Metric on $\mathcal{P}(\mathbb{R}^d)$

Important reminder / result:

__Theo__ : $\mathcal{P}(\mathbb{R}^d)$ is a Polish space.

endowed with the weak convergence topology

i.e; • $\exists$ a distance / metric such that

$(\mathcal{P}(\mathbb{R}^d), dist)$ is a complete metric space
separable

• $(\mu_n) \Rightarrow \mu \iff dist(\mu_n, \mu) \xrightarrow[n \to +\infty]{} 0$

__example__ : dist not unique

• Levy-Prokhorov metric

$$d_{LP}(\mu, \nu) = \inf \{ r > 0 : \mu(F) \leq \nu(F + B(0,r)) + r \quad \forall F \subset \mathbb{R}^d \text{ closed} \}.$$

• Fortet–Mourier metric

$d_{FM}(\mu, \nu) = \sup \{ \mu(f) - \nu(f) : f$ bounded Lip $\}$.

Not really practical : hard to evaluate / estimate.

Another "more practical" option

The Wasserstein distance.

It is defined through coupling / transference plan.

γ is a coupling between μ and ν if

• $\gamma \in \mathcal{P}(\mathbb{R}^{2d})$

• $\gamma(\mathbb{R}^d \times \cdot) = \mu(\cdot)$

• $\gamma(\cdot \times \mathbb{R}^d) = \nu(\cdot)$.

The Wasserstein distance of order $p \geq 1$ on $\mathbb{R}^d$ is defined as

$$W_p^p(\mu, \nu) = \inf \left\{ \int \|x - y\|^p \, d\gamma(x,y) : \gamma \text{ is a coupling of } \mu, \nu \right\}.$$

In fact, we may change $(x,y) \longmapsto |x-y|$
by any distance ρ on $\mathbb{R}^d$.

$$W_\rho(\mu,\nu) = \inf \left\{ \int \rho(x,y)\, d\gamma(x,y) : \gamma \text{ coupling of } \mu, \nu \right\}.$$

In particular for $\rho(x,y) = \mathbb{1}_{x \neq y}$
we obtain the total variation distance

Question : How to compute W_ρ ?

In general not possible but possible
for $\mathbb{R}$.

Recall that if X r.v on $\mathbb{R}$
with distribution δ_X
$F^{\leftarrow}(U)$ has the same distribution
as X with $U \sim \text{Unif}[0,1]$

$$F^{\leftarrow}(u) = \inf\{x : F(x) \geq u\}$$

Therefore for two r.v X et Y with distribution μ, ν respect

$(F_X^{\leftarrow}(U_1), F_Y^{\leftarrow}(U_2))$ coupling of μ, ν

if $U_1, U_2 \sim \text{Unif}\ [0,1]$

$(F^{\leftarrow}(U), F^{\leftarrow}(U))$, is in fact

an optimal coupling for W_1.

Theo :

$$W_1(\mu,\nu) = \int_0^1 | F_\mu^{\leftarrow}(u) - F_\nu^{\leftarrow}(u)| \, du$$

As a result, we may estimate $W_1(\mu,\nu)$

by a MC procedure.

$$W_1(\mu,\nu) \approx \frac{1}{M} \sum_{i=1}^{M} |F_\mu^{\leftarrow}(U_i) - F_\nu^{\leftarrow}(U_i)|$$

$(U_i) \overset{iid}{\sim} \text{Unif}\ [0,1]$.

For $\mathbb{R}^d$, we may consider the SW_1

Def :

$$SW_1(\mu, \nu) = \int_{\mathbb{S}^{d-1}} W_1(\mathrm{proj}_v \# \mu, \mathrm{proj}_v \# \nu) \, d\sigma(v)$$

where $\mathbb{S}^{d-1} = \{ v \in \mathbb{R}^d : \|v\| = 1 \}$

σ unif distrib° on $\mathbb{S}^{d-1}$

$\mathrm{proj}_v \, . \, x \longmapsto \langle x, v \rangle$.

It can be shown that SW_1 is a metric on $\mathbb{R}^d$

Monte Carlo estimation : double MC ...

$$SW_1(\mu, \nu) \simeq \frac{1}{M_1} \sum_{i=1}^{M_1} \hat{W}_1(\mu_i, \nu_i)$$

(V_i) iid $\mathrm{Unif}(\mathbb{S}^{d-1})$ $\left(G_i / \|G_i\| \right)$

$$\hat{W}_1(\mu_i, \nu_i) = \frac{1}{M_2} \sum |F^{\leftarrow}_{\mu_i}(U_j)$$

$- F_{\nu_n}(U_j)|$

$\mu_n = \text{proj}_{v_n} \# \mu \qquad \nu_n = \text{proj}_{v_n} \# \nu .$

$(U_j)_j$ iid $\text{Unif}(0,1)$.

Otherwise how to proceed ?

We can consider empirical version of

$$\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}$$

$$\hat{\nu}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i}$$

! random quantity.

And approximate $W_p^p(\mu,\nu)$ by $W_p^p(\hat{\mu}_m, \hat{\nu}_m)$

Case $\underline{p=2}$

Also we can consider a bit more general pb

$$\tilde{\mu}_x = \sum_{i=1}^{m} \varpi_i^{\mu} \delta_{x_i} \qquad \tilde{\nu}_x = \sum_{i=1}^{m} \varpi_i^{\nu} \delta_{y_i}$$

$$W_2^2(\tilde{\mu}_m, \tilde{\nu}_m) = \inf_{\gamma} \sum_{ij} \gamma(\{x_i, y_j\}) \| x_i - y_j \|^2$$

$$\sum_j \gamma(\{x_i, y_j\}) = \varpi_i^{\mu}$$

$$\sum_i \gamma(-) = \varpi_j^{\nu}$$

then this problem is eq to

(*) maximize $x M y$ over $M \in \mathbb{R}^{d \times d}$

where M is the matrix

$$M_{ij} = \gamma(\{x_i, y_j\})$$

(C_1) so should satisfies $\sum M_{ij}$ $\sum M_{ij}$

and $x = (x_1 - x_m)^T$

$y = (y_1 - y_m)^T$.

Pb (*) linear programming pb

ref

$\exists$ algorithm to solve it but complexity $O(n^3)$

It has been suggested to relax the problem.

maximize $xMy + \varepsilon \psi(M)$

over M satisfying (E_1)

where ψ strictly convex and $\varepsilon > 0$

$\Rightarrow$ Sinkhorn algorithm.

ψ is a divergence notion that we now present.

Divergence.

While we would like to only consider metric since they formalize the concept of being closed.

As we saw, they can be difficult to estimate while we need them to evaluate $\theta \longmapsto D(\mu_\theta, \mu^*)$

We consider here $\int$ $D : \mathcal{P}(\mathbb{R}^d)^2 \to \mathbb{R}_+$ which do not satisfy most axiom of a metric:

- symmetric
- triangle ineq

Still, while they may induce "strange"

geometry and their use has to be made with some care, they will satisfy

$$D(\mu, \nu) = 0 \iff \mu = \nu.$$

We provide here two examples of divergence: the KL and the Fisher.

R The KL.

We define the KL divergence

$$KL(\nu | \mu) = \int \log\left(\frac{d\nu}{d\mu}\right) d\nu.$$

$= \infty$ if $\nu \not\ll \mu$.

- $-\log$ is strictly convex

therefore $KL(\nu | \mu) \geq 0$ with $=$ if $\mu = \nu$.

More prop on Fisher session.

The Fisher divergence.

$$I(\nu \mid \mu) = \int \left\| \nabla \log \frac{d\nu}{d\lambda} - \nabla \log \frac{d\mu}{d\lambda} \right\|^2 d\nu$$

$$= +\infty \quad \text{if } \nu \neq \mu$$