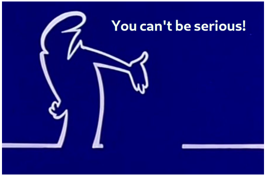
Optimal Transport in a Nutshell

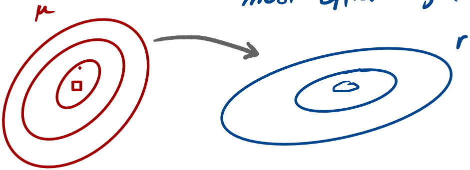
Part 2



ICMS workshop on
'Connections between interacting particle dynamics and data science'

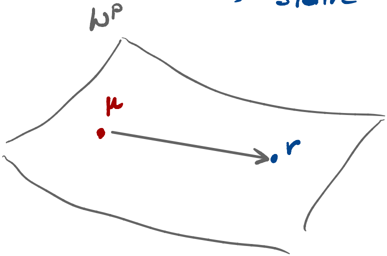
Recap

OT problem: How can we move mass from a source to a target most efficiently?



- Monge map $T(x)$
- Kantorovich plan $\pi(x,y)$
- Dual Kantorovich potential φ

$\rightarrow$ static formulation of OT



TODAY

- Solutions to (KP) define metric
- Dynamic OT Benamou-Brenier
- Wasserstein gradient flows
- Computational OT

Wasserstein distance: Given two probabilities μ and $\nu \in \mathscr{P}_p(X)$ we define

$$d_{W_p}(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \left(\int_{X \times X} |x-y|^p \, d\pi(\mu,\nu) \right)^{\frac{1}{p}} .$$

Vaserstein

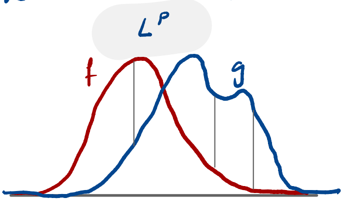
Monge-Kantorovich

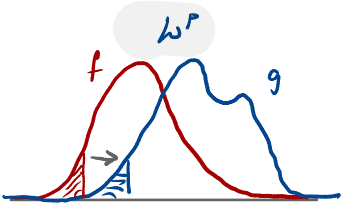
Earth-mover distance

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

$\mathcal{P}_p(X) = \{ \mu \in \mathcal{P}(X) : \int |x|^p \, d\mu(x) < \infty \}$

$W^p(X) = (\mathcal{P}_p(X), d_{W^p})$





$\|f-g\|_{L^p} = \int |f-g|^p \, dx$

vertical distance

horizontal distance

very natural way to describe interacting particle systems

used everywhere

measure to compare distributions

Calculating the Wasserstein distance

- In 1D: Let $\mu, \nu \in \mathscr{P}(\mathbb{R})$ with cdf F and G respectively. Let the cost be a function of the distance, that is $c(x, y) = h(x - y)$, where $h : \mathbb{R} \to [0, \infty)$ is convex and continuous. Then the Kantorovich cost is given by

$$\min_{\pi \in \Pi(\mu,\nu)} K(\pi) = \int_0^1 h\left(F^{-1}(t) - G^{-1}(t)\right) dt.$$

- Between Gaussians: Let $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$ be two Gaussian in $\mathbb{R}^d$, then the map

$$T : x \to m_\nu + A(x - m_\mu) \Leftarrow$$

Gradient of φ

$$\varphi(x) = \frac{1}{2}\langle x - m_\mu, A(x - m_\mu)\rangle + \langle m_\nu, x\rangle$$

with $A = \Sigma_\nu^{-\frac{1}{2}} \left(\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\mu^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_\mu^{-\frac{1}{2}} = A^T$ satisfies $T_\# \rho_\mu = \rho_\nu$.

<u>1D</u> For μ the CDF $F(x) = \int_{-\infty}^{x} d\mu(x)$

Pseudo-inverse $F^{-1}(t) = \min_{x \in \mathbb{R}} \{ x \in \mathbb{R} \cup \{-\infty\} \mid F(x) \geq t \}$

$$d_{W^p}^{\,p} = \| F^{-1} - G^{-1} \|^p_{L^p[0,1]} = \int_0^1 | F^{-1}(t) - G^{-1}(t) |^p \, dt$$

in general $\to$ computationally expensive!

Properties of the Wasserstein distance

- The Wasserstein distance is a metric on $\mathscr{P}_p(X)$.
- Equivalence of W_p distances: for $p \leq q$ Jensen's inequality implies

d_{W^p}

$$\left(\int d(x,y)^p d\pi\right)^{\frac{1}{p}} \leq \left(\int d(x,y)^q d\pi\right)^{\frac{1}{q}},$$

  and therefore $W_p(\mu,\nu) \leq W_q(\mu,\nu)$.
- Convergence in the Wasserstein space:

$$\mu_n \rightharpoonup \mu \Leftrightarrow d_{W^p}(\mu_n,\mu) \to 0.$$

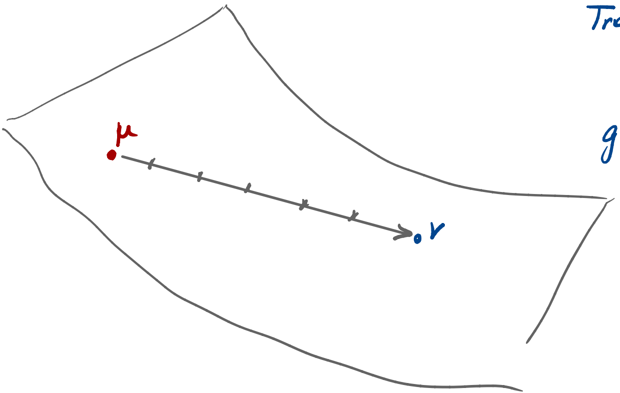
Metric: ◦) symmetric ✓

◦) $d_{W^p}(\mu,\nu) = 0 \Rightarrow \exists!\, \pi$ s.t $\int d(x,y)^p d\pi = 0$

$\pi = (Id, Id)_\# \mu$ ✓

◦) Triangle inequality gluing lemmas...

$W^p = (\mathcal{P}_p(\Omega), d_{W^p})$



Transport plans/maps $\hat{=}$ geodesics in Wasserstein space

Geodesics $\hat{=}$ shortest path wrt. Wasserstein distance

$\Rightarrow$ how can we characterise them?

Curves in metric spaces

- A curve $\omega\colon [0,1] \to X$ is called *absolutely continuous* if there exists a $g \in L^1([0,1])$ such that $d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} g(s)ds$ for every $t_0 < t_1$.
- Consider a curve $\omega\colon [0,1] \to X$. Its *length* is defined as

$$\text{Length}(\omega) := \sup\{\sum_{k=0}^{n-1} d(\omega(t_k), \omega(t_{k+1}))\colon n \geq 1, 0 = t_0 < t_1 \ldots t_n = 1\}.$$

  If $\omega \in AC(X)$ then $\text{Length}(\omega) = \int_0^1 |\omega'(t)| dt$.
- A curve $\omega\colon [0,1] \to X$ is a *geodesic* between x_0 and $x_1 \in X$, if it minimises the length among all curves ω connecting x_0 and x_1.
- A curve $\omega : [0,1] \to X$ is a *constant speed geodesic*, if

$$d(\omega(t), \omega(s)) = |t-s| d(\omega(0), \omega(1)) \qquad \text{for all } s, t \in [0,1].$$

Let $\mu, \nu \in \mathcal{P}_p(X)$, $X \subset \mathbb{R}^d$ compact & convex. Let $\pi \in \Pi(\mu,\nu)$ be OT plan. Define

$$\mu_t = ((Id - t)x + ty))_\# \pi \quad \Leftarrow \text{ McCann interpolation}$$

μ_t ... constant speed geodesic between μ and

Characterise μ_t

- Consider particles initially distributed according to ρ_0
- Move by a given velocity field v_t

Eulerian description $\rho(x,t)$ ... density of particles

$$\Rightarrow \partial_t \rho_t + \nabla(\rho_t v_t) = 0 \qquad \Leftarrow \text{ conservation of mass}$$
$$\rho_0 = \rho_0(x)$$

Dynamic formulation

Given $\mu, \nu \,\hat{=}$ initial & final distribution of particles at $t=0$ & $t=1$

If $\exists!\, v_t$ that is sufficiently smooth

$$\partial_t \rho + \nabla \cdot (\rho_t v_t) = 0$$
$$\rho_0 = \mu \quad \text{and} \quad \rho_1 = \nu$$

Benomou & Brenier : (1990)

$$\underbrace{A_p(\mu,\nu) = \int_0^1 \int \| v_t \|^p \, d\mu(x) \, dt}_{\text{energy loction function}}$$

Theorem

Let $X \subset \mathbb{R}^d$ be compact and consider two probability measures $\mu_i \in \mathscr{P}(X)$, $i = 1, 2$ with densities ϱ_i wr.t. the Lebesgue measure. Let $V(\varrho_0, \varrho_1)$ denote the set of all $(\varrho_t, v_t)_{t\in[0,1]}$ satisfying

- *The map $t \in [0,1] \to \rho_t$ is continuous in $\mathscr{P}(x)$ in the weak topology*
- *The continuity equation $\partial_t \varrho + \nabla \cdot (\varrho_t \cdot v_t) = 0$ holds in the weak sense for $v_t \in L^2(\mu_t)$ with initial and terminal conditions given by*

$$\varrho_{t=0} = \varrho_0 \text{ and } \varrho_{t=1} = \varrho_1 .$$

Then

$$\min_{\pi\in\Pi(\mu,\nu)} K(\pi) = \inf_{(\varrho,v)\in V(\varrho_0,\varrho_1)} \int_0^1 \int_\Omega |v_t|^2 \varrho_t(x) dx dt.$$

$\Rightarrow$ most prominent way do calculate OT plan til 2000'

$\Rightarrow$ Augmented Lagrange

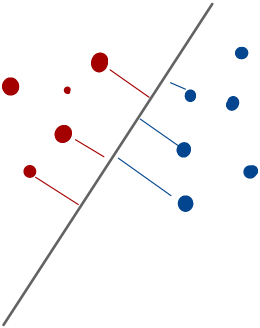
$$\inf \int_0^1 \int |v|^2 \rho \, dx \, dt$$

s.t. $\partial_t \rho + \nabla \cdot (\rho v) = 0 \quad \rho(x,0) = \rho_0, \; \rho(x,1) = \rho_1$

The sliced Wasserstein distance between two probability densities μ and ν in $\mathscr{P}(\mathbb{R}^d)$ is given by

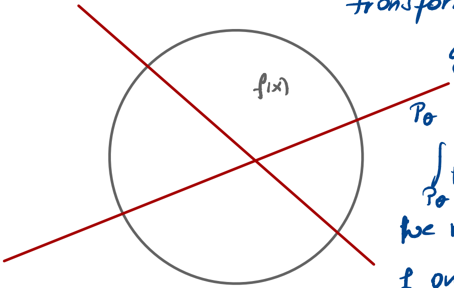
$$d_{SW_2}(\mu,\nu) := \left(\int_{S^d} d^2_{W_2}((P_\theta)_\#\mu,(P_\theta)_\#\nu)d\theta\right)^{\frac{1}{2}},$$

where $S^d = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ and $P_\theta : \mathbb{R}^d \to \mathbb{R}$ is the projection.



$\Rightarrow$ project on different lines

$\Rightarrow$ Connection to Radon transform



given P_θ

$\int_{P_\theta} f$ can

we reconstruct f on the whole domain

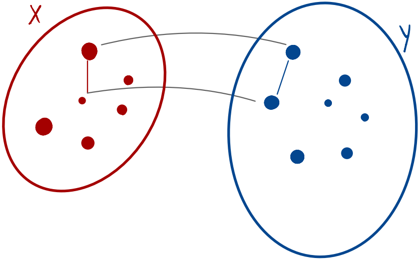
The Gromov-Wasserstein distance

Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \in \mathscr{P}(\mathbb{R}^p)$ and $\nu = \sum_{j=1}^{m} b_j \delta_{x_j} \in \mathscr{P}(\mathbb{R}^q)$ with $p \leq q$ with $\sum_{i=1}^{n} a_i = 1$ and $\sum_{j=1}^{m} b_j = 1$. Let $c_X \colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$, respectively $c_Y \colon \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^+$ denote the cost. Then the Wasserstein-Gromov distance is defined as

$$d^2_{GW_2}(c_X, c_Y, \mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} J(c_X, c_Y, \pi)$$

where

$$J(c_X, c_Y, \pi) = \sum_{i,j,k,l} |c_X(x_i, x_k) - c_Y(y_j, y_l)|^2 \pi_{ij}\pi_{kl}.$$



=) captures topological features

=) more costly.

Gradient flows

- Metric space: Consider a function $F : \mathbb{R}^d \to \mathbb{R}$ and a $x_0 \in \mathbb{R}^d$:

$$x'(t) = -\nabla F(x(t))$$
$$x(0) = x_0$$

Giorgi

⇒ Minimising movement scheme: let $\tau > 0$, define a sequence of points (x_k^τ) s.t.

$$x_{k+1}^\tau \in \operatorname{argmin} F(x) + \frac{|x - x_k^\tau|^2}{2\tau}.$$

$d(x, x_k^\tau)^2$

⇐ less regularity on F

- Wasserstein space The JKO version of the game ....

$$(*) \quad \varrho_{k+1}^\tau \in \operatorname{argmin}_\varrho F(\varrho) + \frac{d_{W_2}^2(\varrho, \varrho_{(k)}^\tau)}{2\tau}.$$

Jordan
Otto
Kinderlehrer
1998

Limiting PDE of (*)

1) Show that (*) has a unique solution
2) Optimality cond. (wrt ρ)

$$\frac{\delta F}{\delta \rho} + \frac{\varphi}{\tau} = \text{const}$$

φ... Kantorovich pot.

3) Transport map $T(x) = x - \nabla \varphi \Rightarrow \frac{T(x) - x}{\tau} = -\frac{\nabla \varphi}{\tau} = \nabla\left(\frac{\delta F}{\delta \rho}\right)$

$\frac{\text{displacement}}{\text{time}} \hat{=}$ velocity

4) As $\tau \to 0$ iterates converge

$$\partial_t \rho - \nabla \cdot \Big(\rho \underbrace{\nabla \Big(\frac{\delta F}{\delta \rho}\Big)}_{=v}\Big) = 0$$

Example:
$$F(\rho) = \int \rho \,(\log \rho - 1)\, dx$$
$$\frac{\delta F}{\delta \rho} = \log \rho - 1 + \rho \cdot \frac{1}{\rho} = \log \rho$$

$$\partial_t \rho = \operatorname{div}\Big(\cancel{\rho} \cdot \frac{1}{\cancel{\rho}} \nabla \rho\Big) = \Delta \rho$$

Fokker Planck, ...

given potential

$$\partial_t \rho = \nabla \cdot \big(\sigma(\rho) \nabla (E'(\rho) + V + W * \rho)\big)$$

non-linear mobility

internal energy

interaction energy

$\Rightarrow$ entropy methods

Connection to data science

- Ensemble / particle methods   SDEs, density of the mean limit satisfies nonlinear FPE

  ⇒ large time behavior $N \to \infty$ # particles

  $t \to \infty$ time

  equilibration behavior aka quasi-invariant Limit

  ⇒ existence, structure of s.

  ⇒ speed to equilibrium

- Generalise this for different costs.

  ⇒ Franca

  ⇒ Andrew Duncan   Stein Wasserstein GV

Entropic regularisation

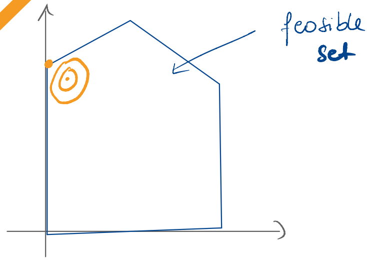
The regularised OT problem reads as

$$\min\left(\sum_{ij} c_{ij}P_{ij} + \overbrace{\varepsilon P_{ij}\log P_{ij}}^{\text{entropic regularisation}}\right), \qquad (OT)_\varepsilon$$

among all admissible transportation plans and for a given $\varepsilon > 0$. The second term corresponds to the variational derivative of the negative entropy

objective

$$H(P) := -\sum_{ij} P_{ij}\left(\log(P_{ij}) - 1\right).$$



◦) $\varepsilon = 0$ minimisers lie on boundary of feasible set

◦) $\varepsilon > 0$ moves the min to interior

$\Rightarrow$ unique

$\Rightarrow$ solution to $(OT)_\varepsilon$ blurred

$(OT)_\varepsilon$ $\Rightarrow$ $\exists!$ unique solution P

$\Rightarrow$ Solutions of $(OT)_\varepsilon$ converge to sol of (OT)

but NO convergence rates

P_ε P

$\Rightarrow$ Convergence of $(OT)_\varepsilon$ deteriorates as $\varepsilon \to 0$

$\Rightarrow$ Solve problem extremely efficiently !

$\Rightarrow$ Sinkhorn - Knopp

Sinkhorn

Given matrix A with pos. entries, can we find D_1 and D_2 s.t $D_1 A D_2$ is doubly stochastic

Idea:

- $\rightarrow$ Iterative prop. fitting (IFPP)
- $\rightarrow$ RAS method
- $\rightarrow$ matrix scaling

Kullback-Leibler divergence between discrete measures

$$KL(P, K) := \sum_{ij} f\left(\frac{P_{ij}}{K_{ij}}\right) K_{ij} \text{ for } f(t) = \begin{cases} t \log t & \text{if } t \geq 0 \\ +\infty & \text{if } t < 0. \end{cases}$$

Gibb's kernel

$K = e^{-C/\varepsilon}$

$(OT)_\varepsilon$ is equivalent to.

$$\min \ \varepsilon \, KL(P|K)$$

$$(OT)_\varepsilon \quad \sum C_{ij} P_{ij} + \varepsilon P_{ij} \log P_{ij}$$

$$KL(P|K) = \frac{P_{ij}}{\cancel{K_{ij}}} \operatorname{Log}\left(\frac{P_{ij}}{K_{ij}}\right) \cancel{K_{ij}} = P_{ij} \log P_{ij} + \frac{1}{\varepsilon} P_{ij} C_{ij}$$

Proposition

The regularised OT problem has a unique solution $P \in \mathbb{R}^{n \times m}$ of the form

$$P_{ij} = u_i K_{ij} v_j,$$

where $u \in \mathbb{R}^n_+$ and $v \in \mathbb{R}^m_+$ are two (unknown) scaling vectors. K .. Gibbs kernel

Dual variable: $\varphi \in \mathbb{R}^n$, $\psi \in \mathbb{R}^m$

$$\mathcal{L}(P, \varphi, \psi) = \langle P, C \rangle - \varepsilon H(P) - \langle \varphi, P \mathbb{1}_m - \mu \rangle$$
$$\langle \psi, P^T \mathbb{1}_n - \nu \rangle$$

OC:

$$\frac{\partial \mathcal{L}}{\partial P_{ij}} = C_{ij} - \varepsilon \log P_{ij} - \varphi_i - \psi_j = 0$$

$$P_{ij} = \underbrace{e^{-\varphi_i/\varepsilon}}_{=: u} \; \underbrace{e^{-C_{ij}/\varepsilon}}_{:= K} \; \underbrace{e^{-\psi_i/\varepsilon}}_{:= v}$$

Sinkhorn algorithm: Let $\mu \in \mathbb{R}^n$ and $\nu \in \mathbb{R}^m$, initialise v^0 and set $\varepsilon > 0$.

- Update $u^{(k+1)} := \frac{\mu}{Kv^k}$
- Update $v^{(k+1)} := \frac{\nu}{K^T u^{(k)}}$.
- Go to first bullet until convergence

Calculate the corresponding OT plan

$$P = \operatorname{diag}(u)\, K \operatorname{diag}(v).$$

◦) (u,v) have to be chosen that constraints are satisfying

$$\operatorname{diag} u \; K \underbrace{\operatorname{dig} v \; \mathbb{1}_m}_{v} = \mu \quad \Rightarrow \quad u \odot (Kv) = \mu$$

element wise mult

$$\operatorname{diag} v \; K^T \underbrace{\operatorname{diag} u \; \mathbb{1}_n}_{u} = \nu$$

◦) $\varepsilon \to 0$ $\quad K = e^{-c/\varepsilon}$ $\Leftarrow$ problematic $\Rightarrow$ log scaling

Schmitzer...

Rapid developments: ◦) Different reg. terms
◦) Connection to proximal alg.

References

- L. Ambrosio, N. Gigli, G. Savare, *Gradient flows: in metric spaces and the space of probability measures*, Springer, 2008
- M. Liero, A. Mielke and G. Savare, *Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures*, Inventiones 211, 2018
- L. Chizat, G. Peyre, B. Schmitzer and F.X. Vialard, *Unbalanced optimal transport: geometry and Kantorovich formulation*, Mathematics of Computation 87, 2018
- R. Jordan, D. Kinderlehrer and F. Otto, *Variational formulation of the Fokker-Planck equation*,SIAM Math Anal, 1998
- J.D Benamou and J. Brenier, *A computational fluid mechanics solutionto the Monge-Kantovich mass transfer problem*, Numer. Math, 2000.
- M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, NIPS 2013
- M. Idel, *A review of matrix scaling algorithms and Sinkhorn's normal for for matrices and positive maps*, 2016.