Probabilistic & computational aspects of OT

Plan :

I What is OT?
- i) The Monge Problem
- ii) Relaxat.° to the Kantorovich Problem.
- iii) Discrete setting

II Kantorovich duality

III Computational aspects : Sinkhorn

IV From Kantorovich to Monge : Brenier's Thm.

I What is OT?

Let X & Y be Polish spaces.

$\mu \in \mathcal{M}_1(X)$ ↦ $\mathbb{P}$-measures on X

$\nu \in \mathcal{M}_1(Y)$

(often : X & Y will be : • open subsets of $\mathbb{R}^n$
• discrete)

i) ↦ Monge's historical problem.
(1781 « Mémoire ~~Théorie~~ sur la théorie des déblais & remblais »

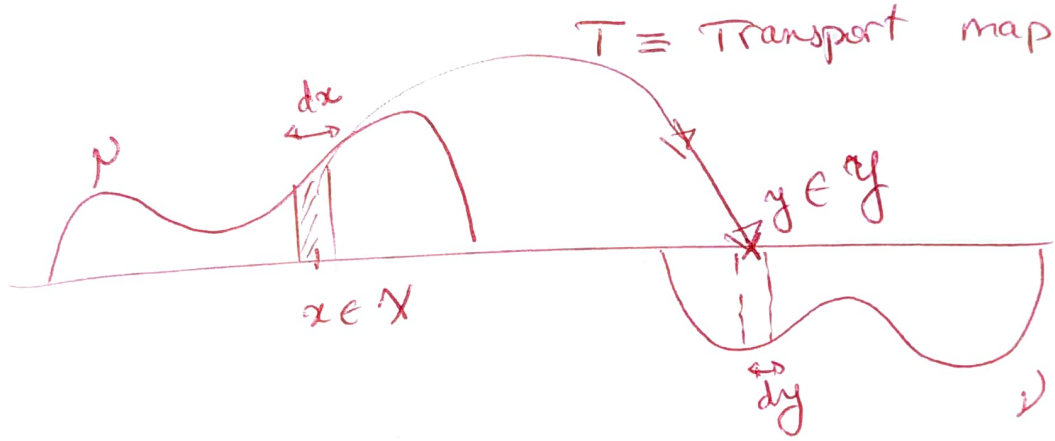
≃ « Dissertat.° on the theory of excavat.° & backfills »

where he asks the question

"What is the optimal way of moving a distribut° μ (of dirt, to excavate) to a distribut° ν (of dirt, to fill a hole) ?"

To that end, let

$$c : X \times Y \longrightarrow \mathbb{R} \sqcup \{+\infty\}$$ Cost funct°



Monge's problem, formalized

Find a transport map $T : X \to Y$ realizing the inf in

$$W_c^{\text{Monge}}(\mu, \nu) = \inf_{T : T_* \mu = \nu} \int_X \mu(dx)\, c(x, T(x))$$

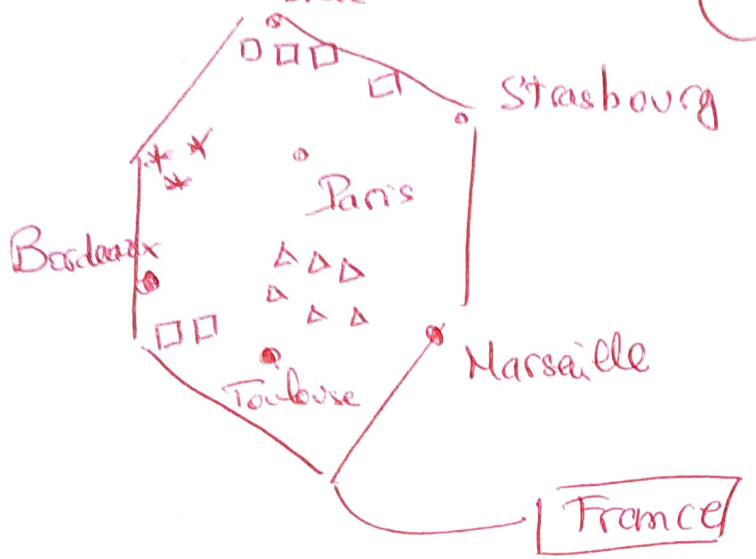
⇝ Difficult. Non-linear constrains on the space of maps T.

ii) ↳ Kantorovich relaxat°

In 1942, Kantorovich proposed a relaxat° of these constrains in a paper

« On the translocat° of masses »

Motivat°: Economics & Central Planning. (↑ Actually same as Monge)



□ Coal mines
✱ Salt mines
△ Quarries

& Kantorovich was thinking metal mines & factories.

Key idea: Allow the splitting of masses.

Not all the mass $\mu(dx)$ at x need to move to $T(x)$ a unique point

but it can ~~move to~~ be split over $\pi_x(dy)$ for $\neq y$ ← conditional probability

$$\int \mu(dx) \int c(x,y)\, \pi_x(dy) = \int \pi(dx\,dy)\, c(x,y)$$

↑ marginal.

Kantorovich Problem (Formalized)

$$W_c(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times Y} \pi(dx,dy)\, c(x,y)$$

$$= \inf_{\mathcal{L}(X)=\mu,\ \mathcal{L}(Y)=\nu} \mathbb{E}\, c(X,Y)$$

couplings ↑ ← marginals

Clearly, we have $W_c^{\text{Monge}}(\mu,\nu) \geq W_c(\mu,\nu)$

because $\left\{ \left(\mu(dx) \otimes_{T(x)} (dy)\right);\ T_\# \mu = \nu \right\} \subsetneq \Pi(\mu,\nu)$

Rmk: Considerable conceptual gain; as Kantorovich's problem is:

- linear in π
- ~~convex in π~~ $\in \Pi(\mu,\nu)$

~~$\Pi(\mu,\nu)$~~ is convex + linear constrain.

$\rightsquigarrow$ Foundat° of linear programming & operat° research

iii) Discrete setup

$$\mathcal{X} = \{x_1, x_2, \cdots, x_n\} \simeq \{1, 2, \cdots, n\} = [\![n]\!]$$

$$\mathcal{Y} = \{y_1, \cdots, y_m\} \simeq [\![m]\!]$$

$$\mu = (\mu_1, \cdots, \mu_n) \in \mathbb{R}_+^n \text{ with } \sum \mu_i = 1$$

$$\nu = (\nu_1, \cdots, \nu_m) \in \mathbb{R}_+^m \text{ with } \sum \nu_j = 1$$

$$\Pi(\mu,\nu) \ni \pi = (\pi_{ij}) \text{ with } \sum_j \pi_{ij} = \mu_i, \quad \sum_i \pi_{ij} = \nu_j$$

$\rightsquigarrow$ Always compact

$$\pi_* = \operatorname*{Argmin}_{\pi \in \Pi(\mu,\nu)} \int \pi(dx\,dy)\, c(x,y) \quad \text{reached!}$$

!! In general, need for Prokhorov!

for general Polish spaces

Even more particular case: $m = n$

$\mu = \nu$ = uniform

$\tilde{\pi}$

$\| \quad \rightsquigarrow n\pi$ is just a bistochastic matrix

$\rightsquigarrow W_c(\mu,\nu) = \inf_{\tilde{\pi} \in \mathcal{P}_n} \left\{ \frac{1}{n} \sum c(x_i, y_j)\, \tilde{\pi}_{ij} \right\}$

$\tilde{\pi} \in \mathcal{P}_n \leftrightsquigarrow$ bistochastic matrices

Fact: (Choquet) Optimum reached on $\mathrm{ext}(\mathcal{P}_n)$

(Birkhoff) $\mathrm{ext}(\mathcal{P}_n) = \mathcal{P}_n$.

$\rightsquigarrow$ Matching Problem

$$W_c(\mu,\nu) = \inf_{\sigma \in \mathcal{P}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) \right\}$$

Solvable using the Hungarian algorithm / Kuhn-Munkres in $\mathcal{O}(n^3)$ (See Wikipedia).

II Kantorovich duality.

This nothing but convex duality which holds at a high level of generality

Thm [Kantorovich (convex) duality]

Let $\begin{cases} X \text{ & } Y & \text{Polish spaces} \\ c : X \times Y \longrightarrow \mathbb{R}_+ \cup \{+\infty\} & \text{lower semi-continuous} \end{cases}$

Then
$$W_c(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \pi(dx\,dy)\, c(x,y).$$
$$= \sup_{(\varphi,\psi) \in \Phi_c} \int \varphi(x)\,\mu(dx) + \int \psi(y)\,\nu(dy)$$

with $\Phi_c = \{ (\varphi,\psi) \in \mathcal{C}_b(X) \times \mathcal{C}_b(Y) \mid \varphi(x) + \psi(y) \leq c(x,y) \quad \forall x \in X \ \mu\text{-a.e.}, \ \forall y \in Y, \ \nu\text{-a.e.} \}$

In the discrete setting $\begin{cases} X \simeq [\![n]\!] \\ Y \simeq [\![m]\!] \end{cases}$

the statement is much more simple:

Corollary (Kantorovich 1942)

$$W_c(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \sum_{\substack{i \\ \| \\ c_{ij}}} c(x_i,y_j)\,\pi_{ij}$$
$$= \sup_{(f,g) \in \mathbb{R}^n \times \mathbb{R}^m} \sum_i f_i \mu_i + \sum_j g_j \nu_j$$
$$f_i + g_j \leq c_{ij} \quad \leftarrow \text{better to saturate}$$

Proof: Identify μ, ν to vectors, π to a matrix, C as well.

Recall $\Pi(\mu,\nu) = \left\{ \pi \in \mathbb{R}_+^{m \times m} \;\middle|\; \begin{array}{l} \pi \mathbb{1}_m = \mu \\ \pi^T \mathbb{1}_m = \nu \end{array} \right\}$

$$\rightsquigarrow \quad W_C(\mu,\nu) = \min_{\substack{\pi \in \mathbb{R}_+^{m \times m} \\ \pi \mathbb{1}_m = \mu \leftarrow n \text{ contrains} \\ \pi^T \mathbb{1}_m = \nu \leftarrow m \text{ ——}}} \langle \pi, C \rangle$$

Linear problem with constraints $\rightsquigarrow$ Introduce Lagrangian.

Lagrange multipliers are $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^m$

$$\mathcal{L}(\pi, f, g) = \langle \pi, C \rangle + \langle f, \pi \mathbb{1}_m - \mu \rangle + \langle g, \pi^T \mathbb{1}_n - \nu \rangle$$

$$\frac{\partial \mathcal{L}}{\partial f_i} = 0 \implies \mu_i = (\pi \mathbb{1}_m)_i$$

$$\frac{\partial \mathcal{L}}{\partial g_j} = 0 \implies \nu_j = (\pi^T \mathbb{1}_m)_j$$

$$\frac{\partial \mathcal{L}}{\partial \pi_{ij}} = 0 \implies 0 = C_{ij} + (f_i + g_j)$$

$$\iff f_i + g_j = C_{ij}$$

Now $W_C(\mu,\nu) = \min_{\pi \in \mathbb{R}_+^{n \times m}} \max_{f,g}$

$$W_C(\mu,\nu) = \langle \pi^*, C \rangle = \sum \pi^*_{ij} (f^*_i + g^*_j)$$

$$= \sum_i f^*_i \underbrace{\left(\sum_j \pi^*_{ij} \right)}_{\mu_i} + \sum_j g^*_j \underbrace{\left(\sum_i \pi^*_{ij} \right)}_{\nu_j} = \langle f^*, \mu \rangle + \langle g^*, \nu \rangle$$

$$= \sup_{f \oplus g \leq C} \langle f, \mu \rangle + \langle g, \nu \rangle$$

Remarks & consequences of Kantorovich duality

① Economical interpretat°

* $W_c(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \sum \pi_{ij} c_{ij}$ ← Minimising cost

$= \sup_{f_i + g_j \le c_{ij}} \sum f_i \mu_i + \sum g_j \nu_j$ ← Maximizing gain

f_i gain in buying one unit

g_j gain in selling one unit

* Cost / Surplus.

If $c(x,y) \le a(x) + b(y)$

$a \in L^1(\mu),\ b \in L^1(\nu)$

$S(x,y) = -c(x,y) + a(x) + b(y)$ "Surplus"

$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\,\pi(dx\,dy) = \mu(a) + \nu(b) - \sup_{\pi \in \Pi(\mu,\nu)} \int s(x,y)\,\pi(dx\,dy)$

② c-concavity and the double convexification trick

Def: [c-concavity]

$c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ given.

$\varphi : X \to \mathbb{R} \cup \{-\infty\}$ is called c-concave if $\exists\, \psi : Y \to \mathbb{R} \cup \{-\infty\}$, $\psi \not\equiv -\infty$

st $\varphi(x) = \inf_{y \in Y} c(x,y) - \psi(y)$. We write $\varphi = \psi^c$.

Important example: $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$

before $c(x,y) = \frac{\|x-y\|^2}{2} = \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} - \underbrace{\langle x, y\rangle}_{s(x,y)}$

Given $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$,

Then $\varphi^c(y) = \inf_{x \in \mathbb{R}^n} \frac{1}{2}\|x-y\|^2 - \varphi(x)$

$\Leftrightarrow \underbrace{\left(\varphi^c - \frac{\|\cdot\|^2}{2}\right)}_{-f^*}(y) = \inf_{x \in \mathbb{R}^n} -\langle x, y\rangle - \underbrace{\left(\varphi - \frac{\|\cdot\|^2}{2}\right)}_{-f}(x)$

$\Leftrightarrow f^*(y) = \sup_{x \in \mathbb{R}^n} \langle x, y\rangle - f(x)$

usual Legendre - Fenchel duality.

$\boxed{W_2(\mu,\nu) = \inf \int \frac{\|x-y\|^2}{2} \pi(dx,dy)}$

very geometric with strong links to Legendren - Fenchel (↝ Part IV).

Now **Fact 1:** $(\varphi^{cc})^c = \varphi^c$

Fact 2: $\mu(\varphi) + \nu(\psi)$ with $\varphi(x) + \psi(y) \leq c(x,y)$

$\leqslant \mu(\varphi) + \nu(\varphi^c)$

$\leqslant \mu(\varphi^{cc}) + \nu(\varphi^c)$

Double convexification trick:

$\rightsquigarrow W_c(\mu,\nu) = \inf_{\pi} \langle \pi, c\rangle = \sup_{\varphi} \mu(\varphi^{cc}) + \nu(\varphi^c)$

③ Kantorovich-Rubinstein ($c=d$ is a distance)

Thm [Kantorovich-Rubinstein]

Let $X=Y$, $c=d$ is a distance.

Then
$$W_1(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int d(x,y)\,\pi(dx\,dy)$$
$$= \sup_{\|\varphi\|_{Lip}\le 1} |\mu(\varphi)-\nu(\varphi)|$$

Proof:
$$W_1(\mu,\nu) \overset{K.\text{ duality}}{=} \sup_{\varphi(x)+\psi(y)\le d(x,y)} \mu(\psi)+\nu(\psi)$$
$$\overset{2x\ \text{convexificate}}{=} \sup_{\varphi \text{ bounded}} \mu(\varphi^d)+\nu(\varphi^{dd}) \quad (\star)$$

But: **Fact 1:** $\varphi^d(y_1) = \inf_{x\in X} d(x,y_1)-\varphi(x)$ is 1-Lip

$\le d(x,y_2)+d(y_1,y_2)$

$\le d(y_1,y_2)+\varphi^d(y_2)$

Fact 2: $-\varphi^d = \varphi^{dd}$

Indeed $-\varphi^d(x) \le \varphi^{dd}(x) = \inf_y d(x,y)-\varphi^d(y)$

$\ge \varphi^d(x)$ as 1-Lip

$\le -\varphi^d(x)$ $x=y$.

from $(\star)$

Therefore $W_1(\mu,\nu) = \sup_{\varphi \text{ bounded}} \mu(\varphi^d)-\nu(\varphi^d)$

φ^d is 1-Lip $\le$ $\sup_{\|\varphi\|_{Lip}\le 1} \mu(\varphi)-\nu(\varphi)$

φ 1-Lip $\Rightarrow (\varphi,-\varphi)\in\Phi_c$ $\|\varphi(x)+\varphi(y)\le d(x,y)\|$ $\sup_{(\varphi,\psi)\in\Phi_c} \nu(\varphi)+\nu(\psi) = W_1(\mu,\nu)$ ∎

III Computational aspects of OT: Sinkhorn

Ref: Cuturi, Peyré "Computational OT"

1) Idea: * Strongly convex optim. problem are better.

* Regularize using entropy / KL:

$$\forall \pi \in \mathbb{R}_+^{n\times m}, \quad H(\pi) = \begin{cases} +\sum \pi_{ij}(\log \pi_{ij} - 1) & \text{if } \pi_{ij} > 0 \\ -\infty & \text{otherwise} \end{cases}$$

Fact: $\nabla^2 H \leq -\mathrm{Id}$. 1-concave.

2) Entropy regularized OT.

Define $W_c^\varepsilon(\mu,\nu) = \inf_{\pi \in \mathbb{R}_+^{n\times m},\ \sum \pi_{ij}=1} \underbrace{\langle C, \pi\rangle}_{\text{linear}} - \varepsilon \underbrace{H(\pi)}_{1\text{-concave}}$

ε-convex.

$$\pi \mathbb{1}_m = \mu$$
$$\pi^T \mathbb{1}_n = \nu$$

$\rightsquigarrow \exists!$ minimiser π_ε^*

Theorem (Properties)

- Duality

$$W_c^\varepsilon(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \langle C, \pi\rangle - \varepsilon H(\pi)$$
$$= \sup_{(f,g)\in \mathbb{R}^n \times \mathbb{R}^m} \mu(f) + \nu(g) - \varepsilon \sum_{ij} e^{\frac{f_i + g_j - C_{ij}}{\varepsilon}}$$

- Unique optimizers f^*, g^*, π^*

- Simple relat° between primal and dual.

(⋆) $$(\pi_\varepsilon^*)_{ij} = \exp\left(\frac{f_i^* + g_j^* - C_{ij}}{\varepsilon}\right)$$

Proof: Write Lagrange multipliees & do like when $\varepsilon = 0$.

More over, $\lim_{\varepsilon \to 0} \pi_\varepsilon^*$ exists and solves the K. problem.

Lemma: $$\pi_\varepsilon^* \xrightarrow[\varepsilon \to 0]{} \operatorname{Argmin}_{\pi} \{ -H(\pi) \mid W_C(\mu, \nu) = \langle C, \pi \rangle$$

~~Proof: write first order condit° Admi~~

Proof: ~~Compare~~ π_ε^* Take a CV subsequence π_ε^* and look how limit behaves when plugged into to minimizat° problem.

(3) Sinkhorn alg. or proportional fitting

Write $K = e^{-\frac{C_{ij}}{\varepsilon}}$ (Gibbs Kernel).

From (⋆) π_ε^* is the only rescaling of columns & rows such that $\pi_\varepsilon^* \in \Pi(\mu, \nu)$

⇝ Iteratively rescale columns, then rows in order to have the correct marginals.

Algorithm:

$P \in \mathbb{R}_+^{n \times m}$

$u^{(0)} \in \mathbb{R}^n$

$v^{(\ell)} \in \mathbb{R}^m$

$u^{(0)} \leftarrow 1 \; ; \; v^{(0)} \leftarrow 1$

While $P = \text{diag}(u^{(\ell)}) K \, \text{diag}(v^{(\ell)})$

satisfies $\| P \mathbb{1}_m - \mu \| + \| P^T \mathbb{1}_n - \nu \| \geq \text{error}$

$u^{(\ell+1)} \leftarrow \mu \,/\, K \cdot v^{(\ell)}$ (1)

$v^{(\ell+1)} \leftarrow \nu \,/\, K^T \cdot u^{(\ell+1)}$ (2)

element-wise

matrix-vector

IV Brenier's them:

- $(\mu, \nu) \in \mathcal{M}_1(\mathbb{R}^n) \times \mathcal{M}_1(\mathbb{R}^n)$ with 2^{nd} moments
- $c(x,y) = \frac{\|x-y\|^2}{2}$ (W_2 !)
- ~~μ has a density~~

$\Leftarrow\!\!\Rightarrow$ $\mu(A) = 0 \;\; \forall A$ with $\dim_H(A) \leq n-1$

$\Leftarrow$ μ has density

Then $\exists!\, \pi$ optimal, $\pi(dx\,dy) = \mu(dx)\, \delta_{T(x)}(dy)$

with $T_\# \mu = \nu$.

Here $T = \nabla\left(\underbrace{\varphi}_{\text{convex}} - \frac{|\cdot|^2}{2}\right)$ gradient of convex fonct°.