \{\text{Probability measures on } \Omega\}$$
$$= \{\mathcal{L}(X) \mid X \text{ is an } \Omega\text{-valued r.v}\}$$

Given $(\mu, \nu) \in \mathcal{M}_1(\Omega) \times \mathcal{M}_1(\Omega')$, then set

- $\Pi(\mu,\nu) := \{\pi \in \mathcal{M}_1(\Omega \times \Omega') \mid (p_1)_* \pi = \mu;\ (p_2)_* \pi = \nu\}$

$$\simeq \left\{ \mathcal{L}(X,Y) \;\middle|\; \begin{array}{l} (X,Y) \text{ r.v on } \Omega \times \Omega' \\ \mathcal{L}(X) = \mu \\ \mathcal{L}(Y) = \nu \end{array} \right\}$$

"Transport plans" (TPE) or "Couplings" (P)

- $c : \Omega \times \Omega' \longrightarrow \mathbb{R}$ a cost function

Optimal transport (stated in the form of the Kantorovich problem) is the study of the optimization problem

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega'} c(x,y)\, \pi(dx, dy)$$

Thm [Brenier]

If $\begin{cases} \mu \text{ has a density } \mu(dx) = dx\, \rho(x) \\ \text{OR} \\ \mu \text{ \& } \nu \text{ have } n \text{ atoms, each of size } 1/n \end{cases}$

Then there exists a unique transport plan π of the form

$$\pi = (id, T)_{\#} \mu \in \Pi(\mu, \nu)$$

where $T : \Omega \to \Omega'$ is called the Brenier map

As such, in these favorable cases:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega'} c(x,y)\, \pi(dx,dy)$$

$$= \inf_{\substack{T : \Omega \to \Omega' \\ T_{\#}\mu = \nu}} \int_{\Omega} c(x, T(x))\, \mu(dx)$$

This form is the Monge problem (older & less general).

Examples:

1/ $\Omega = \Omega' = \mathbb{R}$

Then $T = F_\nu^{\langle -1 \rangle} \circ F_\mu$ (Left inverse)

where $\begin{cases} F_\nu^{\langle -1 \rangle}(x) := \inf\{t \geq 0 \mid F_\nu(x) \geq t\} \\ F_\mu(t) = \mu(]-\infty, t]) \end{cases}$

2/ If $\begin{cases} \mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \\ \nu = \frac{1}{n} \sum_{j=1}^{n} \delta_{y_j} \end{cases}$

Then the Brenier map identifies to a permutation

$T \simeq \sigma : \{1, 2, \dots, n\} \longrightarrow \{1, 2, \dots, n\}$

via $T(x_i) = y_{\sigma(i)}$. Also, upon writing

$C_{ij} = C(x_i, y_j)$, the transport problem reduces to

$$\boxed{\inf_{\sigma \in \mathfrak{S}_n} \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)}}$$

The Assignment/Matching problem.

This is the most relevant example in ML. Computers work with discrete masses)

I.2/ Topology & Geometry: Geodesics

Here $c(x,y) = |x-y|^p$

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\, \pi(dx,dy) \right)^{1/p}$$

Thm: If $\{\Omega, \Omega'$ compact $p \geq 1$ then

W_p is a distance over $\mathcal{M}_1(\Omega)$ & the induced topology is equivalent to the topology of weak convergence

Better than that, $(\mathcal{M}_1(\Omega), W_p)$ is an arc length space: curves have a good notion of speed & geodesics can be defined

Thm [McCann interpolation]

If μ, ν belong to $(\mathcal{M}_1(\Omega), W_p)$, then ~~there~~

- there is a geodesic curve between them

$$\begin{cases} \mu_t : [0, \underline{1}] \longrightarrow \mathcal{M}_1(\Omega) \\ \mu_0 = \mu \,;\; \mu_1 = \nu \end{cases}$$

- when μ_t is parametrized by Arclength & π optimal coupling then

$$\mu_t = (P_t)_{\#} \pi \quad \& \quad P_t : \begin{array}{ccc} \Omega \times \Omega & \longrightarrow & \Omega \\ (x,y) & \longmapsto & x + t(y-x) \end{array}$$

(Linear interpolation)

In particular, if π is given by a Brenier map,

$$\mu_t = [(1-s)\,\mathrm{id} + s\,T]_* \mu$$

And in the ever more particular case of the assignment problem: $\mu = \frac{1}{n}\sum_i \delta_{x_i}$

$$\nu = \frac{1}{n}\sum_j \delta_{y_j}$$

Then $\mu_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)}$ where $x_i(t) = (1-t)\,x_i + t\,T(x_i)$

$= (1-t)\,x_i + t\,y_{\sigma(i)}$

↑ In the tutorial, there will be a drawing of that!

I-3/ Duality

As a particular case of convex duality, which transforms a primal problem to its convex dual.

Thm: Kantorovich Duality:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\,\pi(dx,dy) = \sup_{f,g} \int f\,d\mu + \int g\,d\nu$$

$$f(x) + g(y) \leqslant c(x,y)$$

In particular for the assignement problem:

$$\inf_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} c_{i,\sigma(i)} = \sup_{f_i + g_j \leqslant c_{ij}} \sum_{i=1}^{n} f_i + \sum_{j=1}^{n} g_j.$$

II The Hungarian algorithm.

If the complexity of testing all combinations is $O(n!)$, the above duality is the basis of the so-called { Hungarian algorithm / Munkres algorithm } whose complexity is $O(n^3)$.

The name "Hungarian algorithm" is in honor of Dénes König & Jenő Egerváry; as the validity of the ~~Hung~~ algorithm relies on graph theory due to them.

Description of the algorithm:

Input: The Cost Matrix $C = (C_{ij})_{1 \le i,j \le n}$

Ex: $C_1 = \begin{pmatrix} 5 & 9 & \textcircled{1} \\ 10 & 3 & \textcircled{2} \\ 8 & 7 & \textcircled{4} \end{pmatrix}$

Step 1: • Remove smallest entry from each row.

$\begin{pmatrix} \textcircled{4} & 8 & 0 \\ 8 & \textcircled{1} & 0 \\ 4 & 3 & 0 \end{pmatrix}$

• Same for columns

$\begin{pmatrix} 0 & 7 & 0 \\ 4 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}$

⇝ Each row & column contain a zero.

Step 2:

• Draw lines through row & columns that have 0 entries with ~~least~~ fewest possible lines

$\begin{pmatrix} 0 & 7 & 0 \\ 4 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}$ Here 3 is smallest

Step 3:

If n lines are drawn, then we can find an assignement of 0.

DONE

If $k < n$ lines are drawn,

call V the indices (i,j) covered at least by column or line

Λ the indices (i,j) covered by both column & line.

$$\Delta = \min_{(i,j) \notin V} c_{ij}$$

Transform c_{ij} ←

$$\begin{cases} c_{ij} - \Delta, & (ij) \notin V \\ c_{ij}, & (ij) \in V, (ij) \notin \Lambda \\ c_{ij} + \Delta, & (ij) \in \Lambda \end{cases}$$

Loop to step 2.

here for C_1

$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

work

Cost = 12

= 5 + 3 + 4

= 8 + 3 + 1

Example:

$C_2 = \begin{pmatrix} 1 & 4 & 5 \\ 5 & 7 & 6 \\ 5 & 8 & 8 \end{pmatrix}$

Step 1 ⇝ $\begin{pmatrix} 0 & 3 & 4 \\ 0 & 2 & 1 \\ 0 & 3 & 3 \end{pmatrix}$

Step 1 ⇝ $\begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & 0 \\ 0 & 1 & 2 \end{pmatrix}$

Step 2 ⇝ $\begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & 0 \\ 0 & 1 & 2 \end{pmatrix}$

Step 3 ⇝ $\begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ $\Delta = 1$

Step 2 ⇝ $\begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 0 & & \\ & 0 & 0 \\ & 0 & \end{pmatrix}$

Cost = 15.

= 6 + 4 + 5

= 6 + 8 + 1

III - Sinkhorn algorithm

Again, in 2013, another little revolution happened in the numerical solving of the optimal transport problem.

1/ Marco Cuturi (2013) had the idea of introducing ~~another~~ a penalized problem, which has better convexity properties:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\,\pi(dx,dy) + \varepsilon \,\underline{\underline{\mathrm{Ent}}}(\pi)$$

↳ Entropy

Written for a discrete problem:

$$\inf_{\pi \in \Pi(\mu,\nu)} \sum_{i,j=1}^{n} c_{ij}\pi_{ij} + \varepsilon\, \pi_{ij} \log \pi_{ij}$$

$$\Longleftrightarrow \begin{cases} \pi \mathbb{1} = \mu \\ \pi^T \mathbb{1} = \nu \end{cases}$$

$$= \inf_{\substack{\pi \in M_n(\mathbb{R}_+), \\ \pi \mathbb{1} = \mu \\ \pi^T \mathbb{1} = \nu}} \sum_{i,j=1}^{n} c_{ij}\,\pi_{ij} + \underbrace{\varepsilon\, \pi_{ij} \log \pi_{ij}}_{\text{gives convexity}}$$

Method of Lagrangian Multipliers

$$\mathcal{L}(\pi, \alpha, \beta) = \sum_{i,j=1}^{n} c_{ij}\,\pi_{ij} + \varepsilon\,\pi_{ij}\log\pi_{ij} - \sum_i \alpha_i\left(\pi\mathbb{1} - \mu\right)_i - \beta_i\left(\pi^T\mathbb{1} - \nu\right)_i$$

(π, α, β) critical point

$$\Longleftrightarrow \begin{cases} \pi\mathbb{1} = \mu\,; \quad \pi^T\mathbb{1} = \nu \\ 0 = \dfrac{\partial \mathcal{L}}{\partial \pi_{ij}} = c_{ij} + \varepsilon\left(\log\pi_{ij} + 1\right) - \alpha_i - \beta_j \end{cases}$$

$$\Longleftrightarrow \begin{cases} \pi\mathbb{1} = \mu\,; \quad \pi^T\mathbb{1} = \nu \\ \log\pi_{ij} = -1 + \dfrac{1}{\varepsilon}\left(\alpha_i + \beta_j - c_{ij}\right) \end{cases}$$

$$\Longleftrightarrow \begin{cases} \pi_{ij} = e^{-1}\, e^{-\frac{c_{ij}}{\varepsilon}}\, e^{\frac{1}{\varepsilon}\alpha_i}\, e^{\frac{1}{\varepsilon}\beta_j} \\ \pi\mathbb{1} = \mu\,; \quad \pi^T\mathbb{1} = \nu \end{cases}$$

$$\Longleftrightarrow \begin{cases} \pi_{ij} = [\text{illegible}]\, u_i\, e^{-\frac{c_{ij}}{\varepsilon}}\, v_j \\ u_i, v_j > 0 \\ \pi\mathbb{1} = \mu\,; \quad \pi^T\mathbb{1} = \nu \end{cases}$$

Hence **Proposition:** The ε-penalized problem has a unique solution of the form

$$\pi_{ij} = u_i\, e^{-\frac{c_{ij}}{\varepsilon}}\, v_j \qquad u_i, v_j > 0$$
$$\pi\mathbb{1} = \mu\,; \quad \pi^T\mathbb{1} = \nu$$

2/ Sinkorn & Knopp (1967)

had the idea much earlier of solving this algorithm via a fixed-point interation

Lemma: $K = (K_{ij} = e^{-\frac{c_{ij}}{\varepsilon}})$ $\quad \pi_{ij} = u_i K_{ij} v_j$

$$\pi \mathbb{1} = \mu \implies u_i = \frac{\mu_i}{(Kv)_i}$$

$$\pi^T \mathbb{1} = \nu \implies v_i = \frac{\nu_i}{(K^T u)_i}$$

Proof: Easy.

Hence the idea:

while $|\pi \mathbb{1} - \mu|_1 + |\pi^T \mathbb{1} - \nu|_1 \geq$ Error

$u \leftarrow (\mu_i / (Kv)_i)_i$

~~$\pi \leftarrow (u_i K_{ij} v_j)$~~

$v \leftarrow (\nu_i / (K^T u)_i)_i$

End while

$\pi \leftarrow (u_i K_{ij} v_j)_{1 \leq i,j \leq n}$