LECTURE 3

ENTROPIC OPTIMAL TRANSPORT AND NUMERICS

- DISCRETE OPTIMAL TRANSPORT

two finite sets X and Y
with cardinality N

$$\mu = \sum_{x \in X} \mu_x \delta_x \qquad \nu = \sum_{y \in Y} \nu_y \delta_y$$

$$\sum \mu_x = \sum \nu_y = 1$$

definition (Discrete OT) μ, ν and a cost $c : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$

$$\inf \Big\{ \sum_{x \in X} \sum_{y \in Y} \gamma_{xy}\, c(x,y) \;\Big|\; \gamma \in \Pi(\mu,\nu) \Big\}$$

γ is a matrix!

element wise $\sum_i \sum_j \gamma_{ij} c_{ij}$

$$\Pi(\mu,\nu) = \left\{ \gamma \in \mathbb{R}^{X \times Y} \;\middle|\; \gamma_{xy} \geq 0 \right.$$

$$\sum_y \gamma_{xy} = \mu_x \quad \forall x \in X$$

$$\left. \sum_x \gamma_{xy} = \nu_y \quad \forall y \in Y \right\}$$

Rmk : linear programming problem

$$\min_y \langle C \mid y \rangle \qquad \Big\|$$
$$Ay = b$$

Rmk : LP has complexity $O(N^3)$

$\rightarrow$ Entropic regularization

- Entropic OT : the discrete case

Idea : introduce an entropic

$$\mathrm{Ent}(\gamma) := \sum_{x,y} e(\gamma_{xy}) \quad \text{to}$$

penalize the non-negative constraint

$$e(r) := \begin{cases} r(\log r \;\; -1) & r > 0 \\ 0 & r = 0 \\ +\infty & r < 0 \end{cases}$$

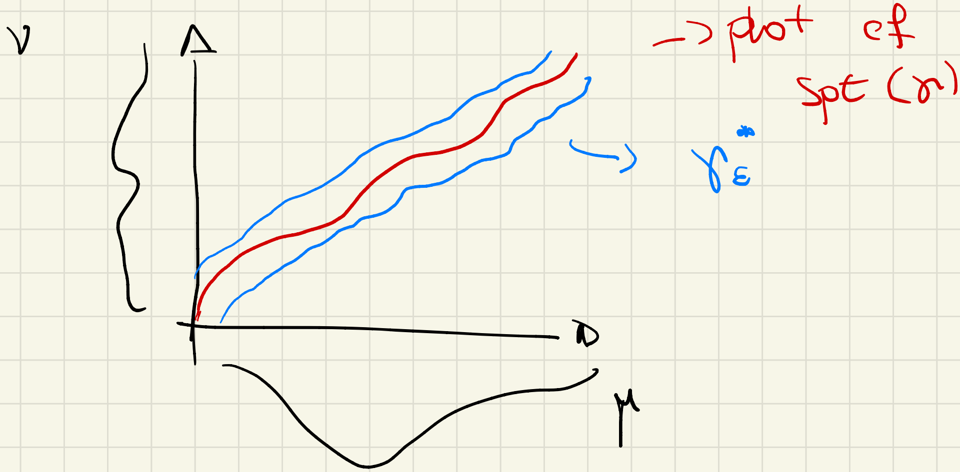
Given a parameter $\varepsilon > 0$ we consider the following minimization pb

$$P_\varepsilon = \inf \left\{ \langle \gamma | c \rangle + \varepsilon \operatorname{Ent}(\gamma) \;\middle|\; \begin{array}{l} \gamma \in \mathbb{R}^{X \times Y} \\ \sum_y \gamma_{xy} = \mu_x \quad \forall x \\ \sum_x \gamma_{xy} = \nu_y \quad \forall y \end{array} \right\}$$

where $\langle \gamma | c \rangle = \sum_{xy} \gamma_{xy} c(x,y)$

<u>thm</u> The problem P_ε has a unique solution γ^ε which belongs to $\Pi(\mu, \nu)$.
Moreover if $\min \{ \min_x \mu_x , \min_y \nu_y \} > 0$ then

$$\gamma_{xy} > 0 \quad \forall (x,y) \in X \times Y$$



thm (Convergence in ε) The unique solution γ_ε^* to P_ε converges to the optimal solution with minimal entropy within the set of all optimal solutions of the OT problem that is

$$\gamma_\varepsilon \xrightarrow[\varepsilon \downarrow 0]{} \operatorname{argmin} \left\{ \mathrm{Ent}(\gamma) \;\middle|\; \gamma \in \Pi(\mu,\nu),\ \langle c | \gamma \rangle = \Pi K_c(\mu,\nu) \right\}$$

proof

Consider a sequence of $(\varepsilon_k)_k$ s.t. $\varepsilon_k \to 0$ $\varepsilon_k > 0$ denote γ_k the solution to P_ε with $\varepsilon = \varepsilon_k$

Since $\Pi(\mu,\nu)$ is bounded and closed we can extract a converging subsequence $\gamma_k \longrightarrow \gamma^* \in \Pi(\mu,\nu)$.

Take a γ optimal for the unreg. problem by optimality of γ_n and γ we have:

$$\langle \gamma_n | c \rangle + \varepsilon_n \mathrm{Ent}(\gamma_n) - \langle \gamma | c \rangle - \varepsilon_n \mathrm{Ent}(\gamma) \leq 0$$

$$\langle \gamma_n | c \rangle - \langle \gamma | c \rangle \leq \varepsilon_n \left(\mathrm{Ent}(\gamma) - \mathrm{Ent}(\gamma_n) \right)$$

$$0 \leq \langle \gamma_n | c \rangle - \langle \gamma | c \rangle$$

Since $\mathrm{Ent}(\cdot)$ is cont. by taking the limit $k \to +\infty$

$$0 \leq \langle \gamma^* | c \rangle - \langle \gamma | c \rangle \leq 0$$

$$\Rightarrow \langle \gamma^* | c \rangle = \langle \gamma | c \rangle$$

Divide by ε_k

$$\mathrm{Ent}(\gamma) - \mathrm{Ent}(\gamma_k) \geq 0$$

$$\Big\downarrow \quad k \uparrow +\infty$$

$$\mathrm{Ent}(\gamma) \geq \mathrm{Ent}(\gamma^*)$$

$\boxtimes$

Lagrangian associated to P_ε

$$\mathcal{L}(\gamma, \varphi, \psi) = \sum_{x,y} \Big(\gamma_{xy}\, c(x,y) + \varepsilon\, e(\gamma_{xy}) \Big)$$
$$+ \sum_x \varphi_x \Big(\mu_x - \sum_y \gamma_{xy} \Big)$$
$$+ \sum_y \psi_y \Big(\nu_y - \sum_x \gamma_{xy} \Big)$$

$$P_\varepsilon = \inf_\gamma \sup_{\varphi, \psi} \mathcal{L}(\gamma, \varphi, \psi)$$

by interchanging inf and sup

$$D_\varepsilon = \sup_{\varphi, \psi} \min_\gamma \mathcal{L}(\gamma, \varphi, \psi)$$

the optimal must satisfy for γ given φ, ψ

$$D_{\gamma_{xy}} \mathcal{L} = 0$$

$$c(x,y) + \varepsilon \log(\gamma_{xy}) - \psi_y - \varphi_x = 0$$

$$\Rightarrow \quad \gamma_{xy} = \exp\left(\frac{\varphi_y + \psi_x - c(x,y)}{\varepsilon} \right)$$

<u>def</u> (Dual problem)

$$D_\varepsilon = \sup_{\varphi, \psi} \Phi_\varepsilon(\varphi, \psi) \quad \text{with}$$

$$\Phi_\varepsilon(\varphi, \psi) = \underbrace{\sum_x \varphi_x \mu_x + \sum_y \psi_y \nu_y}_{\text{classic OT}} - \sum_{xy} \varepsilon \exp\left(\frac{\varphi_x + \psi_y - c(x,y)}{\varepsilon} \right)$$

- Weak duality holds $P_\varepsilon \geq D_\varepsilon$
- <u>thm</u> (strong duality) Strong duality holds and the maximum in the dual problem is attained, that is $\exists \varphi \in \mathbb{R}^X, \psi \in \mathbb{R}^Y$ s.t

$$P_\varepsilon = D_\varepsilon = \Phi_\varepsilon(\varphi, \psi)$$

cor If (φ, ψ) is the solution to D_ε, then the solution γ_ε^* to P_ε is given by

$$\gamma^*_{xy,\varepsilon} = \exp\left(\frac{\varphi_x + \psi_y - c(x,y)}{\varepsilon}\right)$$

the optimal coupling can be written as

$$\gamma_{xy} = D_\varphi \, e^{-\frac{c(x,y)}{\varepsilon}} \, D_\psi$$

when D_φ and D_ψ are the diagonal matrices associated to $e^{\varphi/\varepsilon}$ $e^{\psi/\varepsilon}$.

The regularized pb. is similar to a MATRIX SCALING PB.

def (MATRIX SCALING PB) let $K \in M_N(\mathbb{R})$ with positive coeff. Find D_ψ and D_φ positive diagonal matrices s.t.

$D_\varphi K D_\psi$ is DOUBLY STOCHASTIC

(that is sum along each row and each column is equal to 1)

Algorithm 1 Sinkhorn-Knopp algorithm for the matrix scaling problem

function SINKHORN-KNOPP(K)
    $D_\varphi^0 \leftarrow \mathbf{1}_N,\ D_\psi^0 \leftarrow \mathbf{1}_N$
    for $0 \leqslant k < k_{\max}$ **do**
        $D_\varphi^{k+1} \leftarrow \mathbf{1}_N./(KD_\psi^k)$
        $D_\psi^{k+1} \leftarrow \mathbf{1}_N./(K^T D_\varphi^{k+1})$
    end for
end function

Algorithm 2 Sinkhorn-Knopp algorithm for the regularised optimal transport problem

function SINKHORN-KNOPP(K_ε, μ, ν)
    $D_\varphi^0 \leftarrow \mathbf{1}_X,\ D_\psi^0 \leftarrow \mathbf{1}_Y$
    for $0 \leqslant k < k_{\max}$ **do**
        $D_\varphi^{k+1} \leftarrow \mu./(KD_\psi^k)$
        $D_\psi^{k+1} \leftarrow \nu./(K^T D_\varphi^{k+1})$
    end for
end function

$$\gamma^* = D_\varphi K D_\psi \qquad K_{xy} = \exp\left(-\frac{c(x,y)}{\varepsilon}\right)$$

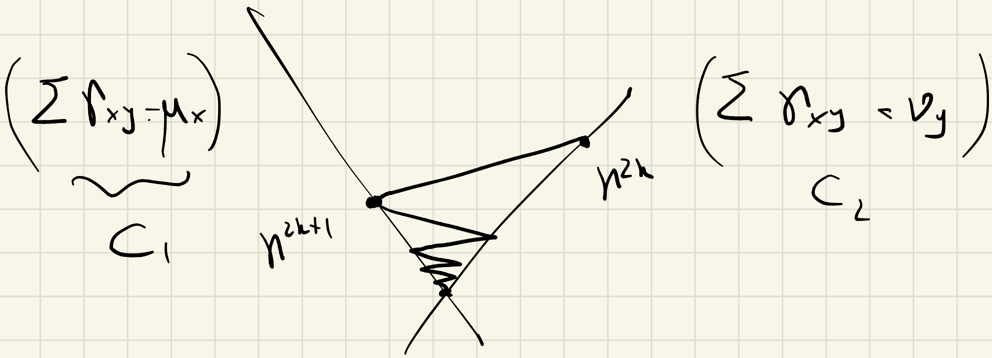
$$\sum_y \gamma^*_{xy} = \mu_x \qquad \forall x$$

$$\Rightarrow \sum_y (D_\varphi)_x K_{xy} (D_\psi)_y = \mu_x$$

$$(D_\varphi)_x = \frac{\mu_x}{\sum (D_\psi)_y K_{xy}} \qquad D_\varphi = \mu./K^T D_\psi$$

$$\gamma^{2k} = D_\varphi^{k}\, K\, D_\psi^{k} \to \nu \text{ as second marginal}$$

$$\gamma^{2k+1} = D_\varphi^{k+1}\, K\, D_\psi^{k} \to \mu \text{ as first marginal}$$



P_ε can be re-written in the following way

$$P_\varepsilon(\mu,\nu) = \inf \left\{ \langle c|\gamma\rangle + H(\gamma|\mu\otimes\nu) \;\middle|\; \begin{array}{l} \gamma \in \mathbb{R}^{X\times Y} \\ \sum_y \gamma_{xy} = \mu_x \\ \sum_x \gamma_{xy} = \nu_y \end{array} \right\}$$

$$H(\rho|\mu) = \sum_x \rho_x \left(\log\left(\frac{\rho_x}{\mu_x}\right) - 1 \right)$$ Relative entropy

or Kullback-Leibler divergence

$$P_\varepsilon(\mu,\nu) = \inf \left\{ H(\gamma \mid K_\varepsilon) \;\middle|\; \begin{array}{l} \gamma \in \mathbb{R}^{X \times Y} \\ \sum_y \gamma_{xy} = \mu_x \\ \sum_x \gamma_{xy} = \nu_y \end{array} \right\}$$

where $K_\varepsilon = \operatorname{diag}(\mu) \exp^{-c/\varepsilon} \operatorname{diag}(\nu)$

- Recasting P_ε in the continuous framework

$$P_\varepsilon(\mu,\nu) = \inf \left\{ \int c \, d\gamma + \varepsilon H(\gamma \mid \mu \otimes \nu) \;\middle|\; \gamma \in \Pi(\mu,\nu) \right\}$$

$$H(\rho \mid \pi) = \begin{cases} \int_{X \times Y} \left(\log \frac{d\rho}{d\pi} - 1 \right) d\rho \, (+1) & \text{if } \rho \ll \pi \\ +\infty & \text{otherwise} \end{cases}$$

$\hookrightarrow$ Static Schrödinger pb.

One can see Sinkhorn as a coordinate ascent algorithm

take a function $f(x,y)$

$$y_{k+1} = \operatorname{argmax}_y f(x_k, y)$$

$$x_{k+1} = \operatorname{argmax}_x f(x, y_{k+1})$$

$y \to \psi$

$x \to \varphi$ from now on $X = Y$

$f \to \Phi_\varepsilon$

Prop The dual pb. reads as

$$D_\varepsilon = \sup \left\{ \Phi_\varepsilon(\varphi, \psi) \;\middle|\; \varphi, \psi \in C_0(X) \right\}$$

$$\Phi_\varepsilon(\varphi,\psi) = \int_X \varphi(x)\, d\mu + \int \psi \, d\nu$$
$$- \varepsilon \int_{X \times Y} \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon} \right) d\mu \otimes d\nu$$

it is strictly concave w.r.t. each argument φ and ψ

and w.r.t. $\varphi + \psi$.

it is Fréchet diff. for $(C_0, \|\cdot\|_\infty)$ topology

Futhermore, if a maximizer exists it is unique up to a constant.

Frechet diff. X, Y normed vect spaces

$O \subset X$ open subset. we say that

$f : O \to Y$ si F-diff. at a point $x \in O$

if $\exists$ a bounded lin. op. $L : X \to Y$ s.t.

$$\lim_{h \to 0} \frac{\| f(x+h) - f(x) - Lh \|_Y}{\|h\|_X} = 0$$

$$Lh = \langle \nabla f(x), h \rangle$$

prop The maximization of Φ_ε w.r.t. each variable can be made explicit and the Sinkhorn algorithm. can be defined as

$$\varphi_{k+1}(x) = -\varepsilon \log \left(\int_X \exp\left(\frac{\psi_n - c(x,y)}{\varepsilon} \right) d\nu \right) := S_\nu(\psi_n)$$

$$\psi_{n+1}(y) = -\varepsilon \log \left(\int_X \exp\left(\frac{\varphi_{n+1} - c}{\varepsilon} \right) d\mu \right) := S_\mu(\varphi_{n+1})$$

Moreover the following properties hold

(i) $\Phi_\varepsilon(\varphi_k, \psi_k) \leq \Phi_\varepsilon(\varphi_{k+1}, \psi_k) \leq \Phi_\varepsilon(\varphi_{k+1}, \psi_{k+1})$

(ii) the continuity modulus of $\varphi_{k+1}, \psi_{k+1}$ is bounded by that $c(x,y)$

(iii) if $\psi_k - c$ $(\varphi_{k+1} - c)$ is bounded by M on the support of ν (μ), then so is φ_{k+1} (ψ_{k+1})

<u>Rmk</u> $\varphi_{k+1} \simeq \min_y c(x,y) - \psi_k(y) \simeq \psi_k^c$

$\psi_{k+1} \simeq \min_x c(x,y) - \varphi_{k+1}(x) = \varphi_{k+1}^c$

$\to$ Log-Sum-Exp function

<u>prop</u> The sequence (φ_k, ψ_k) converges in $(C, \|\cdot\|_\infty)$ to the unique (up to a constant) couple of potentials (φ, ψ) which maximizes Φ_ε

proof: Shifting the potentials by an additive constant, one can have a couple (φ, ψ) which have uniformly bounded modulus of continuity and such that $\varphi(x_0) = 0$ for a given $x_0 \in X$.

From previous prop. (φ_k, ψ_k) are uniformly bounded and have uniformly modulus of continuity

$\Rightarrow$ we can extract a converging subsequence $(\varphi_k, \psi_k) \to (\bar{\varphi}, \bar{\psi})$

by continuity of Φ_ε and the monotonicity of the sequence.

$$\Phi_\varepsilon(\bar{\varphi}, S_\mu(\bar{\varphi})) \leq \Phi_\varepsilon(S_\nu \circ S_\mu(\bar{\varphi}), S_\mu(\bar{\varphi}))$$

$$= \Phi_\varepsilon(\bar{\varphi}, S_\mu(\bar{\varphi}))$$

one has
$$S_\nu(\bar{\psi}) = \bar{\varphi}$$
$$S_\mu(\bar{\varphi}) = \bar{\psi}$$

pseudo-norm of uniform convergence

$$\|f\|_{o,\infty} = \frac{1}{2}\left(\sup f - \inf f\right) = \inf_{a\in\mathbb{R}} \|f+a\|_\infty$$

<u>prop</u> The LSE function is convex and

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{o,\infty} \leq \|\varphi_1 - \varphi_2\|_{o,\infty}$$

<u>then</u> the map $S = S_\nu \circ S_\mu$ is a contraction for $\|\cdot\|_{o,\infty}$. The sequence (φ_1, φ_2) linearly converges to the unique maximizer of Φ_ε

$$\rightarrow \quad 2\left(1 - e^{-2\frac{\|c\|_{o,\infty}}{\varepsilon}}\right)$$