Geometría de espacios de diagramas de persistencia

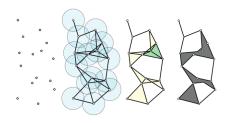
Mauricio Che

Durham University

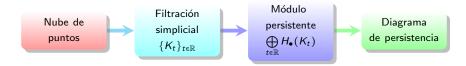
Seminario de TDA Universidad Autónoma de Madrid 12 de Febrero de 2024 Parte de estas diapositivas está basada en proyectos en conjunto con Galaz-García, Guijarro, Membrillo Solis y Valiunas.

TDA en 5 minutos

Topological Data Analysis (TDA):



Tubería del TDA:



• La idea de este proceso es obtener invariantes que nos permitan reconocer propiedades topológicas del conjunto de datos y su relevancia (persistencia).

 La idea de este proceso es obtener invariantes que nos permitan reconocer propiedades topológicas del conjunto de datos y su relevancia (persistencia).

 Los diagramas de persistencia son objetos geométricos que podemos comparar entre sí y, como resultado, nos permiten discernir entre conjuntos de datos. La idea de este proceso es obtener invariantes que nos permitan reconocer propiedades topológicas del conjunto de datos y su relevancia (persistencia).

 Los diagramas de persistencia son objetos geométricos que podemos comparar entre sí y, como resultado, nos permiten discernir entre conjuntos de datos.

• Esta comparación se hace a través de métricas/funciones distancia entre diagramas de persistencia.

Comparando diagramas de persistencia

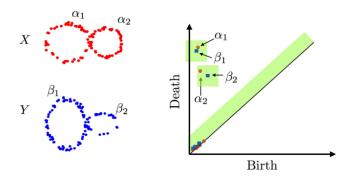


Figura: Diagramas de persistencia de nubes de puntos que presentan una 'forma' similar.

• Cada diagrama de persistencia es un conjunto de puntos $\{(a_i,b_i)\}_{i\in I}\subset\mathbb{R}^2_\geq:=\{(x,y)\in\mathbb{R}^2:x\geq y\}$, donde cada (a_i,b_i) representa una clase de homología que 'nace' en tiempo a_i y 'muere' en tiempo b_i .

- Cada diagrama de persistencia es un conjunto de puntos
 {(a_i, b_i)}_{i∈I} ⊂ ℝ²_≥ := {(x, y) ∈ ℝ² : x ≥ y}, donde cada (a_i, b_i)
 representa una clase de homología que 'nace' en tiempo a_i y 'muere'
 en tiempo b_i.
- Una métrica razonable entre diagramas de persistencia debería indicar que dos diagramas son cercanos cuando podemos sobreponer uno en el otro sin mucho esfuerzo.

- Cada diagrama de persistencia es un conjunto de puntos $\{(a_i,b_i)\}_{i\in I}\subset\mathbb{R}^2_{\geq}:=\{(x,y)\in\mathbb{R}^2:x\geq y\}$, donde cada (a_i,b_i) representa una clase de homología que 'nace' en tiempo a_i y 'muere' en tiempo b_i .
- Una métrica razonable entre diagramas de persistencia debería indicar que dos diagramas son cercanos cuando podemos sobreponer uno en el otro sin mucho esfuerzo.

• Además, los puntos que están cerca de la diagonal $\Delta = \{(x,y) \in \mathbb{R}^2 : x = y\}$ representan clases de homología que persisten por muy poco tiempo. Queremos que nuestra métrica no se vea demasiado afectada por dichos puntos.

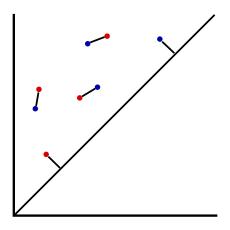


Figura: Biyección entre diagramas

Espacio de diagramas de persistencia sobre $(\mathbb{R}^2_{\geq}, \Delta)$

 $\mathcal{D}(\mathbb{R}^2_{\geq}, \Delta) \coloneqq \{\sigma : \sigma \in \mathbb{R}^2_{\geq} \text{ es un multiconjunto contable}\} / \sim$

donde $\sigma \sim \tau$ si σ y τ difieren únicamente por puntos en Δ .

Espacio de diagramas de persistencia sobre $(\mathbb{R}^2_{\geq}, \Delta)$

 $\mathcal{D}(\mathbb{R}^2_\geq,\Delta) \coloneqq \{\sigma: \sigma \in \mathbb{R}^2_\geq \text{ es un multiconjunto contable}\}/\sim$

donde $\sigma \sim \tau$ si σ y τ differen únicamente por puntos en Δ .

p-distancias entre diagramas de persistencia

Para un $p \in [1, \infty]$ fijo y $\sigma, \tau \in \mathcal{D}(\mathbb{R}^2_{\geq}, \Delta)$, definimos

$$d_{p}(\sigma,\tau) = \begin{cases} \inf_{\phi} \left(\sum_{x \in \sigma} \|x - \phi(x)\|_{\infty}^{p} \right)^{1/p} & \text{if } p \in [1,\infty), \\ \inf_{\phi} \sup_{x \in \sigma} \|x - \phi(x)\|_{\infty} & \text{if } p = \infty, \end{cases}$$
 (1)

donde ϕ corre sobre las biyecciones entre σ y $\tau,$ i.e. biyecciones entre sus representantes.

Espacios $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$

Para $p \in [1, \infty)$, la función d_p define una métrica en

$$\mathcal{D}_p(\mathbb{R}^2_{\geq},\Delta) = \left\{ \sigma \in \mathcal{D}(\mathbb{R}^2_{\geq},\Delta) : \sum_{x \in \sigma} \mathsf{dist}(x,\Delta)^p < \infty \right\}.$$

Si $p = \infty$, tenemos una semi-métrica en

$$\mathcal{D}_{\infty}(\mathbb{R}^2_{\geq},\Delta) = \left\{\sigma \in \mathcal{D}(\mathbb{R}^2_{\geq},\Delta) : \sup_{x \in \sigma} \{ \mathsf{dist}(x,\Delta) \} < \infty \right\}.$$

Un poquito de motivación práctica

• La distancia bottleneck d_{∞} es isométrica a la distancia interleaving entre módulos de persistencia (Cohen-Steiner–Edelsbrunner–Harer, '07; Lesnick, '15).

Un poquito de motivación práctica

- La distancia bottleneck d_{∞} es isométrica a la distancia interleaving entre módulos de persistencia (Cohen-Steiner–Edelsbrunner–Harer, '07; Lesnick, '15).
- Resultados de estabilidad para las p-distancias respecto a perturbaciones en las nubes de puntos correspondientes (Cohen-Steiner-Edelsbrunner-Harer, '07; Cohen-Steiner-Edelsbrunner-Harer-Mileyko, '10; Chazal-De Silva-Oudot, '14; Skraba-Turner, '23).

Un poquito de motivación práctica

- La distancia bottleneck d_{∞} es isométrica a la distancia interleaving entre módulos de persistencia (Cohen-Steiner–Edelsbrunner–Harer, '07; Lesnick, '15).
- Resultados de estabilidad para las p-distancias respecto a perturbaciones en las nubes de puntos correspondientes (Cohen-Steiner-Edelsbrunner-Harer, '07; Cohen-Steiner-Edelsbrunner-Harer-Mileyko, '10; Chazal-De Silva-Oudot, '14; Skraba-Turner, '23).
- Estadística en los espacios $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$ (Mileyko–Mukherjee–Harer, '11; Turner–Mileyko–Mukherjee–Harer, '14).

¿Qué sabemos de $\mathcal{D}_p(\mathbb{R}^2_{\geq},\Delta)$?

¿Qué sabemos de $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$?

• (Mileyko–Mukherjee–Harer, '11): $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$ es completo y separable para cualquier $p \in [1, \infty)$.

¿Qué sabemos de $\mathcal{D}_{ ho}(\mathbb{R}^2_{\geq},\Delta)$?

• (Mileyko–Mukherjee–Harer, '11): $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$ es completo y separable para cualquier $p \in [1, \infty)$.

• (Turner–Mileyko–Mukherjee–Harer, '14; Turner, '20): Siempre es posible encontrar biyecciones óptimas entre diagramas en $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$.

¿Qué sabemos de $\mathcal{D}_{\rho}(\mathbb{R}^2_{\geq},\Delta)$?

- (Mileyko–Mukherjee–Harer, '11): $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$ es completo y separable para cualquier $p \in [1, \infty)$.
- (Turner–Mileyko–Mukherjee–Harer, '14; Turner, '20): Siempre es posible encontrar biyecciones óptimas entre diagramas en $\mathcal{D}_p(\mathbb{R}^2_{\geq}, \Delta)$.
- (Turner–Mileyko–Mukherjee–Harer, '14; Turner, '20; Chowdhury, '21): $\mathcal{D}_p(\mathbb{R}^2_\geq, \Delta)$ es un espacio geodésico para cualquier $p \in [1, \infty)$. Más aún, en $\mathcal{D}_p((\mathbb{R}^2_\geq, \|\cdot\|_p), \Delta)$ y $\mathcal{D}_p((\mathbb{R}^2_\geq, \|\cdot\|_2), \Delta)$ todas las geodésicas están dadas por combinaciones convexas.



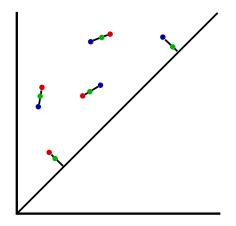


Figura: Combinaciones convexas de diagramas

• (Turner–Mileyko–Mukherjee–Harer, '14): $\mathcal{D}_2((\mathbb{R}^2_{\geq},\|\cdot\|_2),\Delta)$ es un espacio de Alexandrov de curvatura no negativa.

• (Turner–Mileyko–Mukherjee–Harer, '14): $\mathcal{D}_2((\mathbb{R}^2_{\geq}, \|\cdot\|_2), \Delta)$ es un espacio de Alexandrov de curvatura no negativa.

 Los espacios de Alexandrov son generalizaciones de variedades Riemannianas con curvatura acotada inferiormente. • (Turner–Mileyko–Mukherjee–Harer, '14): $\mathcal{D}_2((\mathbb{R}^2_{\geq},\|\cdot\|_2),\Delta)$ es un espacio de Alexandrov de curvatura no negativa.

 Los espacios de Alexandrov son generalizaciones de variedades Riemannianas con curvatura acotada inferiormente.

 En general se trata de espacios métricos con singularidades, pero con suficiente estructura para usar diferentes técnicas de la geometría diferencial.

Diagramas de persistencia generalizados

Pares métricos

(X,A) donde X es un espacio métrico, $A \subset X$ es un cerrado no vacío.

Diagramas de persistencia generalizados

Pares métricos

(X, A) donde X es un espacio métrico, $A \subset X$ es un cerrado no vacío.

Espacio de diagramas de persistencia sobre (X, A)

 $\mathcal{D}(X,A) \coloneqq \{\sigma : \sigma \in X \text{ es un multiconjunto contable}\}/\sim$

donde $\sigma \sim \tau$ si difieren únicamente por puntos en A.

p-distancias entre diagramas de persistencia generalizados

Para un $p \in [1, \infty]$ fijo y $\sigma, \tau \in \mathcal{D}(X, A)$, definimos

$$d_{p}(\sigma,\tau) = \begin{cases} \inf_{\phi} \left(\sum_{x \in \sigma} d(x,\phi(x))^{p} \right)^{1/p} & \text{if } p \in [1,\infty), \\ \inf_{\phi} \sup_{x \in \sigma} d(x,\phi(x)) & \text{if } p = \infty, \end{cases}$$
 (2)

donde ϕ corre sobre las biyecciones entre σ y $\tau,$ i.e. biyecciones entre sus representantes.

Espacios $\overline{\mathcal{D}_p(X,A)}$

Para $p \in [1, \infty)$, la función d_p define una métrica en

$$\mathcal{D}_p(X,A) = \left\{ \sigma \in \mathcal{D}(X,A) : \sum_{x \in \sigma} \operatorname{dist}(x,A)^p < \infty \right\},\,$$

y cuando $p = \infty$, tenemos una semimétrica en

$$\mathcal{D}_{\infty}(X,A) = \left\{ \sigma \in \mathcal{D}(X,A) : \sup_{x \in \sigma} \{ \operatorname{dist}(x,A) \} < \infty \right\}.$$

Teorema (C.–Galaz-García–Guijarro–Membrillo Solis, '22)

Teorema (C.-Galaz-García-Guijarro-Membrillo Solis, '22)

Sea (X,A) un par métrico y $p \in [1,\infty)$. Entonces se cumple lo siguiente:

1. Si X es completo entonces $\mathcal{D}_p(X,A)$ es completo.

Teorema (C.–Galaz-García–Guijarro–Membrillo Solis, '22)

- 1. Si X es completo entonces $\mathcal{D}_p(X,A)$ es completo.
- 2. Si X es separable, entonces $\mathcal{D}_p(X,A)$ es separable.

Teorema (C.–Galaz-García–Guijarro–Membrillo Solis, '22)

- 1. Si X es completo entonces $\mathcal{D}_p(X,A)$ es completo.
- 2. Si X es separable, entonces $\mathcal{D}_p(X,A)$ es separable.
- 3. Si X es propio (i.e. satisface el teorema de Heine-Borel), entonces existen biyecciones óptimas entre cualesquiera dos diagramas de $\mathcal{D}_p(X,A)$.

Teorema (C.-Galaz-García-Guijarro-Membrillo Solis, '22)

- 1. Si X es completo entonces $\mathcal{D}_p(X,A)$ es completo.
- 2. Si X es separable, entonces $\mathcal{D}_p(X,A)$ es separable.
- 3. Si X es propio (i.e. satisface el teorema de Heine-Borel), entonces existen biyecciones óptimas entre cualesquiera dos diagramas de $\mathcal{D}_p(X,A)$.
- 4. Si X es propio y geodésico, entonces $\mathcal{D}_p(X,A)$ es geodésico. Más aún, bajo las misma hipótesis, en $\mathcal{D}_2(X,A)$ todas las geodésicas están dadas por combinaciones convexas.

Teorema (C.-Galaz-García-Guijarro-Membrillo Solis, '22)

- 1. Si X es completo entonces $\mathcal{D}_p(X,A)$ es completo.
- 2. Si X es separable, entonces $\mathcal{D}_p(X,A)$ es separable.
- 3. Si X es propio (i.e. satisface el teorema de Heine-Borel), entonces existen biyecciones óptimas entre cualesquiera dos diagramas de $\mathcal{D}_p(X,A)$.
- 4. Si X es propio y geodésico, entonces $\mathcal{D}_p(X,A)$ es geodésico. Más aún, bajo las misma hipótesis, en $\mathcal{D}_2(X,A)$ todas las geodésicas están dadas por combinaciones convexas.
- 5. Si X es propio y de Alexandrov con curvatura no negativa, entonces $\mathcal{D}_2(X,A)$ es de Alexandrov con curvatura no negativa.

Teorema (C.-Galaz-García-Guijarro-Membrillo Solis-Valiunas, '22)

Sea (X,A) un par métrico. Entonces lo siguiente se cumple:

Sea (X,A) un par métrico. Entonces lo siguiente se cumple:

1. $\mathcal{D}_{\infty}(X,A)$ es un espacio métrico si y sólo si $X \setminus A$ es un espacio discreto.

- 1. $\mathcal{D}_{\infty}(X,A)$ es un espacio métrico si y sólo si $X \setminus A$ es un espacio discreto.
- 2. Si X es completo, entonces $\mathcal{D}_{\infty}(X,A)$ es completo.

- 1. $\mathcal{D}_{\infty}(X,A)$ es un espacio métrico si y sólo si $X \setminus A$ es un espacio discreto.
- 2. Si X es completo, entonces $\mathcal{D}_{\infty}(X,A)$ es completo.
- 3. Si $B_D(A) \setminus B_\delta(A) \subseteq X$ es totalmente acotado para todo $D > \delta > 0$, entonces $\mathcal{D}_{\infty}(X,A)$ es separable.

- 1. $\mathcal{D}_{\infty}(X,A)$ es un espacio métrico si y sólo si $X \setminus A$ es un espacio discreto.
- 2. Si X es completo, entonces $\mathcal{D}_{\infty}(X,A)$ es completo.
- 3. Si $B_D(A) \setminus B_\delta(A) \subseteq X$ es totalmente acotado para todo $D > \delta > 0$, entonces $\mathcal{D}_\infty(X,A)$ es separable.
- 4. Si X es un espacio propio y $\sigma, \tau \in \mathcal{D}_{\infty}(X, A)$, entonces existen diagramas $\sigma', \tau' \in \mathcal{D}_{\infty}(X, A)$ tales que $d_{\infty}(\sigma, \sigma') = d_{\infty}(\tau, \tau') = 0$, y una biyección óptima entre σ' y τ' .

- 1. $\mathcal{D}_{\infty}(X,A)$ es un espacio métrico si y sólo si $X \setminus A$ es un espacio discreto.
- 2. Si X es completo, entonces $\mathcal{D}_{\infty}(X,A)$ es completo.
- 3. Si $B_D(A) \setminus B_\delta(A) \subseteq X$ es totalmente acotado para todo $D > \delta > 0$, entonces $\mathcal{D}_\infty(X,A)$ es separable.
- 4. Si X es un espacio propio y $\sigma, \tau \in \mathcal{D}_{\infty}(X, A)$, entonces existen diagramas $\sigma', \tau' \in \mathcal{D}_{\infty}(X, A)$ tales que $d_{\infty}(\sigma, \sigma') = d_{\infty}(\tau, \tau') = 0$, y una biyección óptima entre σ' y τ' .
- 5. Si X es propio y geodésico, entonces $\mathcal{D}_{\infty}(X,A)$ es geodésico.

Medias de Fréchet

Medias de Freéchet

Dada una medida de probabilidad de Borel μ en un espacio métrico X, una media de Fréchet para μ es un punto $p \in X$ que minimiza la función

$$p \mapsto \int_X d(x,p)^2 d\mu(y)$$

Medias de Fréchet

Medias de Freéchet

Dada una medida de probabilidad de Borel μ en un espacio métrico X, una media de Fréchet para μ es un punto $p \in X$ que minimiza la función

$$p \mapsto \int_X d(x,p)^2 d\mu(y)$$

(Mileyko–Mukherjee–Harer, '11): Sea μ una medida de probabilidad de Borel sobre $\mathcal{D}_p(\mathbb{R}^2_\geq, \Delta)$, con $p \in [1, \infty)$. Si μ tiene soporte compacto o tiene decaimiento a infinito mayor que máx $\{2, p\}$, entonces existen medias de Fréchet para μ .

Sea μ una medida de probabilidad de Borel sobre $\mathcal{D}_p(X,A)$ con $p \in [1,\infty)$. Si μ tiene soporte compacto o decaimiento a infinito mayor que máx $\{2,p\}$, entonces existen una media de Fréchet par μ .

Transporte parcial óptimo

 Podemos describir las p-distancias de diagramas de persistencia en términos de una variante del problema del transporte óptimo: el transporte parcial óptimo.

Transporte parcial óptimo

- Podemos describir las p-distancias de diagramas de persistencia en términos de una variante del problema del transporte óptimo: el transporte parcial óptimo.
- Con el transporte parcial óptimo podemos comparar medidas con masa total diferente usando un conjunto auxiliar de donde podemos tomar masa faltante o a donde podemos echar masa excedente.

Transporte parcial óptimo

- Podemos describir las p-distancias de diagramas de persistencia en términos de una variante del problema del transporte óptimo: el transporte parcial óptimo.
- Con el transporte parcial óptimo podemos comparar medidas con masa total diferente usando un conjunto auxiliar de donde podemos tomar masa faltante o a donde podemos echar masa excedente.
- Originalmente se plantea este formalismo en el contexto de las ecuaciones diferenciales parciales con condiciones de frontera de Dirichlet (Figalli-Gigli, '10).

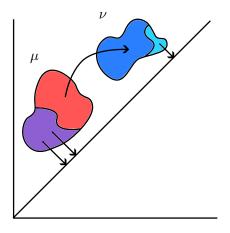


Figura: Transporte parcial entre dos medidas μ y ν en $\mathcal{M}(\mathbb{R}^2_{\geq}, \Delta)$.

Consideremos los conjuntos

$$\mathcal{M}(\mathbb{R}^2_{\geq})$$
 = {medidas de Borel no-negativas localmente finitas en \mathbb{R}^2_{\geq} }

y

$$\mathcal{M}(\mathbb{R}^2_{\geq}, \Delta) = \{ \mu \in \mathcal{M}(\mathbb{R}^2_{\geq}) : \mu(\Delta) = 0 \}.$$

Consideremos los conjuntos

У

 $\mathcal{M}(\mathbb{R}^2_{\geq})$ = {medidas de Borel no-negativas localmente finitas en \mathbb{R}^2_{\geq} }

$$\mathcal{M}(\mathbb{R}^2_{\geq},\Delta)=\{\mu\in\mathcal{M}(\mathbb{R}^2_{\geq}):\mu(\Delta)=0\}.$$

Dadas $\mu, \nu \in \mathcal{M}(\mathbb{R}^2_{\geq}, \Delta)$, un *plan de transporte parcial* entre μ y ν es una medida $\pi \in \mathcal{M}(\mathbb{R}^2_{>} \times \mathbb{R}^2_{>})$ tal que

$$\pi(E \times \mathbb{R}^2) = \mu(E)$$
 y $\pi(\mathbb{R}^2 \times F) = \nu(F)$

para cualesquiera $E, F \subset \mathbb{R}^2_{\geq} \setminus \Delta$ conjuntos de Borel. Denotamos por $\mathsf{Adm}(\mu, \nu)$ el conjunto de planes de transporte parcial entre μ y ν .

La *p*-distancia en $\mathcal{M}(\mathbb{R}^2_>,\Delta)$ está dada por

$$\widetilde{W}_p(\mu,\nu) = \inf_{\pi \in \operatorname{Adm}(\mu,\nu)} \left(\int_{X \times X} \|x - y\|^p \ d\pi(x,y) \right)^{1/p}.$$

La *p*-distancia en $\mathcal{M}(\mathbb{R}^2_{\geq}, \Delta)$ está dada por

$$\widetilde{W}_{p}(\mu,\nu) = \inf_{\pi \in \operatorname{Adm}(\mu,\nu)} \left(\int_{X \times X} \|x - y\|^{p} \ d\pi(x,y) \right)^{1/p}.$$

 \widetilde{W}_{p} define una métrica en el conjunto

$$\mathcal{M}_{p}(\mathbb{R}^{2}_{\geq}, \Delta) = \left\{ \mu \in \mathcal{M}(\mathbb{R}^{2}_{\geq}, \Delta) : \int_{\mathbb{R}^{2}_{\geq}} \operatorname{dist}(x, \Delta)^{p} \ d\mu(x) < \infty \right\}.$$

La p-distancia en $\mathcal{M}(\mathbb{R}^2_{\geq},\Delta)$ está dada por

$$\widetilde{W}_{p}(\mu,\nu) = \inf_{\pi \in \operatorname{Adm}(\mu,\nu)} \left(\int_{X \times X} \|x - y\|^{p} \ d\pi(x,y) \right)^{1/p}.$$

 $\widetilde{W}_{\!\scriptscriptstyle p}$ define una métrica en el conjunto

$$\mathcal{M}_p(\mathbb{R}^2_{\geq}, \Delta) = \left\{ \mu \in \mathcal{M}(\mathbb{R}^2_{\geq}, \Delta) : \int_{\mathbb{R}^2_{\geq}} \operatorname{dist}(x, \Delta)^p \ d\mu(x) < \infty \right\}.$$

Teorema (Lacombe-Cuturi-Oudot, '18; Divol-Lacombe, '21)

La función $\mathcal{D}_p(\mathbb{R}^2_{\scriptscriptstyle\geq},\Delta) o \mathcal{M}_p(\mathbb{R}^2_{\scriptscriptstyle\geq},\Delta)$ dada por

$$\sigma \mapsto \sum_{x \in \sigma \setminus \Delta} \delta_x$$

es un encaje isométrico.

Trabajo en curso

• Es posible definir el problema del transporte parcial óptimo para pares métricos.

Trabajo en curso

 Es posible definir el problema del transporte parcial óptimo para pares métricos.

• Los espacios $\mathcal{M}_p(X,A)$ resultantes heredan propiedades geométricas de los correspondientes pares métricos (X,A).

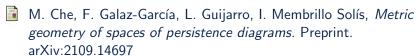
Trabajo en curso

• Es posible definir el problema del transporte parcial óptimo para pares métricos.

• Los espacios $\mathcal{M}_p(X,A)$ resultantes heredan propiedades geométricas de los correspondientes pares métricos (X,A).

• Se tienen encajes isométricos $\mathcal{D}_p(X,A) \to \mathcal{M}_p(X,A)$.

References



M. Che, F. Galaz-García, L. Guijarro, I. Membrillo Solis, M. Valiunas, Basic Metric Geometry of the Bottleneck Distance. To appear in Proceedings of the AMS. Preprint arXiv:2205.09718 ¡Muchas gracias por su atención!