Continuous transformations of probability measures and their transport representations

Hugo Lavenant

Bocconi University

Workshop “Mathematical Foundations of Artificial Intelligence”
UNAM, Mexico City (Mexico), April 21, 2026

Joint work with:

Giuseppe Savaré

Based on a question of:

Gabriel Peyré

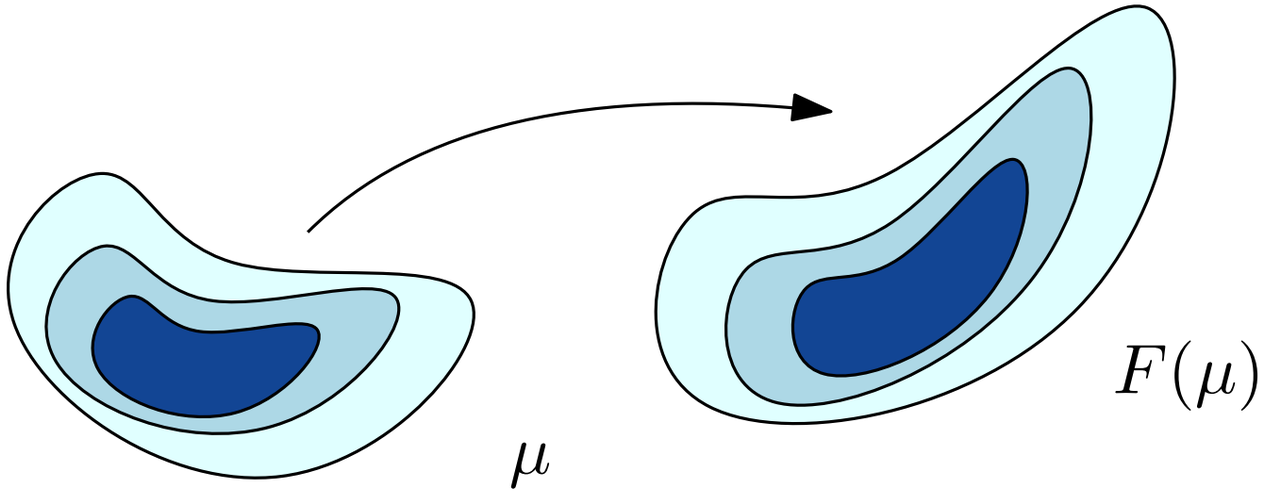
[On arxiv since this morning]

The question

Input.

$$F : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$$

Can be generalized to $\mathcal{P}(X) \to \mathcal{P}(Y)$ with X, Y Polish space and X geodesic.

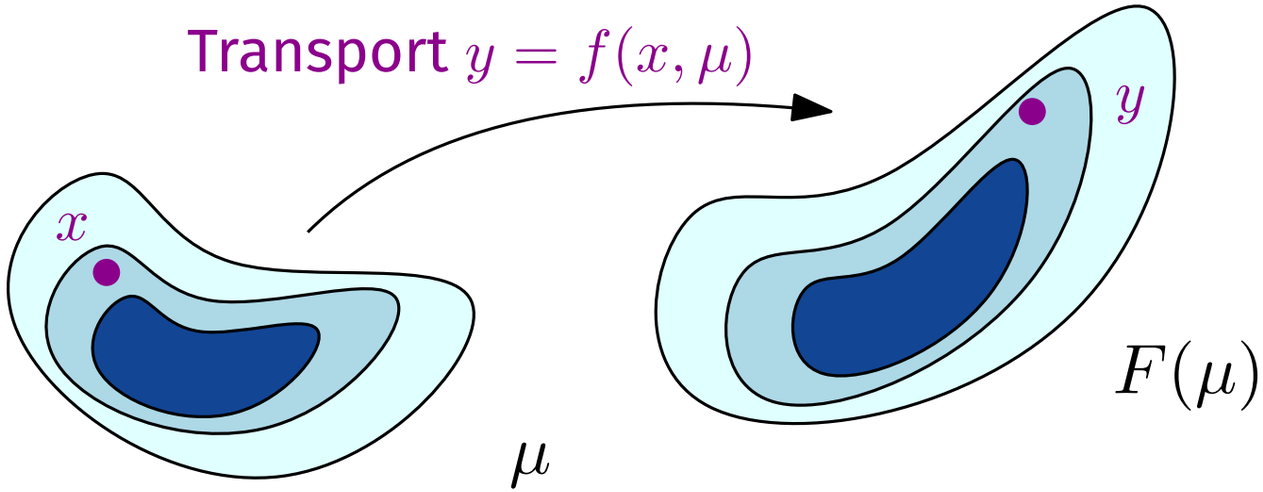


The question

Input.

$$F : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$$

Can be generalized to $\mathcal{P}(X) \to \mathcal{P}(Y)$ with X, Y Polish space and X geodesic.



Is there $f : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ such that, for all μ,

$$F(\mu) = f(\cdot, \mu)_{\#}\mu$$

$g_{\#}\mu$ image measure of μ by map $g : \mathbb{R}^d \to \mathbb{R}^d$

and can f be chosen continuous if F is continuous?

Why? approximation of maps by transformers

An (encoder) transformer can be seen as a map $F : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ with

$$F(\mu) = f(\cdot, \mu)_{\#}\mu$$

$x \in \mathbb{R}^d$ token

$\mu = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}$ collection of tokens.

$f = f(x, \mu)$ to describe the attention mechanism

Useful for the many tokens limit $m \to +\infty$.

[Vuckovic et al, 2020], [Sander et al, 2022], [Geshkovski et al, 2025], etc.

Why? approximation of maps by transformers

An (encoder) transformer can be seen as a map $F : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ with

$$F(\mu) = f(\cdot, \mu)_{\#}\mu$$

$x \in \mathbb{R}^d$ token

$\mu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ collection of tokens.

$f = f(x, \mu)$ to describe the attention mechanism

Useful for the many tokens limit $m \to +\infty$.

Result. If $(\mu_j, \nu_j)_j$ finite collection, under assumption we can find transformer with

$$F(\mu_j) \simeq \nu_j \text{ for all } j.$$

[Geshkovski, Rigollet & Ruiz-Balet, 2024]

Result. If Ω compact and $f : \Omega \times \mathcal{P}(\Omega) \to \mathbb{R}^d$ jointly continuous, it can be approximated by transformer architecture.

[Furuya, de Hoop & Peyré, 2024]

Why? approximation of maps by transformers

An (encoder) transformer can be seen as a map $F : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ with

$$F(\mu) = f(\cdot, \mu)_{\#}\mu$$

When does a transformation F decomposes like this in the first place?

$\mu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ collection of tokens.

$f = f(x, \mu)$ to describe the attention mechanism

Useful for the many tokens limit $m \to +\infty$.

Result. If $(\mu_j, \nu_j)_j$ finite collection, under assumption we can find transformer with

$$F(\mu_j) \simeq \nu_j \text{ for all } j.$$

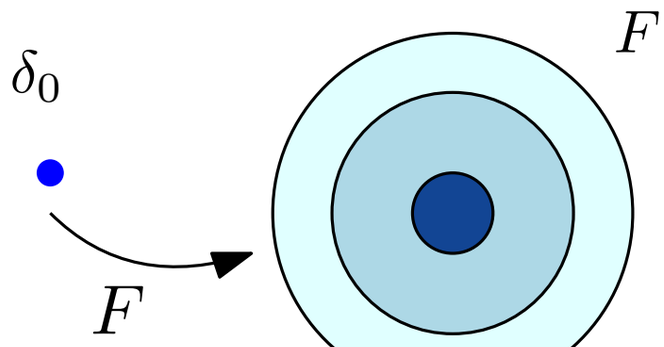
[Geshkovski, Rigollet & Ruiz-Balet, 2024]

Result. If Ω compact and $f : \Omega \times \mathcal{P}(\Omega) \to \mathbb{R}^d$ jointly continuous, it can be approximated by transformer architecture.

[Furuya, de Hoop & Peyré, 2024]

Obstruction and the non splitting assumption

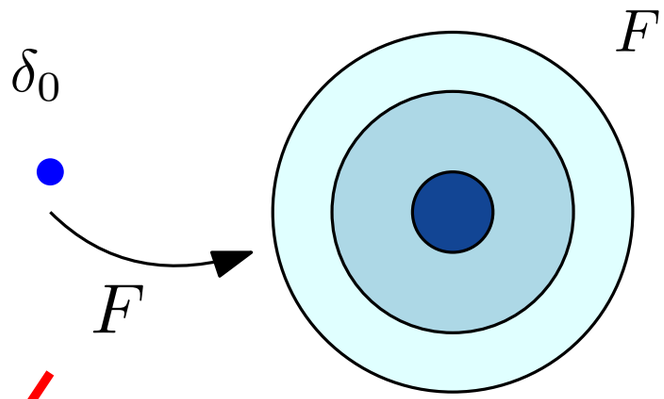
$F(\mu) = \mu * \gamma$, and γ has a density.



No transport representation.

Obstruction and the non splitting assumption

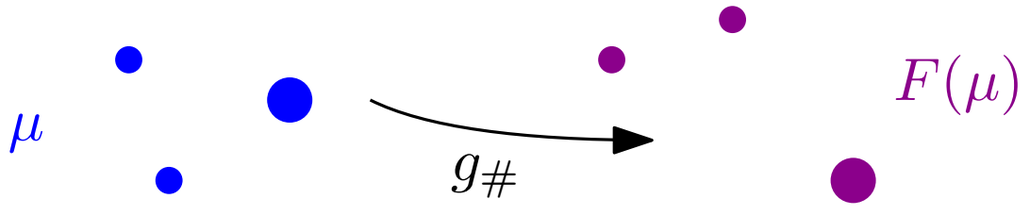
$F(\mu) = \mu * \gamma$, and γ has a density.



No transport representation.

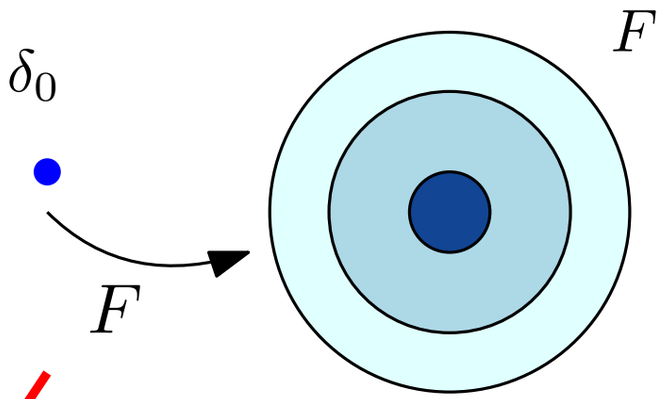
Assumption (non-splitting on empirical measures). For $\mu = \sum_{i=1}^{m} a_i \delta_{x_i}$ with a_i rational, there exists $g : \mathbb{R}^d \to \mathbb{R}$ with

$$g_{\#}\mu = F(\mu).$$



Obstruction and the non splitting assumption

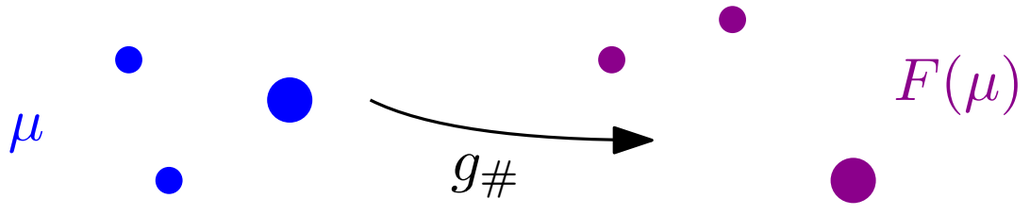
$F(\mu) = \mu * \gamma$, and γ has a density.



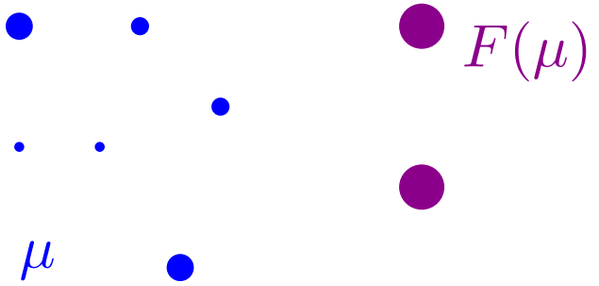
No transport representation.

Assumption (non-splitting on empirical measures). For $\mu = \sum_{i=1}^m a_i \delta_{x_i}$ with a_i rational, there exists $g : \mathbb{R}^d \to \mathbb{R}$ with

$$g_{\#}\mu = F(\mu).$$



Remark. Take $\mu = \sum_{i=1}^m a_i \delta_{x_i}$ and $F(\mu) = \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$.



$F(\mu) = g_{\#}\mu$ if and only if there exists I_1, I_2 partition of $\{1, \dots, m\}$ with $\sum_{i \in I_1} a_i = \sum_{i \in I_2} a_i$: **partition problem** (NP hard).

Continuous F

$\mathcal{P}(\mathbb{R}^d)$ endowed with narrow (a.k.a. weak) topology.

Theorem. If $F : \mathcal{P}(\mathbb{R}) \to \mathcal{P}(\mathbb{R})$ continuous and non-splitting on empirical measures, there exists $f : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ **measurable** such that

$$F(\mu) = f(\cdot, \mu)_{\#}\mu \text{ for all } \mu.$$

Continuous F

$\mathcal{P}(\mathbb{R}^d)$ endowed with narrow (a.k.a. weak) topology.

Theorem. If $F : \mathcal{P}(\mathbb{R}) \to \mathcal{P}(\mathbb{R})$ continuous and non-splitting on empirical measures, there exists $f : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ **measurable** such that

$$F(\mu) = f(\cdot, \mu)_{\#}\mu \text{ for all } \mu.$$

But f cannot be chosen to be continuous

Example of discontinuous f but continuous F

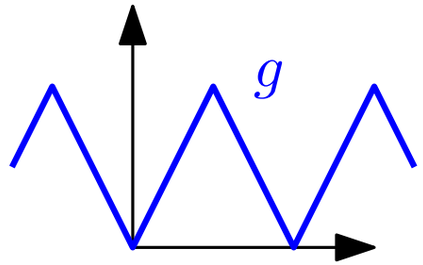
Define for $x \in [0,1]$ and $\mu \in \mathcal{P}([0,1])$

$$f(x,\mu) = g\left(\frac{x}{W(\mu,\lambda)^{1/2}}\right),$$

and

$$F(\mu) = f(\cdot,\mu)_{\#}\mu.$$

g 1-periodic and Lipschitz function with $g_{\#}\lambda = \lambda$



λ Lebesgue measure on $[0,1]$

$W(\mu,\lambda)$ Wasserstein distance between μ and λ

Example of discontinuous f but continuous F

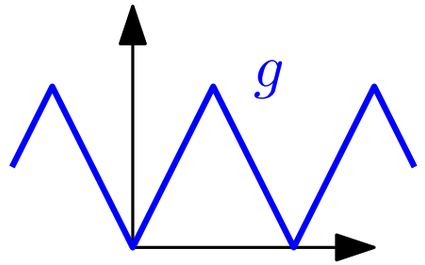
Define for $x \in [0,1]$ and $\mu \in \mathcal{P}([0,1])$

$$f(x,\mu) = g\left(\frac{x}{W(\mu,\lambda)^{1/2}}\right),$$

and

$$F(\mu) = f(\cdot,\mu)_{\#}\mu.$$

g 1-periodic and Lipschitz function with $g_{\#}\lambda = \lambda$



λ Lebesgue measure on $[0,1]$

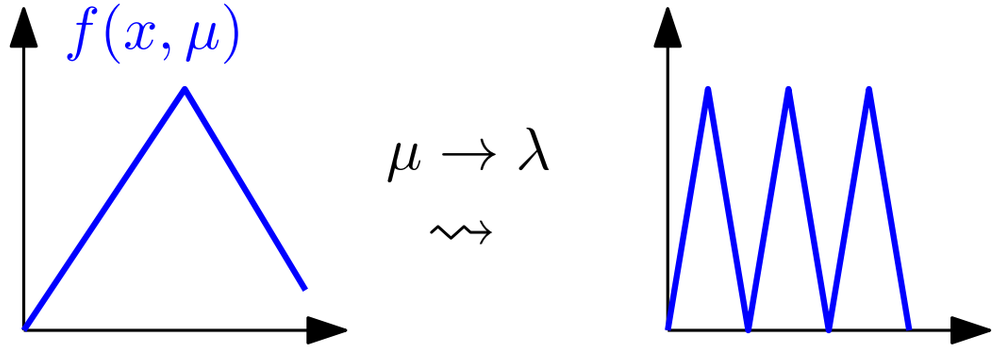
$W(\mu,\lambda)$ Wasserstein distance between μ and λ

1. f and F are continuous for $\mu \neq \lambda$.

2. As $\mu \to \lambda$, $f(\cdot,\mu)$ diverges but

$$W(F(\mu),\lambda) \lesssim W(\mu,\lambda)^{1/2}$$

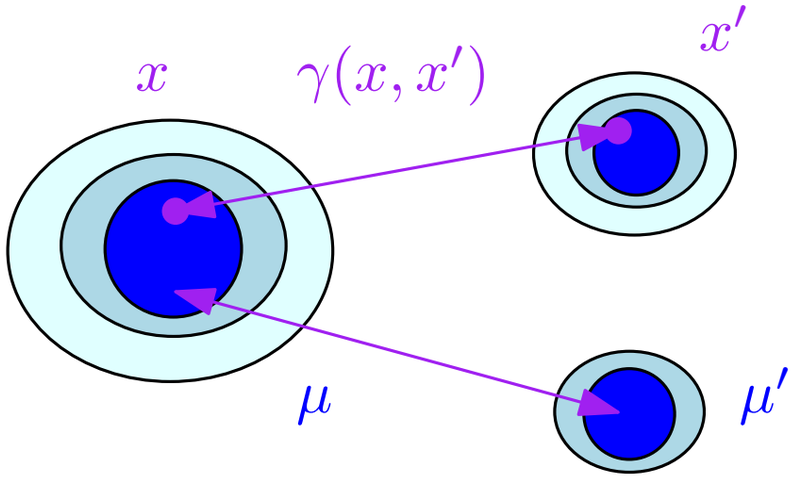
so F continuous with $F(\lambda) = \lambda$



Lipschitz F

Recall. $W_p(\mu,\mu')^p = \min_\gamma \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-x'\|^p \mathrm{d}\gamma(x,x') \right.$
$\left. \gamma \text{ has marginals } \mu,\mu' \right\}$

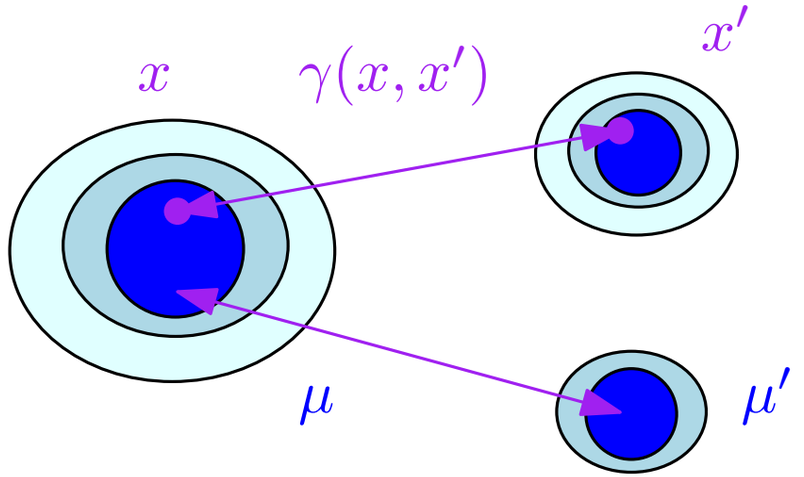
W_p distance on $\mathcal{P}_p(\mathbb{R}^d)$ measures with finite p moments.



Lipschitz F

Recall. $W_p(\mu,\mu')^p = \min_\gamma \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-x'\|^p \mathrm{d}\gamma(x,x') \right.$
$\left. \gamma \text{ has marginals } \mu,\mu' \right\}$

W_p distance on $\mathcal{P}_p(\mathbb{R}^d)$ measures with finite p moments.



Theorem. Assume $F : \mathcal{P}_p(\mathbb{R}^d) \to \mathcal{P}_p(\mathbb{R}^d)$ is non-splitting on empirical measures and **Lipschitz**.

Then there exists $f : \mathbb{R}^d \times \mathcal{P}_p(\mathbb{R}^d) \to \mathbb{R}^d$ such that $f(\cdot,\mu)_\# \mu = F(\mu)$ and

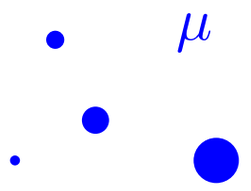
$$x_n \to x,\ \mu_n \to \mu, \quad x_n \in \mathrm{supp}(\mu_n), \quad x \in \mathrm{supp}(\mu) \quad \Rightarrow \quad f(x_n,\mu_n) \to f(x,\mu)$$

cannot be removed

Element of proof 1): generic measures

Definition. A measure $\mu = \sum_{i=1}^{m} a_i \delta_{x_i}$ is generic if:

$$I_1, I_2 \subseteq \{1, \dots, m\}, \quad \sum_{i \in I_1} a_i = \sum_{i \in I_2} a_i \quad \Rightarrow \quad I_1 = I_2$$

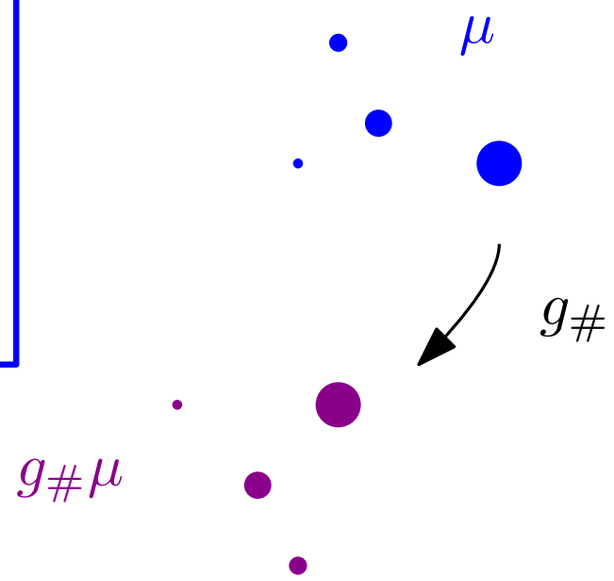


Element of proof 1): generic measures

Definition. A measure $\mu = \sum_{i=1}^{m} a_i \delta_{x_i}$ is generic if:

$$I_1, I_2 \subseteq \{1, \dots, m\}, \quad \sum_{i \in I_1} a_i = \sum_{i \in I_2} a_i \quad \Rightarrow \quad I_1 = I_2$$

Lemma. If μ is generic and $g_\# \mu = g'_\# \mu$ then $g = g'$ on $\mathrm{supp}(\mu)$.

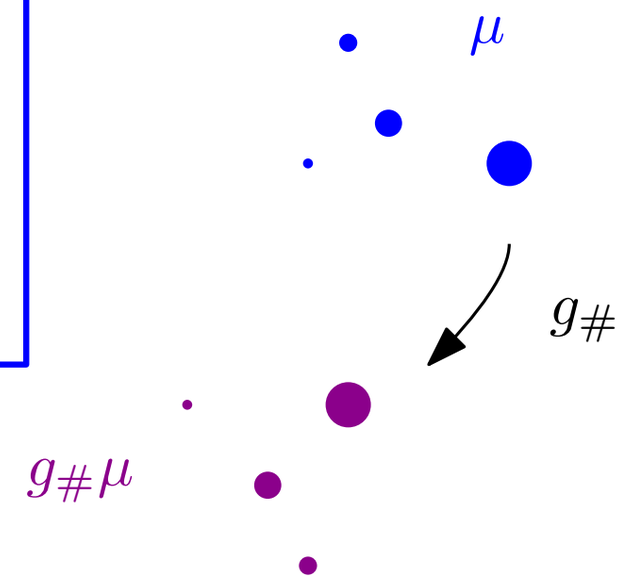


Element of proof 1): generic measures

Definition. A measure $\mu = \sum_{i=1}^{m} a_i \delta_{x_i}$ is generic if:

$$I_1, I_2 \subseteq \{1, \dots, m\}, \quad \sum_{i \in I_1} a_i = \sum_{i \in I_2} a_i \quad \Rightarrow \quad I_1 = I_2$$

Lemma. If μ is generic and $g_{\#}\mu = g'_{\#}\mu$ then $g = g'$ on $\mathrm{supp}(\mu)$.



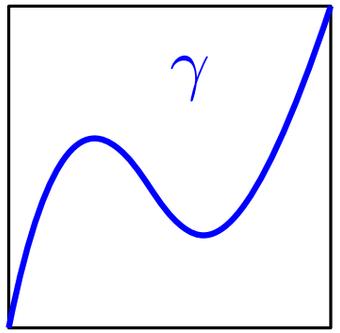
generic measures dense in $\mathcal{P}_p(\mathbb{R}^d)$

Consequence: $f(\cdot, \mu)$ uniquely defined for μ generic measure. There exists **at most one** f globally defined and continuous in μ.

Element of proof 2): the space $\mathrm{TL}_p(\mathbb{R}^d;\mathbb{R}^d)$

$$(\mu, f(\cdot,\mu)) \quad\longrightarrow\quad \gamma = (\mathrm{id}\times f(\cdot,\mu))_{\#}\mu$$

$\mu \in \mathcal{P}_p(\mathbb{R}^d)$ $\qquad$ $f(\cdot,\mu) \in L^p_\mu$ $\qquad$ $\gamma \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$

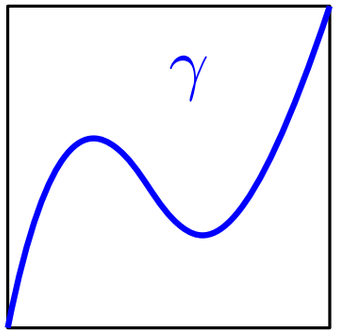


[García Trillos & Slepčev, 2016, 2018], [Thorpe et al, 2017].

Element of proof 2): the space $\mathrm{TL}_p(\mathbb{R}^d;\mathbb{R}^d)$

$(\mu, f(\cdot,\mu)) \quad \rightarrow \quad \gamma = (\mathrm{id} \times f(\cdot,\mu))_{\#}\mu$

$\mu \in \mathcal{P}_p(\mathbb{R}^d) \qquad f(\cdot,\mu) \in L^p_\mu \qquad \gamma \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$



Distance on the space TL_p of pairs $(\mu, f(\cdot,\mu))$.

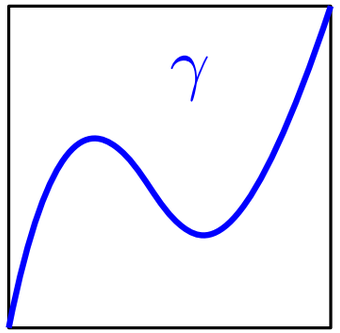
$\leftarrow \quad W_p$ on $\mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$

[García Trillos & Slepčev, 2016, 2018], [Thorpe et al, 2017].

Element of proof 2): the space $\mathrm{TL}_p(\mathbb{R}^d;\mathbb{R}^d)$

$(\mu, f(\cdot,\mu)) \quad \rightarrow \quad \gamma = (\mathrm{id} \times f(\cdot,\mu))_{\#}\mu$

$\mu \in \mathcal{P}_p(\mathbb{R}^d) \qquad f(\cdot,\mu) \in L^p_\mu \qquad \gamma \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$



Distance on the space TL_p of pairs $(\mu, f(\cdot,\mu))$.

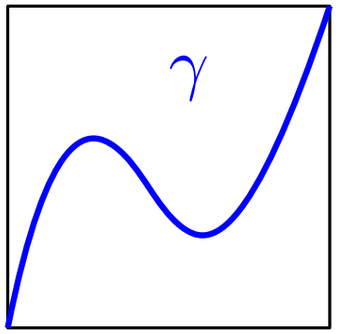
$\leftarrow \quad W_p$ on $\mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$

1. If F L-Lipschitz, then $\mu \mapsto (\mu, f(\cdot,\mu))$ uniquely defined on generic measures, and $(1+L^p)^{1/p}$ Lipschitz with respect to $\mathrm{TL}_p(\mathbb{R};\mathbb{R}^d)$.

Element of proof 2): the space $\mathrm{TL}_p(\mathbb{R}^d;\mathbb{R}^d)$

$(\mu, f(\cdot,\mu)) \quad \rightarrow \quad \gamma = (\mathrm{id} \times f(\cdot,\mu))_{\#}\mu$

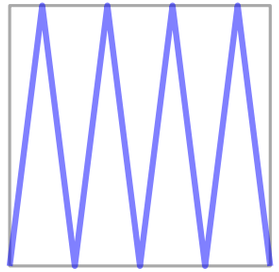
$\mu \in \mathcal{P}_p(\mathbb{R}^d)$ $\quad$ $f(\cdot,\mu) \in L^p_\mu$ $\quad$ $\gamma \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$



Distance on the space TL_p of pairs $(\mu, f(\cdot,\mu))$.

$\leftarrow$ W_p on $\mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$

not complete



1. If F L-Lipschitz, then $\mu \mapsto (\mu, f(\cdot,\mu))$ uniquely defined on generic measures, and $(1+L^p)^{1/p}$ Lipschitz with respect to $\mathrm{TL}_p(\mathbb{R};\mathbb{R}^d)$.

2. The space TL_p is not complete. We prove that $f(\cdot,\mu)$ is L-Lipschitz on $\mathrm{supp}(\mu)$.

Concluding remark

e.g. $\mathcal{A}$ class of transformers

Corollary. For Ω compact, assume $\mathcal{A}$ class of functions $\Omega \times \mathcal{P}_p(\Omega) \to \mathbb{R}^d$ which can approximate uniformly any continuous function.

Then if $F : \mathcal{P}(\Omega) \to \mathcal{P}_p(\mathbb{R}^d)$ is non-splitting on empirical measures and Lipschitz, for any $\varepsilon > 0$ there exists $g \in \mathcal{A}$ such that, with $F_g(\mu) = g(\cdot, \mu)_{\#}\mu$

$$\sup_{\mu} W_p(F(\mu), F_g(\mu)) \leq \varepsilon.$$

Concluding remark

e.g. $\mathcal{A}$ class of transformers

Corollary. For Ω compact, assume $\mathcal{A}$ class of functions $\Omega \times \mathcal{P}_p(\Omega) \to \mathbb{R}^d$ which can approximate uniformly any continuous function.

Then if $F : \mathcal{P}(\Omega) \to \mathcal{P}_p(\mathbb{R}^d)$ is non-splitting on empirical measures and Lipschitz, for any $\varepsilon > 0$ there exists $g \in \mathcal{A}$ such that, with $F_g(\mu) = g(\cdot, \mu)_{\#}\mu$

$$\sup_{\mu} W_p(F(\mu), F_g(\mu)) \leq \varepsilon.$$

Thank you for your attention