UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Transporte Óptimo y Baricentro con distancia de Fermat

Nicolás Gustavo Chehebar

Director: Dr. Pablo Groisman

Fecha de Presentación: 16 de Julio de 2021

Agradecimientos

A Pablo Groisman, quien además de un gran docente fue un excelente director, por el conocimiento, pasión y entusiasmo que me transmitió, por las horas de llamada en las que tuve el gusto de pensar y aprender con él, por el apoyo en todos los aspectos más allá de la matemática. Gracias Patu por ser un gran guía en este proceso.

A Mariela Sued y Andrés Farall, por leer esta tesis y haber aceptado ser jurados de la misma.

A Facundo Sapienza y Esteban Tabak, por su gran ayuda a la hora de trabajar sobre los temas de esta tesis y por todo lo que he aprendido con ellos.

A los docentes que tuve, por todo lo que me enseñaron. A Gustavo Krimker, por ser mi primer profesor, mantener mi interés y enseñarme a pensar.

A Mamá y Papá por apoyarme en absolutamente todo. A Mariano por el apoyo y por la ayuda a toda hora. A toda la familia, amigos y quienes me acompañaron estos años, por estar siempre.

Índice general

Introducción

Muchas aplicaciones requieren saber como transportar información de un contexto en otro; este problema recobró interés por el avance de técnicas de Machine Learning: se puede entrenar con datos obtenidos a partir de una distribución de probabilidad un algoritmo para aprender algo (un predictor, clasificador, etc.), con el objetivo de aplicarlo luego en nuevas muestras. Sin embargo, las nuevas muestras pueden poseer cualidades estadísticas totalmente distintas, lo que hace que no tenga mucho sentido utilizar el algoritmo entrenado en otro dominio. Lo que sería deseable es lograr que el algoritmo pueda aplicarse en este dominio distinto al cual pertenecen las nuevas muestras, lo que es conocido como "Domain Adaptation". Esta es una de las tantas motivaciones para querer transformar datos que son muestras de una distribución "fuente" en datos que sean muestras de otra distribución "objetivo": de esta manera uno transforma los nuevos datos para que tengan cualidades estadísticas como los datos de entrenamiento, dominio en el cual sí es adecuado usar el algoritmo previamente entrenado.

Dadas una distribución fuente y una objetivo, hay muchas funciones/transportes que transforman a una en otra, pero uno desea elegir una en particular que mueva a los puntos "lo menos posible"; esto es el problema de transporte óptimo, que revisaremos en el Capítulo 1. Un ejemplo sencillo sería preguntarse cómo transportar una $\mathcal{N}(0,1)$ en una $\mathcal{N}(1,1)$, cosa que se puede hacer vía una transformación que a cada x le asigne $x+1$ o vía otra que le asigne $-x+1$. Si bien suena razonable que trasladar 1 todos los puntos sea lo mejor, no es tan inmediato justificar que este transporte es el óptimo. Matemáticamente uno busca el transporte que minimice un costo que, típicamente, es una distancia ya que esto refleja que uno quiere mover poco los puntos. Así, uno puede a partir de una distancia entre los puntos definir un distancia entre distribuciones llamada distancia de Wasserstein: si el costo de transportar una distribución en otra es alto, estarán lejos; si el costo es bajo, estarán cerca. El tener esta distancia permite cuantificar cuán lejos están dos distribuciones entre sí, lo que presenta muchísimas aplicaciones principalmente en procesamiento de imágenes, donde esta distancia mide por ejemplo cuan lejos están las distribuciones de colores. Además teniendo esta distancia podemos definir el baricentro de un conjunto de distribuciones, es

decir una distribución que represente a todas las distribuciones, pudiendo definir también distribuciones intermedias de forma satisfactoria, cosa que un simple promedio no logra.

La teoría de transporte óptimo ha sido un campo muy estudiado desde hace ya más de 200 años, cuando Monge introdujo el problema. Ha recobrado importancia con la relajación dada por Kantorovich gracias al desarrollo de la investigación operativa y programación lineal, teniendo aplicaciones bélicas y económicas con el objetivo de optimizar recursos. Más aún, recobró importancia en el último tiempo por sus aplicaciones en el campo de Machine Learning. El problema ha sido estudiado principalmente con la distancia euclídea, sin embargo esta distancia no siempre es adecuada. Si suponemos que todos los datos se encuentran en una superficie $\mathcal{M} \subseteq \mathbb{R}^D$ (de dimensión quizás mucho menor que D), sería más adecuado considerar una distancia intrínseca de la superficie que una del espacio ambiente. Más aún, sería adecuado dar más importancia a regiones de $\mathcal{M}$ más densas, es decir donde tenemos más datos. La distancia de Fermat, que revisaremos en el Capítulo 2, es una distancia que tiene en cuenta dicha superficie y la densidad de los datos en ella.

El objetivo y aporte de este trabajo es estudiar el problema de transporte óptimo cambiando la noción de distancia entre los puntos, considerando una que tenga en cuenta la estructura de los datos. Siendo la distancia de Fermat adecuada para ello, consideraremos el problema de transporte óptimo donde el costo está dado por esta distancia en vez de la euclídea. Así, buscamos un transporte que "mueva poco los puntos" pero con esta noción de distancia diferente. Esto nos permitirá que los puntos al transportarse "viajen" por regiones pobladas de la superficie en vez de en línea recta que sería el caso euclídeo. Esto se puede ver en la imagen de la izquierda de la Figura 1, donde en vez de trasladar las nubes de puntos de la izquierda para obtener las de la derecha, estas viajan por los puentes. También el costo del transporte nos dará una distancia entre distribuciones distinta a la dada por el costo euclídeo, que será adecuada para definir baricentros y distribuciones intermedias que vivan en $\mathcal{M}$ en vez de en $\mathbb{R}^D$. Esto se puede ver en la imagen de la derecha de la siguiente figura, donde definimos un baricentro sobre la superficie y no como una distribución que esté en el medio de las dos, que sería el caso euclídeo.

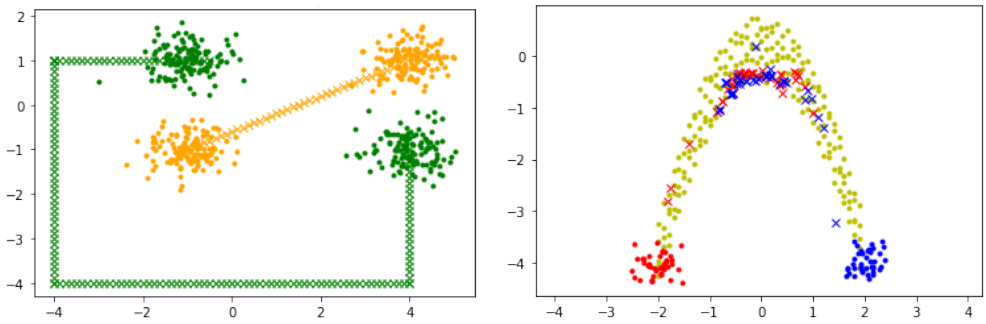


Figura 1: $\mathcal{M}$ está representada por todos los puntos. Izquierda: transportamos las nubes de puntos de la izquierda en las de la derecha, los puntos de igual color representan que fueron transportados unos en otros. Derecha: Representamos con cruces el baricentro de las distribuciones representadas por puntos azules y rojos.

En el Capítulo 3 se presentan las mayores contribuciones originales de esta tesis, donde se revisan, discuten e implementan diversos enfoques algorítmicos para la resolución de este problema, que se pueden encontrar en `https://github.com/nicocheh/FermatOT`. La implementación de algoritmos que logren esto requiere definir estimaciones adecuadas del gradiente para la distancia de Fermat y de densidades en la superficie (que lo haremos vía núcleos pero utilizando distancia de Fermat en ellos en vez de la euclídea). Se muestran resultados satisfactorios en datos sintéticos.

Reconocimiento: Esta tesis presenta parte del trabajo realizado junto con Pablo Groisman, Facundo Sapienza y Esteban Tabak, con quienes he tenido el gusto de pensar sobre estos temas.

Capítulo 1

Transporte Óptimo y Distancia de Wasserstein

En este capítulo introducimos el problema de transporte óptimo y algunas de sus derivaciones, además de resumir algunos de los principales resultados obtenidos en este área.

1.1. Transporte Óptimo

Definiremos primero de manera informal el problema de transporte óptimo como lo hizo Monge originalmente al formularlo: un obrero tiene una pala y debe mover una montaña de arena con una forma dada. Su objetivo será mover esa montaña para construir en otro lugar una nueva forma que podría ser, por ejemplo, un castillo. Para construir el castillo podría llevar la arena de un lugar a otro de muchísimas formas, pero desea minimizar su esfuerzo que lo vamos a cuantificar como la distancia que recorrió cargando arena. Podemos normalizar la masa (la cantidad de arena) tanto de la montaña como del castillo a uno, de forma que se representen distribuciones de probabilidad, que en adelante llamaremos simplemente distribuciones. Esto es un claro ejemplo del problema de transporte óptimo: uno quiere transportar una distribución conocida (la montaña inicial) en otra que también conoce (el castillo que desea construir) minimizando un costo (la distancia entre los puntos).

Como vimos, Monge pensó a la distribución fuente (source) como tierra o masa que habría que mover a otro lugar de forma que quede como lo indica la distribución objetivo

(target). Podemos pensar que toda la masa que se encuentra en un punto se moverá toda junta, a la masa de otro punto: esta es la principal característica de la formulación de Monge. Lo que se quiere minimizar es el costo de ese transporte, que está dado por una función $c(x, y)$ que indica el costo de mover masa de x a y.

El problema podría formularse entonces como minimizar el costo de transportar una distribución μ a una ν (ambas conocidas) sujeto a que a cada punto x se le asigna un $T(x)$ que es su "transportado".

$$\min_{T_{\#}\mu=\nu} \int c(x, T(x))d\mu. \tag{1.1}$$

Donde $T_{\#}\mu = \nu$ se refiere a que después de transportar con T la distribución μ obtendremos ν. Matemáticamente esto es que ν es la medida push-forward de μ vía T, es decir que $\nu(E) = \mu(T^{-1}(E))$ para todo medible E. Típicamente esta función de transporte se define entre dos puntos de $\mathbb{R}^d$ o bien entre dos espacios $\mathcal{X}$ y $\mathcal{Y}$ que son los espacios en los que definimos las medidas μ y ν respectivamente, que en general serán subconjuntos de $\mathbb{R}^d$ (no necesariamente ambos con el mismo d) . Notemos que aquí la función T ofrece un transporte "punto a punto" es decir que conceptualmente toda la masa que hay en x se moverá a un único punto $y = T(x)$.

De otra manera, podemos pensar que la masa del punto x se puede distribuir a varios lugares y no toda a un único $y = T(x)$: esta es la relajación del problema que propuso Kantorovich varios años más tarde. Así, el transporte deja de ser una función T y pasa a estar descripto por una distribución conjunta $\pi(x, y)$ (definida en el espacio producto $\mathcal{X} \times \mathcal{Y}$) que tiene como marginales a μ en la coordenada x y a ν en y, es decir que:

$$\pi(A \times \mathcal{Y}) = \mu(A), \qquad \pi(\mathcal{Y} \times B) = \nu(B), \qquad \forall A \subseteq \mathcal{X}\ ,\ B \subseteq \mathcal{Y}.$$

Y en caso de que μ, ν y π sean densidades (respecto de la medida de Lebesgue):

$$\mu(x) = \int \pi(x, y)dy, \qquad \nu(y) = \int \pi(x, y)dx. \tag{1.2}$$

Conceptualmente, $\pi(x, y)$ nos indica cuánta masa de x esta yendo a y, donde imponemos las restricciones de (1.2) que nos dicen que todo lo que sale de x es $\mu(x)$ y que todo lo que llega a y es $\nu(y)$. Denominaremos $\Pi_{\mu,\nu}$ al conjunto de todas las distribuciones conjuntas

que cumplen esto. Luego la formulación de Kantorovich del problema de transporte óptimo es la siguiente:

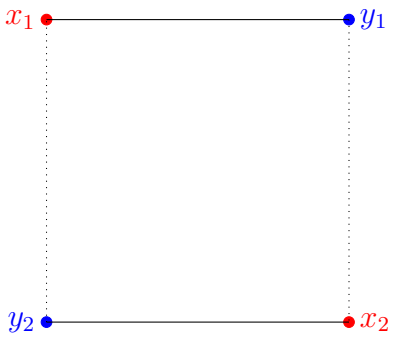
$$\min_{\pi \in \Pi_{\mu,\nu}} \int c(x,y) d\pi(x,y). \tag{1.3}$$

La formulación de Kantorovich (1.3) se trata de una relajación de la de Monge (1.1) porque permite que la masa de x se transporte a distintos lugares y no sólo a uno específico $T(x)$. Más precisamente, toda función T que cumpla las restricciones de la formulación de Monge induce una conjunta $\pi(x,y) = \mu(x)\delta_{T(x)}(y)$ que pertenece a $\Pi_{\mu,\nu}$ ya que cumple (1.2).

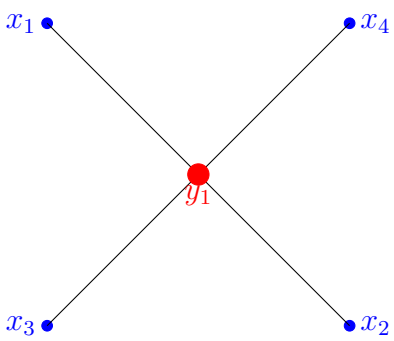
En general, diremos que una solución es factible si cumple las restricciones pedidas, es decir $\pi \in \Pi_{\mu,\nu}$ en la formulación de Kantorovich y $T_{\#}\mu = \nu$ en la formulación de Monge. Por ejemplo, a la hora de transportar una $\mathcal{N}(0,1)$ en una $\mathcal{N}(1,1)$, el transporte $T(x) = x + 1$ es factible y óptimo (aunque aún no lo hemos justificado), $T(x) = -x + 1$ es factible pero no es óptimo y $T(x) = x + 2$ no es factible ya que manda a la $\mathcal{N}(0,1)$ a una $\mathcal{N}(2,1)$ que no es la distribución objetivo.

Veremos a continuación dos ejemplos que muestran diferencias entre ambas formulaciones, poniendo énfasis en la existencia y unicidad de transportes óptimos.

Ejemplo 1.1.1 (Unicidad)**.** Como podemos ver en este ejemplo tenemos dos distribuciones discretas soportadas en los vértices de un cuadrado. Si tomamos el costo dado por la distancia euclídea hay dos posibles funciones de asignación que resuelven el problema en la formulación de Monge (1.1), una con lineas continuas y la otra con lineas punteadas.



Ejemplo 1.1.2 (Existencia)**.** En este ejemplo tenemos una distribución discreta soportada en cuatro vértices y otra soportada en solo un punto. Podemos ver que existe solución de la formulación de Monge (1.1), que es asignar la masa de cada x_i al único punto de la distribución objetivo (la indicada con lineas). De hecho esta es la única solución que existe ya que toda la masa



debe terminar en y_1. Si consideramos el transporte al revés (desde los y hacia los x) no hay solución posible ya que no hay ninguna función de asignación: esto es porque sin importar a donde asignemos y_1 quedarán al menos tres x_i sin masa. En general si se trata de distribuciones discretas, la cantidad de puntos de la distribución fuente debe ser mayor o igual que la cantidad de puntos de la distribución objetivo para que haya alguna función de asignación factible. Notemos en cambio que la formulación de Kantorovich (1.3) sí presenta una distribución conjunta factible ya que podemos dividir la masa que sale de cada punto de la fuente.

El caso discreto (sample-based) es el aplicado en el cómputo ya que típicamente tenemos puntos obtenidos de alguna muestra. Podemos considerar el mismo problema en el caso en que las medidas sean discretas, es decir que sean de la forma:

$$\mu = \sum_{i=1}^{n} a_i \delta_{\{x_i\}}, \qquad \nu = \sum_{j}^{m} b_j \delta_{\{y_j\}}, \tag{1.4}$$

donde δ_x es la delta de Dirac centrada en x, que podemos definirla como objeto límite de muchas formas (como limite de funciones pico, por ejemplo) o en el sentido de Schwartz: como un funcional lineal en un espacio de funciones (las funciones test). La propiedad fundamental que nos interesará de este objeto es que es una "función" que integra 1 y esta concentrada en x, o sea que es la "función" de densidad de la distribución de probabilidad discreta que vale x con probabilidad 1 (es en realidad una distribución temperada, pero bajo ciertas condiciones estas se pueden identificar con funciones y por eso a veces se llama "función" a δ_x).

En este caso el problema es totalmente discreto; podríamos considerar el problema semidiscreto donde una distribución sea discreta y la otra continua o, como hemos visto, el problema continuo donde ambas son continuas. La formulación hasta ahora descripta nos permite trabajar con todo tipo de medidas, desde discretas hasta las que tienen densidad respecto de la medidas de Lebesgue y en general medidas de Radón. En particular, la formulación de Monge (1.1) en el caso discreto sería:

$$\min_{T} \left\{ \sum_i a_i c(x_i, T(x_i)) \ : \ b_j = \sum_{i:T(x_i)=y_j} a_i \ \ \forall j = 1, \ldots, m \right\}, \tag{1.5}$$

donde la condición impuesta es la versión discreta de $T_{\#\mu} = \nu$. Típicamente se considera $a_i = 1/n$ y $b_j = 1/m$ ya que los puntos serán muestras y a priori no hay motivo para dar

más importancia a una que a otra. En particular, si $n = m$ tenemos que T se trata de una biyección. Si además de $n = m$ las distribuciones son uniformes, o sea $a_i = 1/n$ y $b_j = 1/m$, podemos asegurar que hay solución factible y de hecho cualquier asignación es solución. En este caso el problema pasa a ser un problema de asignación (matching) con una matriz de costos $C = c(x_i, y_j)$, es decir hallar la permutación que minimice el costo. Este problema se formula como

$$\min_{\sigma \in Perm(n)} \sum_{i=1}^{n} \frac{1}{n} C_{i,\sigma(i)}, \tag{1.6}$$

donde $Perm(n) = \{\sigma : \{1, \dots, n\} \longrightarrow \{1, \dots, n\}\}$, es el conjunto de permutaciones posibles de n elementos. Notemos que podríamos eliminar el $1/n$ de (1.6) ya que multiplicar por una constante positiva no cambia el mínimo. Dada la matriz $C \in \mathbb{R}^{n \times n}$, este problema pasa a ser uno de optimización combinatoria.

Por otra parte, la formulación (1.3) en el caso discreto se corresponde con hallar una matriz $P \in \mathbb{R}_+^{n \times m}$ que cumpla que $P\mathbb{1}_m = a$ y que $P^T \mathbb{1}_n = b$ (donde $\mathbb{1}_k$ es un vector columna de dimensión k que tiene un 1 en todas sus entradas). Estas condiciones garantizan que la masa que sale de i es la que había (o sea a_i) y que la que llega a j es la que tiene que llegar (o sea b_j). En analogía con π, P_{ij} determina qué cantidad de masa enviamos de la posición i a la posición j. La formulación de Kantorovich en el caso discreto es la siguiente:

$$L_C(a, b) = \min_P \left\{ P : C \quad , \quad P \in \mathbb{R}_+^{n \times m} \, , \, P\mathbb{1}_m = a \, , \, P^T \mathbb{1}_n = b \right\}, \tag{1.7}$$

donde $A : B$ representa el producto punto a punto de matrices, es decir $A : B = \sum_{i,j} A_{i,j} B_{i,j}$. Esto sí se trata de un problema de programación lineal que puede ser resuelto con solvers lineales. Cuando $n = m$, el problema de asignación (1.6) es un caso particular de esta formulación ya que toda permutación induce una matriz $(P_\sigma)_{i,j} = \mathbb{1}_{\{j=\sigma(i)\}}/n$, por lo que el mínimo de la formulación de Kantorovich es menor que el de la de Monge y de ahí que es una relajación. Sin embargo, cuando las distribuciones son ambas uniformes estos problemas son equivalentes como nos muestra la siguiente proposición.

Proposición 1.1.3. *[Kantorovich=Monge] Si $m = n$ y $a = b = \mathbb{1}_n/n$ tenemos que la formulación de Kantorovich discreta (1.7) tiene solución y se corresponde con una permutación que es solución del problema de asignación (1.6), que es la formulación de Monge discreta para pesos uniformes.*

Demostración. Si consideramos el conjunto de todas las matrices P que son factibles para el problema de Kantorovich discreto (1.7), sabemos que se trata de un polítopo convexo ya que imponemos restricciones lineales sobre $P \in \mathbb{R}^{n\times n}$. Luego, el teorema de Birkhoff-Von Neumman de matrices bi-estocásticas [3] nos da que sus vértices son equivalentes al conjunto de matrices de permutación. Dicho teorema nos dice que todas las matrices bi-estocásticas (o sea aquellas en las que las filas y columnas suman uno - en nuestro caso suman $1/n$ y es análogo reescalando -) son combinación convexa de matrices de permutación. Sumado al teorema de Straszewicz (Teorema 18.6 de [4]), que nos dice que todo polítopo es la cápsula convexa de sus vértices, deducimos que los vértices del polítopo son matrices de permutación. Luego, por el teorema fundamental de la programación lineal, el mínimo de la función objetivo se alcanza (si es finito) en un vértice, o sea en una permutación.

Daremos la idea de la demostración del teorema de Birkhoff-Von Neumann siguiendo a [3], donde se puede encontrar con todo detalle. La demostración será por contrarrecíproco, es decir que basta ver que si una matriz no es de permutación entonces no es un vértice del polítopo, que como es convexo equivale a que dicha matriz este en el medio de un segmento que une dos matrices del polítopo. Como una matriz bi-estocástica que no es de permutación tiene un valor que está entre 0 y 1 (en nuestro caso $1/n$) sabemos que en esa fila debe haber otro valor en ese rango. Luego de ese otro valor podemos mirar su columna y encontrar otro y luego mirar la fila de este último. Así encontraremos una secuencia de pares que están en la misma fila o columna que están entre 0 y 1. Esa secuencia en algún momento se corta a si misma y es fácil ver que debe ser de longitud par. La prueba concluye notando que podemos tomar ε suficientemente chico (más que la menor diferencia de los elementos de la matriz seleccionados hasta 0 ó 1) tal que al sumar ε al primer elemento del primer par y restar ε al segundo elemento de dicho par, sumar ε al segundo elemento del segundo par, restar ε al segundo elemento del tercer par, sumar ε al segundo del cuarto y así sucesivamente, logremos modificar todos los elementos de los pares y obtener una matriz que sigue estando en el polítopo. Repitiendo el proceso pero ahora con $-\varepsilon$ obtenemos otra matriz en el polítopo. La matriz original es el promedio de las dos obtenidas, por lo que se encuentra en el segmento que une a estas dos matrices que están en el polítopo que como vimos concluía la demostración. □

Hay un resultado similar que nos habla de la relación de la formulación de Monge y Kantorovich en los casos de medidas en general bajo ciertas condiciones; pero veamos antes ciertas observaciones sobre las diferencias de estas formulaciones:

Observación 1.1.4. La formulación de Monge no presenta ni unicidad ni existencia como muestran los Ejemplos 1.1.1 y 1.1.2 respectivamente.

Observación 1.1.5. La formulación de Kantorovich es simétrica en el sentido de que toda distribución conjunta factible al transportar de μ a ν tiene su correspondiente cuando se transporta de ν a μ. En particular, en el caso discreto (1.7) esto es que si P es un transporte factible en un sentido, P^T es factible para transportar en el sentido opuesto. Es decir, el transporte en la formulación de Kantorovich es reversible, mientras que en la formulación de Monge no necesariamente.

Observación 1.1.6. La formulación de Kantorovich siempre presenta una solución factible ya que $\Pi_{\mu,\nu}$ es no vacío ya que la medida producto siempre pertenece a dicho conjunto.

La función de costo más básica y estudiada es el caso en que esta representa cuan lejos están los puntos, es decir la distancia euclídea entre dos puntos elevada al cuadrado, o sea $c(x, y) = ||x - y||^2$. Nos referiremos a este costo como el "costo euclídeo". En este caso y pidiendo ciertas condiciones sobre una de las distribuciones, podemos asegurar que la formulación de Kantorovich tiene solución y se corresponde con la de Monge. Más aún el transporte T de Monge es el único gradiente de una función convexa que push-forwardea μ a ν. El siguiente teorema generaliza ampliamente la Proposición 1.1.3 y es uno de los resultados más notables de la teoría de transporte óptimo.

Teorema 1.1.7 (equivalencia Kantorovich-Monge, Teorema de Brenier)**.** *Si ambas distribuciones están soportadas en $\mathbb{R}^d$ y la source μ tiene densidad respecto de la medida de Lebesgue, existe un único π óptimo de la formulación (1.3) de Kantorovich con costo euclídeo $c = ||x - y||^2$. Además, este π tiene soporte en $(x, T(x))$ donde T es la única función óptima de la formulación (1.1) de Monge. Más aún, esa función es de la forma $T(x) = \nabla\phi(x)$ donde ϕ es una función convexa. Dicha ϕ es también la única función convexa, salvo constantes, que cumple que T push-forwardea μ a ν, es decir que cumple $(\nabla\phi)_{\#}\mu = \nu$.*

Demostración. Daremos la idea de la demostración que consiste en notar que $\int c d\pi = e - 2\int \langle x, y\rangle d\pi$ donde $e = \int ||x||^2 d\mu(x) + \int ||y||^2 d\nu(y)$ es una constante. Luego, el minimizar la integral de costo equivale al problema de $\max_{\pi\in\Pi_{\mu,\nu}} \int \langle x, y\rangle d\pi(x, y)$. De este problema podemos considerar su dual, donde minimizamos sobre funciones ϕ, ψ el funcional $\int \phi d\mu + \int \psi d\nu$ bajo las restricciones de que $\phi(x) + \psi(y) \geq \langle x, y\rangle$. Esta restricción la podemos reescribir si pasamos restando $\phi(x)$ como $\psi(y) \geq \phi^*(y) := \sup_x \langle x, y\rangle - \phi(x)$. Esta ϕ^* es conocida como la transformada de Legendre y es supremo de lineales, por lo que se trata de una función convexa. Luego, para minimizar en ψ podemos elegir ϕ^* y obtenemos que el problema es

$$\min_{\phi} \int \phi d\mu + \int \phi^* d\nu. \tag{1.8}$$

Si repetimos el mismo argumento dos veces, obtenemos el mismo problema que en (1.8) pero con ϕ^{**} en vez de con ϕ. Así el problema se trata de minimizar en ϕ^{**}, pero ahora tenemos la certeza de que la función sobre la que minimizamos será convexa (pues es una transformada de Legendre). Renombrando (sacando dos *), tenemos el problema (1.8) pero ahora con la garantía de que ϕ es convexa. Como es convexa, es diferenciable c.t.p y, al tener μ densidad respecto de la medida de Lebesgue, también será diferenciable c.t.p respecto de μ.

Por la dualidad, la igualdad $\phi(x)+\phi^*(y) = \langle x, y\rangle$ vale casi seguramente en π, es decir en su soporte. Esto nos da que para puntos (x, y) en el soporte de π tenemos $\nabla\langle x, y\rangle = \nabla\phi(x)$, ya que allí se maximiza $\langle x, y\rangle - \phi(x)$, de donde se deduce que $y = \nabla\phi(x)$. Luego, el π óptimo será $(\mathrm{Id}, \nabla\phi)_{\#}\,\mu$, donde Id representa la función identidad en el espacio correspondiente. La unicidad se obtiene de que $\nabla\phi(x) = \nabla\langle x, y\rangle$ en los lugares donde es diferenciable (o sea μ-c.t.p).

Se incluyó esta idea de demostración para notar la importancia del problema dual en la literatura de transporte óptimo, aunque no ahondaremos en él. Una demostración más rigurosa que incluya demostraciones de los resultados de dualidad se puede encontrar en [5] (ver Teorema 1.22). □

Comentario 1.1.8. El teorema anterior es generalizable a funciones de costo del tipo $c = h(x-y)$ con h estrictamente convexa [5]. Cambiaremos la transformada de Legendre por una $c-$transformada con la diferencia esencial de que la función de costo no es necesariamente

diferenciable por lo que obtendremos que en el soporte de π vale $\nabla\phi(x) \in \partial h(x-y)$ donde ∂ refiere al conjunto subgradiente (que podemos definir pues h es convexa). Usando que h es estrictamente convexa, sabemos que existirá $(\nabla h)^{-1}$ y se recupera $y = x - (\nabla h)^{-1} \nabla\phi(x)$, $\mu - a.e.$.

Observación 1.1.9. El Teorema de Brenier 1.1.7 nos dice que el transporte óptimo de μ a ν es $(\mathrm{Id}, T)_{\#}\mu$ si μ tiene densidad. Si además ν tiene densidad, podemos aplicar el teorema en el transporte óptimo de ν a μ obteniendo que el transporte será $(S, \mathrm{Id})_{\#}\nu$. Esto nos da que en particular $x = S(y)$ y que $y = T(x)$, o sea que $x = S(T(x))$ lo que nos muestra que $S = T^{-1}$. Esto último es que el transporte óptimo es reversible, es decir que la inversa del transporte óptimo en un sentido (de μ a ν) es el transporte óptimo en el sentido opuesto (de ν a μ).

Sabiendo entonces que basta encontrar una función convexa ϕ que cumpla $(\nabla\phi(x))_{\#}\,\mu = \nu$, podemos reescribir $(T(x))_{\#}\,\mu = \nu$ como:

$$\int_{\mathcal{Y}} h(y)d\nu(y) = \int_{\mathcal{X}} h(T(x))d\mu(x), \qquad \forall h \in \mathcal{C}(\mathcal{Y}),$$

donde $\mathcal{C}(\mathcal{Y})$ es el conjunto de las funciones continuas en $\mathcal{Y}$. A su vez suponiendo que μ y ν tienen densidad ρ_μ y ρ_ν y suponiendo también suavidad y biyectividad de T obtenemos que debe valer el cambio de variables $\rho_\mu(x) = |\det(\mathcal{D}T)|\rho_\nu(T(x))$. Además conocemos la forma de esa T que es, por el Teorema de Brenier 1.1.7, $T(x) = \nabla\phi(x)$; luego reemplazando dicha T en el cambio de variables nos queda la siguiente ecuación para ϕ:

$$\rho_\mu(x) = |\det(\mathcal{D}^2\phi)|\rho_\nu(\nabla\phi(x)). \tag{1.9}$$

Esta ecuación es conocida como la ecuación de Monge-Ampère. El operador $|\det(\mathcal{D}^2\cdot)|$ se puede interpretar como una modificación no lineal del laplaciano. De hecho, si hacemos Taylor en una ϕ tal que $\nabla\phi = \mathrm{Id} + \varepsilon\nabla\psi$ se recupera (a orden ε) el laplaciano.

Observación 1.1.10. Si bien resolver esta ecuación con una ϕ convexa resuelve el problema de transporte óptimo, se trata de una tarea para nada simple ya que tenemos dos términos no lineales del lado derecho. El obtener una solución convexa es también una complicación extra. Más aún, la situación es más difícil al considerar otras funciones de costo.

Comentario 1.1.11. Para transporte óptimo basta con considerar una versión más débil de la ecuación (1.9) sin requerir que ϕ sea $\mathcal{C}^2$. Caffarelli estudio dicha ecuación y obtuvo que si ambas densidades están acotadas y el dominio es convexo, entonces la solución de transporte óptimo de Brenier es efectivamente $\mathcal{C}^2$ y solución clásica de la ecuación de Monge-Ampère (1.9) (Teorema 4.14 de [2]).

A continuación resolvemos explícitamente el problema en una dimensión, que nos permite ganar intuición y clarifica como son dichos transportes.

Ejemplo 1.1.12 (Caso 1D)**.** En este caso se puede describir explícitamente la función de Monge que optimiza el transporte. Buscamos una función que sea el gradiente de una función convexa, lo que en $\mathbb{R}$ es una función creciente. Llamemos F_μ, F_ν a las funciones de distribución acumulada de μ y ν respectivamente. Podemos notar que si una función es efectivamente un transporte de μ a ν y es creciente, entonces por el Teorema de Brenier 1.1.7 se trata del transporte óptimo. Construyamos primero dicha función. Intuitivamente, gracias a las funciones de distribución acumulada, podemos saber dado un punto x "cuanto pesa" para μ el intervalo $(-\infty, x]$ (aplicando F_μ) y luego encontrar un punto $T(x)$ que haga que para ν el intervalo $(-\infty, T(x)]$ pese eso (aplicando la inversa de F_ν). Esto es precisamente lo que hace la siguiente función:

$$T(x) = F_\nu^{-1}\left(F_\mu(x)\right).$$

La Figura 1.1 nos muestra como actúa esta función de transporte T.

Recordemos que la función de distribución acumulada determina a la distribución y que no es necesariamente inversible, pero como es continua a derecha podemos definir su inversa como $F^{-1}(x) = \text{ínf}\,\{t \in \mathbb{R} : F(t) \geq x\}$. De esta definición se deduce que $F^{-1}(x) \leq a \iff F(a) \geq x$, por lo que $|\,\{x \in [0,1] : F_\nu^{-1}(x) \leq a\}\,| = |\,\{x \in [0,1] : F_\nu(a) \geq x\}\,| = F_\nu(a)$ de donde se deduce que F_ν^{-1} push-forwardea la medida de Lebesgue en el $[0,1]$ a ν. Si ahora logramos probar que F_μ push-forwardea μ en la medida de Lebesgue del $[0,1]$, habremos demostrado que T push-forwardea la medida μ en ν por composición.

Para ver eso, vamos a asumir que F_μ es continua, que no es más que asumir que es "atomless", o sea que no tiene un punto con densidad infinita. Si tomamos un $a \in [0,1]$ el conjunto $\{x : F_\mu(x) \leq a\}$ es cerrado y es $[-\infty, x_a]$ con $F_\mu(x_a) = a$. Luego, $|[0,a]| =$

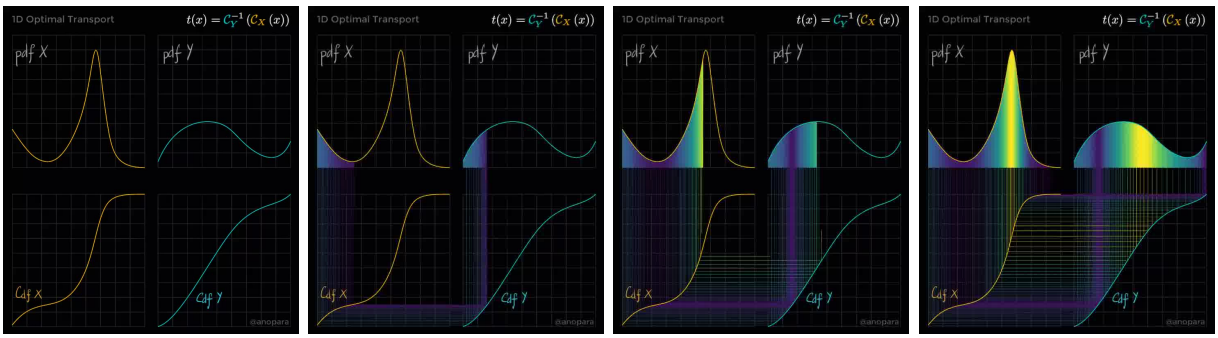


Figura 1.1: Representación del mapa T que transporta la distribución X en Y. En los gráficos de arriba se ven las densidades y en los de abajo las distribuciones acumuladas. Primero se aplica la distribución acumulada de X que se representa en una linea vertical (de arriba a abajo) entre los gráficos de la izquierda. La linea horizontal une las imágenes de cada acumulada. Luego se aplica la inversa de la acumulada de Y que se representa como una linea vertical (de abajo a arriba) entre los gráficos de la derecha. Gif de Anastasia Opara.

$a = F_\mu(x_a) = \mu(\{x : F_\mu(x) \leq a\}) = \mu\left(F_\mu^{-1}([0,a])\right)$ que muestra justamente que $(F_\mu)_{\#}\mu = \mathcal{L}[0,1]$ donde $\mathcal{L}[0,1]$ es la medida de Lebesgue en el $[0,1]$.

Como tanto $F_\mu(\cdot)$ y $F_\nu^{-1}(\cdot)$ son funciones crecientes, es claro que T será creciente y por ende el gradiente de una función convexa. Además probamos que es una función de transporte, o sea que $T_{\#}\mu = \nu$. Por último, utilizando el Teorema de Brenier 1.1.7 obtenemos que este es el transporte óptimo.

Con este resultado del transporte óptimo en una dimensión podemos justificar el ejemplo de la Introducción donde comentamos que una translación $T(x) = x + 1$ es el transporte óptimo de $\mathcal{N}(0,1)$ a $\mathcal{N}(1,1)$. Este transporte tendría costo total $1 = \int ||x + 1 - x||^2 d\mu(x)$ donde $d\mu$ es la densidad de $\mathcal{N}(0,1)$, mientras que el otro transporte propuesto $T(x) = -x+1$ tiene costo $\int ||-x+1-x||^2 d\mu(x) > 1$, mostrando nuevamente que no es el óptimo. Veremos ahora como otro ejemplo el transporte óptimo entre normales en dimensiones más altas.

Ejemplo 1.1.13 (Transporte entre normales en $\mathbb{R}^d$)**.** Consideremos μ y ν dos normales con medias m_μ, m_ν y varianza Σ_μ, Σ_ν que suponemos son definidas positivas, o sea son gaussianas no degeneradas/singulares. Al tratarse de ambas normales sabemos que podemos encontrar algún cambio lineal que traslade (sumar) y reescale (multiplicar por alguna matriz) que nos transforme una distribución en otra, como podemos ver en la Figura 1.2. Para que este sea efectivamente un transporte óptimo debería ser el gradiente de una fun-

ción convexa, pero en el caso de una función de la forma $T(x) = A(x-b)+c$ tenemos que es gradiente de $\frac{1}{2}(x-b)^T A(x-b)+cx$ que es una función convexa siempre que A sea definida positiva.

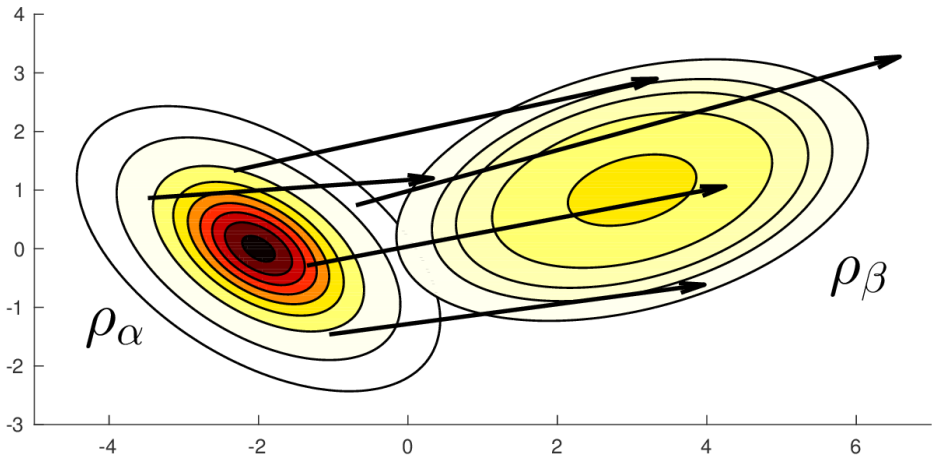


Figura 1.2: Dos normales de densidad ρ_α y ρ_β. Las flechas unen puntos x con su transportado $T(x)$ según transporte óptimo T de (1.11). Observamos también las curvas de nivel de las funciones de densidad de ambas normales. Los colores representan los valores donde las funciones de densidad son más grandes (oscuro) o más pequeñas (claro). Figura 2.12 de [1]

Busquemos entonces tal T, primero trasladaremos al origen, rotaremos allí y luego trasladaremos de vuelta a normal objetivo, propondremos entonces $T(x) = A(x-m_\mu)+m_\nu$. Para que transporte μ a ν debe cumplir el cambio de variable que dio origen a (1.9) que en el caso de las normales es ver que

$$\rho_\nu(T(x)) = \frac{e^{-(T(x)-m_\nu)^T\Sigma_\nu^{-1}(T(x)-m_\nu)}}{\det(2\pi\Sigma_\nu)^{1/2}}|\det(\mathcal{D}T)|, \qquad T(x) = A(x-m_\mu)+m_\nu.$$

$$\implies \rho_\nu(T(x)) = \frac{e^{-(x-m_\mu)^T A^T\Sigma_\nu^{-1}A(x-m_\mu)}}{\det(2\pi\Sigma_\nu)^{1/2}}|\det(A)| \stackrel{?}{=} \rho_\mu(x)|\det(T(x))|. \tag{1.10}$$

Donde para satisfacer la igualdad con ? nos queda elegir A que haga que se cumpla. Si tomamos

$$A = \Sigma_\mu^{-1/2}\left(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2}\right)^{1/2}\Sigma_\mu^{-1/2}, \qquad T(x) = A(x-m_\mu)+m_\nu, \tag{1.11}$$

tenemos que A se trata de una matriz simétrica definida positiva (pues es de la forma $B^{1/2}CB^{1/2}$ con B, C simétricas definidas positivas, donde usamos también que B^{-1} y $B^{1/2}$ son también definidas positivas) y también satisface (1.10) ya que su determinante es $\det(A) = \left(\frac{\det(\Sigma_\nu)}{\det(\Sigma_\mu)}\right)^{1/2}$ y además $A^T\Sigma_\nu^{-1}A = \Sigma_\mu^{-1}$.

Ahora que hemos ejemplificado varios transportes, podemos pensar al proceso de transporte como "continuo". O sea, estamos buscando una familia continua de distribuciones parametrizadas en el tiempo con $t \in [0, 1]$ de forma que a tiempo 0 tengamos la distribución source y a tiempo 1 la objetivo.

Definición 1.1.14 (Interpolante de McCann)**.** Si T es el transporte óptimo entre μ_0 y μ_1 distribuciones definidas en un mismo espacio $\mathcal{X}$, se define el interpolante de McCann como

$$\mu_t = (tT + (1-t)\,\mathrm{Id})_{\#}\,\mu_0. \tag{1.12}$$

Observación 1.1.15. El transporte $tT+(1-t)\,\mathrm{Id}$ utilizado para obtener el interpolante de McCann es el transporte óptimo entre μ_0 y μ_t con $t \in [0, 1]$ ya que sigue siendo el gradiente de una función convexa porque T lo es y la Id también. Además, el costo de transporte escala como t^2 ya que

$$\int |x - [(1-t)x + tT(x)]\,|^2 d\mu_0(x) = t^2 \int |x - T(x)|^2 d\mu_0(x),$$

por lo que el costo de transportar de μ_0 a μ_t es el costo de transportar de μ_0 a μ_1 multiplicado por t^2.

La interpolación de McCann nos permite definir distribuciones intermedias en un nuevo sentido que muchas veces puede ser el deseado, mientras que las que podrían definirse como distribuciones intermedias haciendo un promedio de las densidades de las distribuciones son mucho mas pobres. Si hacemos un simple promedio la distribución fuente irá teniendo menor peso mientras la objetivo obtiene mayor peso, lo que no es muy informativo como distribución intermedia; sin embargo, la interpolación de McCann irá trasladando y deformando una distribución en otra, como muestra la Figura 1.3.

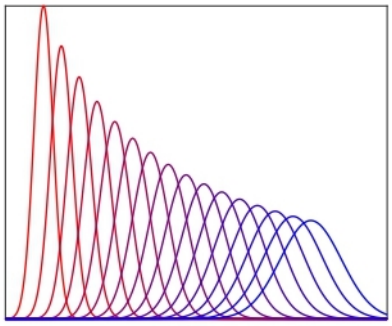


Figura 1.3: Interpolantes de McCann entre dos normales a distintos tiempos, en rojo la distribución source y en azul la objetivo. Figura 2.14 de [1].

Los interpolantes de McCann son muy importantes ya que nos permiten definir distribuciones intermedias. Así, podemos pensar que transportar una distribución en otra es componer los transportes entre distintas distribuciones intermedias. Ahora el problema se basa en resolver varias veces el problema de transporte óptimo "local" entre dos distribuciones mucho mas cercanas. Al tratarse de dos distribuciones cercanas, el transporte será similar a la identidad lo que hace al problema mucho más tratable, este enfoque fue el dado por Kuang y Tabak [16] para proponer un algoritmo de transporte óptimo y baricentro de Wasserstein que introduciremos luego. En esencia, la interpolación de McCann nos permite, sabiendo resolver el problema de transporte óptimo local, poder resolver el problema en general entre dos distribuciones "lejanas".

El problema de transporte óptimo y encontrar distribuciones intermedias puede tener muchas aplicaciones [18]. Inicialmente las aplicaciones fueron militares y económicas, orientadas a optimizar recursos. Mas recientemente, tiene numerosas aplicaciones en problemas de ciencias de datos, principalmente en Machine Learning y procesamiento de imágenes. Se puede pensar como una buena forma de transformar un conjunto de datos en otro. Si pensamos a las imágenes como una distribución de la cual muestreamos pixels $\in \mathbb{R}^3$ donde cada coordenada representa la intensidad de cada tono de color RGB (Rojo, Verde y Azul), podemos hacer "color transfer" [16] [17] entre una imagen y otra. Si una imagen S (source) es $\{x_i\}_{i=1}^n$ y la queremos transportar en una imagen T (target) que es $\{y_i\}_{i=1}^n$, tenemos que resolver el problema (1.6). Al resolverlo y obtener una imagen U dada por los transportados $\{y_{\sigma(i)}\}_{i=1}^n$, tendremos una imagen que luce como S pero que tiene el estilo de colores de T, lo que puede verse en la siguiente figura.

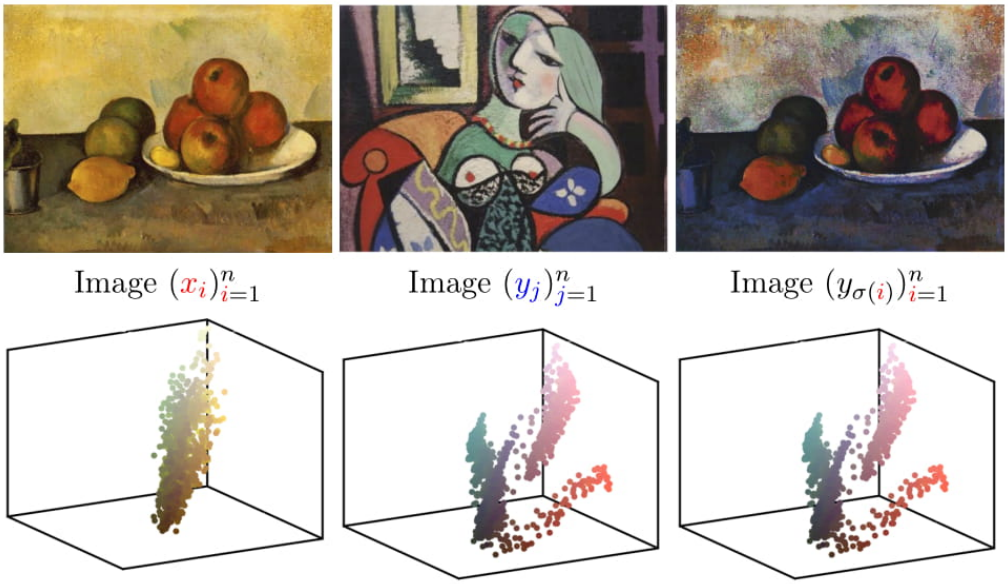


Figura 1.4: Ejemplo de "color transfer". En la fila de arriba las imágenes S, T y U y en la de abajo los pixels como puntos de $\mathbb{R}^3$. Imagen de Gabriel Peyré, Figura 6 de [18]

Por otra parte, podemos calcular el costo de transporte entre un par de distribuciones para ver si estas se parecen o no. Esto tiene por ejemplo aplicaciones en análisis de texto [20]: podemos pensar a estos como un conjunto de palabras $\{x_i\}_{i=1}^n$ y $\{y_i\}_{i=1}^m$ y tenemos dos distribuciones como en (1.4) donde a_i y b_i son proporcionales a la cantidad de apariciones de la palabra (como se puede ver en la Figura 1.5). Así, tendremos que resolver el problema de transporte óptimo (1.5) donde el costo del transporte óptimo nos dirá cuantitativamente cuan parecidos son los dos discursos. En este problema también aparece una gran dificultad en elegir el costo, que dirá "cuán lejos" están dos palabras. Nuevamente a partir de un costo o distancia punto a punto, podemos obtener un costo o distancia entre dos conjuntos de puntos que podemos pensar como distribuciones.

Otra interesante aplicación de transporte óptimo es en el contexto de adaptación de dominio ("domain adaptation"), que se ha visto fuertemente potenciado por los avances y aplicaciones de Machine Learning en los últimos años. En muchas ocasiones se entrena algún algoritmo, función o clasificador en un contexto que podemos pensar como un espacio Y, para que sea eficiente con nuevos datos de $y \in Y$. Uno desea poder extender su funcionamiento a nuevos contextos distintos, pudiendo aplicarlo en un nuevo dominio

Figura 1.5: Distribución de palabras que define un discurso, el tamaño de cada palabra es proporcional a su cantidad de apariciones. Imagen de Gabriel Peyré, Figura 8 de [18]

X. Usar transporte óptimo para adaptación de dominio [32] consiste en transportar datos del dominio X al dominio adecuado Y, siendo una gran forma de generalizar a nuevos dominios.

1.2. Distancia de Wasserstein

Como hemos visto el problema de transporte óptimo nos permite cuantificar el costo de trasladar una distribución en otra. A partir de una función de costo punto a punto, podemos obtener un costo de transportar una distribución en otra. En el caso discreto, podemos vía transporte óptimo generalizar un costo punto a punto a un costo entre histogramas o conjuntos de puntos, que representan distribuciones discretas.

Tener un costo entre distribuciones es muy útil ya que nos permite cuantificar cuáles de ellas "se parecen" (están cerca) y cuales no (están lejos). Esto nos lleva a preguntarnos si este costo de transporte óptimo puede ser una distancia en el espacio de distribuciones. Inicialmente se estudió el problema de transporte óptimo con costo euclídeo, o sea $c(x, y) = ||x - y||^2$ y podemos tomar como distancia entre distribuciones a la raíz cuadrada del costo del transporte óptimo de μ a ν, es decir:

$$W_2(\mu, \nu) = \left(\min_{\pi \in \Pi_{\mu,\nu}} \int ||x - y||^2 d\pi(x, y) \right)^{1/2}. \tag{1.13}$$

Esta función W_2 es la 2-distancia de Wasserstein, que se generaliza a W_p cambiando el costo por el costo euclídeo elevado a la p y normalizando con $1/p$. Esta distancia es también

llamada "Earth Mover Distance" ya que justamente se obtiene de resolver el problema de transporte óptimo que mide el costo de mover la tierra de la distribución source μ hasta llegar a ν, que comentamos al inicio de este capítulo. En general, las distancias de Wasserstein pueden generalizarse a otras funciones de costo y, bajo ciertas condiciones, estas pueden ser efectivamente distancias en el espacio de distribuciones. Para comenzar, veamos esto en el caso discreto, es decir cuando las distribuciones están soportadas en los mismos n puntos y quedan definidas por un vector en $\mathbb{R}^n$ que asigna los pesos a_i de μ en (1.4).

Definición 1.2.1. Una matriz $D \in \mathbb{R}^{n\times n}$ define una distancia en un conjunto de n puntos si:

1. Es positiva y simétrica, i.e, $D \in \mathbb{R}_+^{n\times n}$ y $D = D^T$.
2. Vale 0 sólo en la diagonal, i.e, $D_{i,j} = 0 \iff i = j$.
3. Vale la desigualdad triangular, i.e, $\forall (i,j,k) \in \{1,\dots,n\}, \quad D_{i,k} \leq D_{i,j} + D_{j,k}$.

Recordemos las tres propiedades que debemos probar para ver que W es una distancia entre las distribuciones soportadas en n puntos:

1. W es positiva y simétrica.
2. $W(\mu,\nu) = 0 \iff \mu = \nu$ para μ, ν como en (1.4).
3. Cumple la desigualdad triangular, o sea $W(\mu,\eta) < W(\mu,\nu) + W(\nu,\eta)$ para μ,ν,η como en (1.4).

Proposición 1.2.2. *Si en la formulación discreta de Kantorovich (1.7), $n = m$, $x_i = y_i$ $\forall i = 1,\dots,n$ y $C = D^p = \left(D_{i,j}^p\right)_{i,j} \in \mathbb{R}^{n\times n}$ para algún $p \geq 1$ con D una matriz que es una distancia, la solución a dicho problema elevada a la $1/p$ es una distancia entre distribuciones soportadas en esos n puntos. Esto es que $W_p(\mu,\nu) = L_C(\mu,\nu)^{1/p}$ es una distancia entre las distribuciones soportadas en los n puntos x_i.*

Demostración. Seguimos la demostración de la Proposición 2.2 de [1].

La propiedad 2. se deduce fácilmente porque al ser D una distancia, D^p tiene ceros en la diagonal por lo que $W_p(\mu,\mu) = 0$ para cualquier μ como en (1.4) ya que tomamos

$P = \text{diag}(a)$. Además si $\mu \neq \nu$, como todos los elementos de fuera de la diagonal de D^p son positivos y al tener que haber una entrada de P no nula que no este en la diagonal (pues $\mu \neq \nu$) tenemos que $W_p(\mu, \nu) > 0$. Esto prueba 2. y la positividad de 1. Además, como D^p es simétrica, $W_p(\mu, \nu)$ también lo será. Luego, tenemos también 1.

Para probar la desigualdad triangular 3., debemos pegar los transportes óptimos de cada uno de los dos problemas: el transporte P de μ a ν y el Q de ν a η, lo que en este caso se puede obtener simplemente multiplicándolas y escalando correctamente para obtener S, una conjunta válida en el transporte de μ a η. Llamemos a a los pesos de cada $\delta_{\{x_i\}}$ para μ, b a los de ν y c a los de η. Luego definimos dicha S de la siguiente forma:

$$S = P \begin{pmatrix} \frac{1}{\mathbb{1}_{\{b_1=0\}} + b_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\mathbb{1}_{\{b_n=0\}} + b_n} \end{pmatrix} Q.$$

En esencia la matriz diagonal por la que reescalamos es la que tiene en la diagonal $1/b_j$ solo que en el caso en que algún $b_j = 0$ lo cambiamos por un 1, para evitar dividir por 0. A este vector de la diagonal lo llamaremos $\hat{b}$. Esta matriz S es factible para la formulación de Kantorovich discreta (1.7) porque al multiplicar por unos a derecha, $Q\mathbb{1}_n = b$ por ser Q transporte de ν a η. Luego al reescalar con la matriz diagonal obtenemos un vector de todos 1 salvo en las posiciones j en que $b_j = 0$, donde vale 0. Estos valores no nos interesan puesto que para esos j, $P_{i,j} = 0 \ \forall i = 1, \ldots, n$ debido a que no llega masa a dichos j pues $b_j = 0$. Luego como multiplicamos por un vector de unos a P (salvo donde tiene entradas 0, pero podríamos cambiarlo por unos y la cuenta sería la misma) obtenemos la distribución source del transporte ya que $P\mathbb{1}_n = a$. Entonces $S\mathbb{1}_n = a$.

Al trasponer y multiplicar de nuevo por un vector de unos el razonamiento es análogo porque P^T transporta de ν a μ y Q^T de η a ν por la Observación 1.1.5. Entonces $S^T\mathbb{1}_n = b$, o sea que se puede verificar que S es un transporte de a a b, es decir de μ a ν, porque cumple las condiciones de (1.7). Veamos entonces que vale la desigualdad triangular:

$$W_p(\mu, \eta) \leq (S : D^p)^{1/p} = \left(\sum_{i,k} D_{i,k}^p \sum_j \frac{P_{i,j} Q_{j,k}}{\hat{b}_j} \right)^{1/p} \leq \left(\sum_{i,j,k} (D_{i,j} + D_{j,k})^p \frac{P_{i,j} Q_{j,k}}{\hat{b}_j} \right)^{1/p},$$

donde la primer desigualdad es porque S es, como ya vimos, un transporte factible mientras que W_p se trata del ínfimo. Luego escribimos a S y el producto punto a punto y en la última

desigualdad usamos la desigualdad triangular de D ya que se trata de una distancia. Para terminar, aplicamos la desigualdad de Minkowski con pesos y usamos que $\sum_k \frac{Q_{j,k}}{\hat{b}_j} \leq 1$ y que $\sum_i \frac{P_{i,j}}{\hat{b}_j} \leq 1$ pues al sumar la j-ésima fila de Q obtenemos la distribución source que es viene dada por b y sumando en la j-ésima columna de P obtenemos la distribución objetivo que es de nuevo la dada por b. Luego, la suma da b_j y lo dividimos por si mismo si es positivo y por 1 si es 0 por lo que será ≤ 1. Entonces

$$W_p(\mu,\eta) \leq \left(\sum_{i,j,k} D_{i,j}^p P_{i,j} \frac{Q_{j,k}}{\hat{b}_j}\right)^{1/p} + \left(\sum_{i,j,k} D_{j,k}^p Q_{j,k} \frac{P_{i,j}}{\hat{b}_j}\right)^{1/p} \leq$$

$$\left(\sum_{i,j,} D_{i,j}^p P_{i,j}\right)^{1/p} + \left(\sum_{j,k} D_{j,k}^p Q_{j,k}\right)^{1/p} = W_p(\mu,\nu) + W_p(\nu,\eta).$$

□

Observación 1.2.3. En el caso $0 < p \leq 1$ la distancia es $W_p(a,b)^p$ ya que D^p es directamente una distancia y es todo análogo salvo que nos salteamos un paso.

Nuevamente, lo que podemos obtener en el caso discreto se puede generalizar al caso continuo esta vez valiéndonos de una herramienta más teórica como es el "Gluing Lemma". La demostración es en esencia parecida sólo que la construcción de un transporte de μ a η a partir de tener uno de μ a ν y otro de ν a η requiere un mayor marco teórico que nos diga como pegar estos dos transportes para obtener uno nuevo, lo que en el caso discreto era tan simple como multiplicar las matrices.

Lema 1.2.4 (Gluing Lemma)**.** *Sean μ_1, μ_2, μ_3 tres medidas soportadas en X_1, X_2, X_3 espacios métricos completos y separables (u homeomorfos a algún espacio así, es decir "Polish Spaces"). Sean $\pi_{1,2}$ y $\pi_{2,3}$ transportes en Π_{μ_1,μ_2} y Π_{μ_2,μ_3} respectivamente. Luego, existe una medida π soportada en el producto de los tres espacios con marginales $\pi_{1,2}$ en $X_1 \times X_2$ y $\pi_{2,3}$ en $X_2 \times X_3$*

Demostración. Esta demostración utiliza fuertemente el concepto de desintegración de una medida el cual mencionaremos pero no profundizaremos (ver más en el capítulo 7 de [2]). La desintegración de la medida muestra rigurosamente y de forma más general que podemos

tomar una probabilidad condicional. Nos dice que podemos describir una medida π en el producto de ciertos espacios $X_1 \times X_2$ como un promedio de medidas en $\{x\} \times X_2$ con $x \in X_1$. En particular si π tiene marginal μ_1 en X_1, existe una única (c.t.p en μ_1) función medible que a cada x le asigna π_x una medida soportada en X_2 tal que, para todo medible $A \subseteq X_1 \times X_2$, vale

$$\pi(A) = \int_{X_1} \pi_x(A_x) d\mu_1(x) \,, \text{ con } A_x = \{y \in X_2 : (x, y) \in A\} \,.$$

Justamente esto es que podemos descomponer la medida como la integral respecto de una medida en x. Esto lo podemos notar también de la siguiente manera:

$$\pi = \int_{X_1} \delta_x \otimes \pi_x d\mu(x),$$

donde $\otimes$ se refiere a la medida producto. Esta notación es adecuada para mostrar que la desintegración de la medida es el proceso opuesto a la construcción de una medida producto. A partir de esto, podemos desintegrar las dos medidas producto $\mu_{1,2}$ y $\mu_{2,3}$ con respecto a su marginal en común μ_2 y obtener que existen funciones $\pi_{1,2:\mu_2}$ y $\pi_{2,3:\mu_2}$ de X_2 al espacio de medidas de X_1 y X_3 respectivamente tales que:

$$\pi_{1,2} = \int_{X_2} \pi_{1,2:\mu_2} \otimes \delta_x d\mu_2(x), \qquad \pi_{2,3} = \int_{X_2} \delta_x \otimes \pi_{2,3:\mu_2} d\mu_2(x)$$

Luego podemos construir la medida π deseada como:

$$\pi = \int_{X_2} \pi_{1,2:\mu_2} \otimes \delta_x \otimes \pi_{2,3:\mu_2} d\mu_2(x),$$

ya que si integramos en X_1 se va $\pi_{1,2:\mu_2}$ y nos queda la medida $\pi_{2,3}$ y si integramos en X_3 se va $\pi_{2,3:\mu_2}$ y nos queda la medida $\pi_{1,2}$ □

Teorema 1.2.5 (Distancia de Wasserstein)**.** *Si en la formulación de Kantorovich (1.3) $\mathcal{X} = \mathcal{Y}$ y para $p \geq 1$, $c(x, y) = d(x, y)^p$ con d una distancia en $\mathcal{X}$, entonces el costo de transporte óptimo elevado a la $1/p$ define una distancia (que depende de d) en el espacio de las distribuciones de $\mathcal{X}$.*

Observación 1.2.6. La distancia del Teorema 1.2.5 queda definida en el conjunto de las distribuciones que tengan costo de transporte óptimo finito. Si d es acotada quedará definida en todas las distribuciones de $\mathcal{X}$. En el caso que d sea $||\cdot||^p$ entonces es una distancia entre distribuciones con p-ésimo momento finito.

Demostración. (del Teorema 1.2.5)

La simetría se debe a la simetría comentada en la Observación 1.1.5 del problema de transporte óptimo en la formulación de Kantorovich (1.3). La positividad y que $W_p(\mu,\mu) = 0$ son triviales. Para ver que $W_p(\mu,\nu) = 0 \implies \mu = \nu$, basta notar que un transporte π (una conjunta) que tenga costo cero debe estar soportado en la recta $y = x$ (y por ende $\int_{X\times Y} \phi(x)d\pi(x,y) = \int_{X\times Y} \phi(y)d\pi(x,y)$) por lo que para cualquier ϕ medible tenemos que:

$$\int_X \phi(x)d\mu(x) = \int_X \phi(x) \int_Y d\pi(x,y) = \int_{X\times Y} \phi(x)d\pi(x,y) =$$

$$= \int_{X\times Y} \phi(y)d\pi(x,y) = \int_Y \phi(y) \int_X d\pi(x,y) = \int_Y \phi(y)d\nu(y),$$

lo que muestra que $\mu = \nu$.

Por último, la desigualdad triangular es análoga a la de la Proposición 1.2.2 que establece lo mismo en el caso discreto, solo que nos valemos del Gluing Lemma 1.2.4 para la demostración de la desigualdad triangular y reemplazamos la desigualdad de Minkowski por su versión integral. □

El hecho de poder comparar distribuciones con una distancia presenta grandes ventajas respecto de otras formas de hacerlo como la divergencia de Kullback-Leibler que no es simétrica ni cumple la desigualdad triangular. Además, con la distancia de Wasserstein podemos comparar distribuciones que se encuentren en distintos espacios. Esta distancia con el caso del costo euclídeo (es decir W_1) presenta numerosas aplicaciones en procesamiento de imágenes, donde se utiliza para comparar las distribuciones de colores (sea en RGB o escala de grises). Recientemente también ha sido utilizada en [19] para comparar distribuciones en el contexto de entrenamiento de GANs ("Generative Adversarial Networks").

1.3. Baricentro de Wasserstein

Es un problema muy interesante y útil representar una media o elegir un buen representante de muchos datos. Típicamente en $\mathbb{R}^d$ podemos definir al baricentro de un conjunto de puntos $\{x_i\}_{i=1,\dots,n}$ con pesos $\{w_i\}_{i=1,\dots,n}$ ($\sum_{i=1}^n w_i = 1$) como $\overline{x} = \sum_{i_1}^n w_i x_i$. Esta definición si bien útil es poco generalizable y depende fuertemente de la estructura de $\mathbb{R}^d$

con la distancia euclídea. Si estuviéramos en una superficie, tomar un promedio pesado es probable que ni caiga en esa superficie. Más aún, si a un mismo espacio lo dotáramos de otra distancia el baricentro no cambiaría para nada. Claramente esta no es una definición que podamos imitar en otros espacios y con otras distancias. Sin embargo al baricentro lo podemos definir también como la solución de:

$$\min_x \sum_{i=1}^{n} w_i ||x_i - x||^2. \tag{1.14}$$

Esta última definición, que es equivalente a la anterior ya que basta notar que es el único punto crítico de una función convexa, nos permite generalizar el baricentro con mucha facilidad. De hecho, otras nociones de centralidad se pueden obtener modificando este problema, por ejemplo la mediana se obtiene con el $||\cdot||$ en vez de $||\cdot||^2$.

Más aún, la podemos generalizar a cualquier espacio métrico con una distancia porque es lo único que utilizamos en la formulación de la minimización: la distancia y los x_i con sus pesos que los suponemos dados. De esta forma podemos obtener lo que se conoce como Medias de Frechet ("Frechet Means"); vale la pena notar que este mínimo no necesariamente es único.

Si ahora consideramos el espacio métrico de las distribuciones definidas en un mismo $\mathcal{X}$, donde hemos definido una distancia de Wasserstein W, y reemplazamos $||\cdot||^2$ por dicha distancia podemos definir el baricentro de Wasserstein de $\mu_1, \ldots, \mu_n$ como

$$\overline{\mu} = \text{argmin}_\mu \sum_{i=1}^{n} w_i W(\mu, \mu_i). \tag{1.15}$$

Este baricentro será a su vez una distribución definida en $\mathcal{X}$ que podemos considerar como la distribución intermedia entre $\mu_1, \ldots$ y μ_n.

Si bien en el caso de puntos en $\mathbb{R}^d$ es sencillo resolver el problema de optimización, se trata de una tarea difícil para medidas en general. Mas aún, podríamos generalizar el problema a definir el baricentro de infinitas distribuciones considerando una distribución M (que juega el rol de los pesos w_i) en el espacio de distribuciones y el baricentro sería:

$$\overline{\mu} = \text{argmin}_\mu \int W(\mu, \nu) dM(\nu). \tag{1.16}$$

Notemos que tomando M una distribución discreta (es decir $M = \sum_{i=1}^{n} w_i \delta_{\mu_i}$) recuperamos (1.15), o sea el baricentro de finitas medidas. Podemos estimar el baricentro de infinitas

distribuciones obtenidas a partir de M calculando el baricentro en su versión empírica (o sea con finitas) ya que se demostró la consistencia en la sección 7 de [10]. Es decir que si tomamos n muestras obtenidas de M (que son distribuciones) y calculamos su baricentro $\overline{\mu}_n$, este converge al $\overline{\mu}$ definido en (1.16).

Veamos inicialmente un caso discreto del baricentro como ejemplo. Consideremos medidas $\mu_1, \ldots, \mu_m$ soportadas en n puntos y por ende definidas con pesos $b_1, \ldots, b_m$, y supongamos que queremos encontrar la distribución baricentro $\overline{\mu}$ (con pesos w_i dados) pero restringiendo $\overline{\mu}$ este soportada en los mismos n puntos, quedando definida por pesos a. Si contamos también con la matriz de costo c que define las distancias dos a dos entre los n puntos en que se soportan las distribuciones, encontrar el a que representa el baricentro es resolver

$$\min_{a, P_i \in \mathbb{R}^{n\times n}} \left\{ \sum_{i=1}^{m} w_i \, P_i : C \ , \ \forall i = 1, \ldots, m \ \ P_i^T \ \mathbb{1}_n = a, \ \ P_i \mathbb{1}_n = b \right\}. \tag{1.17}$$

Aquí estamos no solo optimizando sobre cada transporte P_i de μ_i al baricentro $\overline{\mu}$, sino también sobre el mismo baricentro que viene dado por los pesos a.

Observación 1.3.1 (k-medias)**.** Si en vez de tener m distribuciones μ_i tenemos una sola μ, imponemos que el baricentro esté soportado en exactamente k puntos de los n y tomamos como costo la norma euclídea al cuadrado, el problema es equivalente al de k-medias donde los centros de cada cluster serán los puntos donde está soportado a. Tomar baricentro de una sola distribución μ exigiendo que pertenezca a un conjunto de distribuciones $\mathcal{D}$ es lo mismo que buscar la distribución de $\mathcal{D}$ más cercana a μ en distancia de Wasserstein.

El problema del baricentro de Wasserstein tiene en algunos casos solución única, lo que fue demostrado por Agueh y Carlier [6] como cuenta el siguiente teorema.

Teorema 1.3.2. *Si alguna de las medidas $\mu_1, \ldots, \mu_m$ (definidas en $\mathbb{R}^d$) vale cero en conjuntos pequeños (es decir en todo boreliano de dimension de Hausdorff $\leq d-1$) el problema del baricentro de Wasserstein (1.15) con el costo euclídeo tiene única solución.*

Demostración. Referimos al paper original de Agueh y Carlier, donde este resultado es la Proposición 3.5 [6]. □

Comentario 1.3.3. En particular, la condición de que una medida se anule en conjuntos pequeños se cumple si la medida tiene densidad. En su demostración proveen además una caracterización de dicho baricentro a partir de analizar el problema dual de (1.15).

Ejemplo 1.3.4 (Baricentro entre dos medidas)**.** En el caso de calcular el baricentro entre μ_0 y μ_1 con pesos $w_i = 1/2$, este equivale a $\mu_{1/2}$ el interpolante de McCann como se indica en (1.12). Si ponemos pesos $w_0 = 1 - t, w_1 = t$ el baricentro será μ_t. Intuitivamente esto es razonable, si al sumar pesadamente los dos transportes de μ_0 y μ_1 al baricentro no se cancelan es que podemos mover la distribución según nos dice la dirección contraria a la que da esa suma y descender la función objetivo del problema del baricentro (1.17). Por ejemplo, si tenemos una $\mathcal{N}(0,1)$ y otra $\mathcal{N}(3,1)$ su baricentro sería una $\mathcal{N}(1{,}5,1)$; si tomamos una $\mathcal{N}(1,1)$, los transportes a ella serían $T_1(x) = x + 1$ y $T_2(x) = x - 2$ que al sumarlos con pesos $1/2$ nos da $x - 1/2$, lo que nos indica que conviene trasladar la normal $\mathcal{N}(1,1)$ hacia los positivos. Esto podemos generalizarlo también al sumar los transportes de varias medidas como indica la siguiente Proposición 1.3.5. La demostración de que el interpolante de McCann es el baricentro entre dos medidas es inmediata a partir de ella.

Proposición 1.3.5. *Si en el problema del baricentro (1.15) (con medidas definidas en $\mathbb{R}^d$ y costo euclídeo) llamamos T_i al transporte óptimo entre μ_i y μ y definimos $T^{(\mu)} := \sum_{i=1}^n w_i T_i$ entonces:*

1. *$T^{(\mu)}$ es el transporte óptimo entre μ y $T^{(\mu)}_{\#}\mu$.*
2. *La condición de optimalidad (necesaria y suficiente) de primer orden del problema del baricentro (1.15) es $T^{(\overline{\mu})} = \mathrm{Id}$ en el óptimo $\overline{\mu}$ (el baricentro).*
3. *Bajo condiciones de regularidad, el baricentro es el único punto fijo de la ecuación $T^{\mu}_{\#}\mu = \mu$.*
4. *Bajo ciertas condiciones en las medidas $\mu_1, \ldots, \mu_n$ la función G que asigna $\mu \longrightarrow T^{\mu}_{\#}\mu$ es una función de descenso para la función objetivo del problema del baricentro (1.15) y realizar estas iteraciones convergen al baricentro, o sea $\mu^l = G(\mu^{l-1}) \longrightarrow \overline{\mu}$.*

Demostración. Para demostrar 1 basta notar que como cada T_i es óptimo es el gradiente de una función convexa por el Teorema de Brenier 1.1.7. Esto implica que T^{μ} es una suma

de gradientes de convexas, por lo que es gradiente de suma de convexas que también es convexa, por lo que es el gradiente de una función convexa. Sumado a que claramente T^μ es un transporte válido de μ a $T^\mu_\# \mu$ y usando de nuevo Brenier, obtenemos que T^μ es transporte óptimo.

Para 2. ver [6]. Para 3. y 4. ver [14] y [15]. □

Ejemplo 1.3.6 (Baricentro de normales)**.** Si en el problema del baricentro de Wasserstein (1.15) $\mu_i = \mathcal{N}(m_i, \Sigma_i)$, entonces el baricentro es una normal cuya media m es el promedio de las medias ($m = \sum w_i m_i$) y su varianza es la única matriz Σ simétrica definida positiva que cumple:

$$\sum w_i \left(\Sigma^{1/2}\Sigma_i\Sigma^{1/2}\right)^{1/2} = \Sigma. \tag{1.18}$$

Así, el baricentro es $\overline{\mu} = \mathcal{N}(m, \Sigma)$. Para probar esto hay que ver primero que existe tal Σ y luego que dicha normal es óptima. Seguiremos la demostración original de Agueh y Carlier de [6].

- Tomemos $\alpha \leq \left(\sum_i^n w_i\sqrt{\alpha_i}\right)^2$ y tomemos $\beta \geq \left(\sum_i^n w_i\sqrt{\beta_i}\right)^2$ donde α_i es el autovalor más chico de Σ_i y β_i el más grande. Llamemos K al conjunto (convexo y compacto) de todas las matrices simétricas Σ (y definidas positivas) cuyos autovalores están entre $[\alpha, \beta]$, es decir $\alpha\mathbb{I} \leq \Sigma \leq \beta\mathbb{I}$, donde $\mathbb{I}$ representa la matriz identidad en el espacio correspondiente. Si definimos la función de la cual buscamos el punto fijo $F(\Sigma) := \sum_{i=1}^n w_i \left(\Sigma^{1/2}\Sigma_i\Sigma^{1/2}\right)^{1/2}$ y vemos que va de K en K, por el Teorema del punto fijo de Brouwer tenemos la existencia de un punto fijo y por ende de la Σ buscada. Por como definimos α y β, es claro que $\alpha\mathbb{I} \leq \sum_{i=1}^n w_i\sqrt{\alpha\alpha_i}\mathbb{I} \leq F(\Sigma) \leq \sum_{i=1}^n w_i\sqrt{\beta\beta_i}\mathbb{I} \leq \beta\mathbb{I}$ por lo que F va de K en K donde este es un compacto convexo y tenemos la existencia de una única Σ simétrica definida positiva que cumpla lo pedido.

- Lo haremos en el caso en que todas las normales tienen media 0 ya que simplifica las cuentas, luego el caso general se deduce de tan solo aplicar traslaciones sobre esto. Sabemos que existe tal Σ, y recordemos que cada transporte óptimo de $\overline{\mu}$ a μ_i es $T_i(x) = A_i x$ con $A_i = \Sigma_i^{1/2}\left(\Sigma_i^{1/2}\Sigma\Sigma_i^{1/2}\right)^{-1/2}\Sigma_i^{1/2}$ como vimos en el ejemplo de transporte entre normales (1.11) ya que estamos tomando la inversa del transporte de μ_i a $\overline{\mu}$. Por la Proposición 1.3.5 basta ver que $\sum_{i=1}^n w_i T_i = \mathrm{Id}$. Pero $\sum_{i=1}^n w_i T_i =$

$\sum_{i=1}^n w_i \Sigma_i^{1/2} \left(\Sigma_i^{1/2} \Sigma \Sigma_i^{1/2}\right)^{-1/2} \Sigma_i^{1/2}$ y el que esta última expresión sea Id se deduce de que Σ era el único punto fijo de F. Para ver esto, usando que $F(\Sigma) = \Sigma$ obtenemos $\Sigma = F(\Sigma) = \sum_{i=1}^n w_i \left(\Sigma^{1/2} \Sigma_i \Sigma^{1/2}\right)^{1/2} = \sum_{i=1}^n w_i \Sigma^{1/2} \Sigma_i^{1/2} \left(\Sigma_i^{1/2} \Sigma \Sigma_i^{1/2}\right)^{-1/2} \Sigma_i^{1/2} \Sigma^{1/2}$ y podemos usar que Σ es invertible por ser definida positiva y multiplicando a izquierda y derecha por $\Sigma^{-1/2}$ obtenemos $\sum_{i=1}^n w_i T_i = \sum_{i=1}^n w_i \Sigma_i^{1/2} \left(\Sigma_i^{1/2} \Sigma^{1/2} \Sigma_i^{1/2}\right)^{-1/2} \Sigma_i^{1/2} =$ $= \mathrm{Id}$, que es lo que queríamos.

En la siguiente figura podemos ver el baricentro de distintas normales.

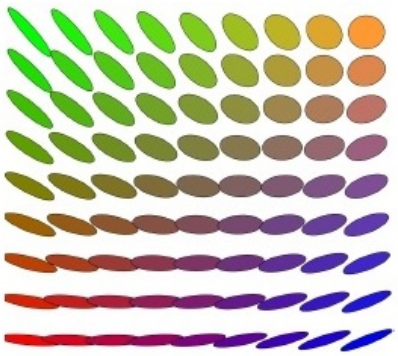


Figura 1.6: Baricentros de las cuatro normales de las esquinas con diferentes pesos w_i. Cada normal 2-dimensional se representa como una elipse alineada con los autovectores de la matriz de covarianza y estirada proporcionalmente a los autovalores. Figura 9.2 de [1].

Nuevamente, encontrar un baricentro puede tener varias aplicaciones. La más inmediata es encontrar a partir de ciertas formas geométricas, otras que sean intermedias, es decir sus baricentros. Un ejemplo de esto es lo que muestra la Figura 1.7 que da baricentros con distintos pesos w_i para las tres distribuciones que se ven en los vértices del triángulo. Cada distribución es una uniforme soportada en la forma geométrica que representa, o sea constante en la forma y 0 fuera de ella. Como los transportes van trasladando las distribuciones y deformándolas a la vez, los baricentros ente dos formas van a encontrarse espacialmente en la recta que las une y van a representar formas intermedias. Si utilizáramos aquí un simple promedio como distribución intermedia no podríamos distinguir formas con sentido, que visualmente tengan componentes de cada una de las formas de las cuales estamos tomando baricentro.

Figura 1.7: Baricentros para formas en tres dimensiones. Imagen de Gabriel Peyré, Figura 7 de [18]

Comentario 1.3.7. En general utilizaremos como costos distancias elevadas al cuadrado. En principio esto no es sólo porque el Teorema 1.3.2 nos da una solución para el caso de la distancia al cuadrado, sino también porque se trata de un costo estrictamente convexo que nos dará única solución, a diferencia de la distancia que es solamente convexa. Un ejemplo claro y sencillo de esto sería tomar dos distribuciones discretas que concentren toda su probabilidad en un punto. Si consideramos un costo de la forma $c(x, y) = ||y - x||$, cualquier punto que se encuentre en el segmento entre x e y es solución del problema del baricentro con pesos $w_i = 1/2$, mientras que si consideramos $c(x, y) = ||y - x||^2$ habrá unicidad de solución y esta será el punto medio como es deseado. Es análogo a por qué utilizamos $||\cdot||^2$ al definir el baricentro de puntos en (1.14). Si el costo define una distancia y no incluimos el cuadrado, dará lo mismo elegir cualquier punto en el segmento o geodésica que los une. Es por esto que en el problema del baricentro de Wasserstein al utilizar como costo la distancia de Fermat (que definiremos en el siguiente capítulo) o cualquier distancia en general, lo haremos elevándola al cuadrado.

Capítulo 2

Distancia de Fermat

Como hemos visto, el caso más estudiado de costo en transporte óptimo es el costo dado por la norma euclídea al cuadrado, que mide la distancia entre dos puntos. Sin embargo, la distancia euclídea puede ser una mala distancia a considerar. Un ejemplo es en el caso de puntos soportados en alguna superficie y más aún si esta es de dimensión baja pero contenida en un espacio de muy alta dimensión. Más aún, en altas dimensiones la distancia euclídea L_2 es poco significativa, trayendo varios problemas: este hecho es conocido como la maldición de la dimensionalidad ("Curse of Dimensionality"). Entre tantos problemas que traen las altas dimensiones, se encuentra el hecho de que para poblar una región se necesitan cada vez más puntos (que crecen exponencialmente). Por ejemplo, esto hace que los vecinos más cercanos no sean significativos por estar muy lejos. Típicamente estando en varias dimensiones los datos se encuentran en alguna superficie de dimensión mucho menor; allí puede ser una mala distancia la del espacio ambiente ya que no tiene en cuenta la superficie que explica los datos.

La distancia de Fermat [8] permite utilizar el conocimiento que uno tiene de alguna superficie dados puntos en ella; más aún, tiene en cuenta la densidad de estos. Esto es fundamental porque refleja una distancia que tiene en cuenta la superficie: es una geodésica pesada según la densidad. Para definir la distancia, definámosla primero en el caso discreto, o sea en el caso en que tenemos puntos solamente. La idea será observar que esta distancia discreta es un estimador consistente de una distancia continua que tiene en cuenta, como dijimos, la densidad y la superficie.

Definición 2.0.1 (Distancia de Fermat discreta)**.** Dado un conjunto de puntos $Q = \{q_i\}_{i=1}^{K}$, la distancia de Fermat entre dos de esos puntos se define como

$$D_{Q,\alpha}(p,q) = \min_{\gamma=(p=q_0,q_1,q_2\ldots,q_k=q)} \sum_{i=0}^{k-1} |q_{i+1} - q_i|^{\alpha}, \tag{2.1}$$

o sea la longitud del camino mínimo γ en el grafo completo de los puntos de Q donde las aristas tienen como peso las distancias euclídeas entre ellos pero elevadas a la α.

Observación 2.0.2. Notemos que la distancia de Fermat define una distancia a partir de otra distancia de base; en este caso está dada por la distancia euclídea $|\cdot|$, pero podría basarse en cualquier distancia $\tilde{d}$. Esa $\tilde{d}$ es la que determinará el peso de las aristas del grafo completo de vértices q_i.

En el caso $\alpha = 1$ se trata de la distancia euclídea, ya que el camino que minimizará es el camino directo entre p y q (cosa que sucede también con $\alpha < 1$). Al variar α esta distancia provee una noción de cercanía distinta entre los puntos. Si $\alpha > 1$, para minimizar (2.1) se evitará que $|q_{i+1} - q_i|$ sea grande; cuanto más grande sea α, más importancia tendrá esto. En superficies contenidas en alta dimensión esto es deseable ya que las distancias lejanas no son significativas y buscamos trabajar con puntos que se encuentran cerca y reflejan bien a la superficie. El buscar moverse por caminos donde se dan saltos pequeños, es moverse por zonas donde hay muchos puntos y por ende la densidad es más alta; la importancia que se le da a pasar por zonas de densidad alta esta dada por α.

Observación 2.0.3. La distancia de Fermat (2.1) define una distancia en el conjunto de vértices del grafo completo Q siempre y cuando el peso de las aristas sea una distancia (que llamaremos la distancia de base).

Si bien definimos la distancia de Fermat para puntos que estén en el grafo, es natural preguntarse cómo se define para puntos fuera del mismo. En muchas ocasiones cuando analicemos un punto, puede ser conveniente incluirlo en el grafo directamente y utilizar la definición previa. Otra opción es utilizar la siguiente definición:

Definición 2.0.4. La distancia de Fermat de la Definición 2.0.1 se generaliza a dos puntos cualesquiera p y q se define como la distancia entre $\tilde{p}$ y $\tilde{q}$ que son los puntos del grafo más cercanos (en distancia euclídea) a p y q respectivamente.

Hemos definido entonces una distancia discreta, que la podemos pensar como una versión muestral de una distancia. Previo a dar la versión continua o poblacional de dicha distancia, motivaremos su definición buscando informalmente un objeto límite al que tienda la distancia muestral.

Si suponemos que los puntos son muestras obtenidas de forma i.i.d según una distribución de densidad f en una superficie $\mathcal{M}$, queremos que, cuando la cantidad de puntos crezca, la distancia converja a alguna distancia continua que refleje no solo las propiedades de la superficie sino también la densidad f. Llamemos q_i a los puntos intermedios de un camino óptimo. Conceptualmente, si la superficie esta muy poblada, tenemos de manera local que $|q_{i+1} - q_i|^{\alpha-1} \approx 1/f^\beta$ con $\beta = (\alpha - 1)/d$ donde d es la dimensión de $\mathcal{M}$. Esto se deduce de pensar a los puntos como un Proceso Puntual de Poisson (PPP) de intensidad f que, bajo hipótesis de continuidad, podemos pensar localmente constante pues los puntos del camino consecutivos están cerca. El pensar un conjunto de puntos i.i.d como un PPP es válido acotando probabilidades de que no sea Poisson con grandes desvíos. Primero, salvo constantes multiplicando, $|q_{i+1} - q_i|^d$ crece como el volumen de la menor bola que contiene en su borde a q_{i+1} y q_i. Luego, salvo constantes, $|q_{i+1} - q_i|^d f$ representa la cantidad esperada de puntos que hay en dicha bola, ya que es la esperanza de una Poisson de parámetro f multiplicada por medida del conjunto. A su vez, esperamos que esta cantidad sea 2 puntos, ya que, si hubiese 3, convendría utilizar el tercer punto que hay en la bola como punto intermedio en el camino, lo que contradice la suposición de que q_i y q_{i+1} eran puntos consecutivos en el camino. Dejando de lado las constantes, podemos pensar que $|q_{i+1} - q_i|^d f \approx 1$ de donde se deduce que $|q_{i+1} - q_i|^{\alpha-1} \approx 1/f^\beta$. Por último, si pensamos que la suma que define la distancia de Fermat (2.1) es la discretización de una integral de línea y separamos del $|q_{i+1} - q_i|^\alpha$ un $|q_{i+1} - q_i|$ para que juegue el papel de diferencial de longitud, tendremos la integral de línea sobre la superficie $\mathcal{M}$ de $1/f^\beta$. Ahora sí, tenemos la intuición de que la distancia de Fermat discreta converge a una distancia de Fermat continua que definimos a continuación.

Definición 2.0.5 (Distancia de Fermat continua)**.** Dados dos puntos $x, y \in S$, una función f positiva y un $\beta \geq 0$, se define la distancia de Fermat entre x e y como

$$\mathcal{D}_{f,\beta}(x, y) = \inf_\gamma \int_\gamma 1/f^\beta, \tag{2.2}$$

donde el ínfimo es sobre todos los caminos γ de x a y contenidos en $\overline{S}$. El ínfimo está dado por una geodésica que minimiza la integral de línea propuesta, que no es más que una geodésica pesada según la densidad.

Exactamente por esto la distancia se llama distancia "de Fermat": en similitud al principio de óptica en que las trayectorias se darán por zonas con menor índice de refracción, aquí se buscará ir por regiones de mayor densidad. El papel del índice de refracción lo juega $1/f^{\beta}$. El buscar ir por regiones de mayor densidad indica que, según el α, se recorrerá más distancia por una zona con mayor densidad a cambio de minimizar la distancia recorrida en una zona de menor densidad. Esto se puede ver en la Figura 2.1 donde se busca ir por regiones de mayor densidad y se observa que a mayor α es más importante tomar caminos que pasen por zonas de alta densidad que caminos cortos en su longitud total. En la versión discreta, pasar por zonas de alta densidad se corresponde con que el camino de saltos pequeños.

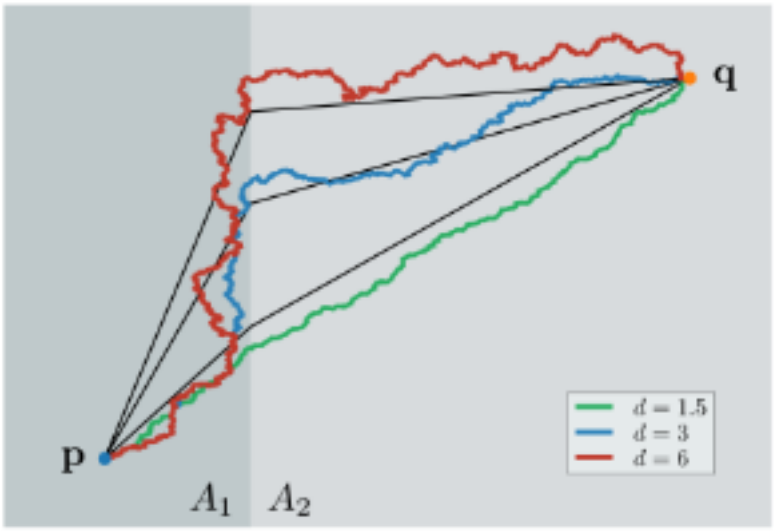


Figura 2.1: Caminos que realizan la distancia de Fermat discreta (en color) y continua (en negro) entre p y q para $\alpha = 1{,}5, 3, 6$. Se sortearon uniformes en dos rectángulos A_1 y A_2, la densidad en la zona gris oscura es el doble que en la gris clara. Figura de Facundo Sapienza, Pablo Groisman y Matthieu Jonckheere. Figura 3 de [8].

Lo que hemos introducido muy informalmente para deducir la distancia continua se formalizó en [7] dando lugar al Teorema 2.0.6, que establece que la distancia de Fermat discreta (2.1) es un estimador consistente de la distancia de Fermat continua (2.2).

Teorema 2.0.6 (Consistencia de la distancia de Fermat)**.** *Sea $\mathcal{M} \subseteq \mathbb{R}^D$ isométrico a un abierto conexo S con bordes $\mathcal{C}^1$ de $\mathbb{R}^d$ y sea $f : \mathcal{M} \longrightarrow \mathbb{R}_+$ una función de densidad continua. Sean $Q_n = \{q_1, \ldots, q_n\}$ n puntos i.i.d con densidad f. Para $\alpha > 1, x, y \in \mathcal{M}$ se tiene que*

$$n^{\beta} D_{Q_n,\alpha}(x, y) \longrightarrow \mu \mathcal{D}_{f,\beta}(x, y) \qquad c.s, \tag{2.3}$$

con $\beta = \dfrac{\alpha - 1}{d}$ y μ una constante que depende solo de α y d. Más aún: si $\mathcal{D}_{f,\beta}(x, y)$ se alcanza con una única geodésica γ^*, entonces los caminos γ_n que realizan el mínimo en $D_{Q_n,\alpha}$ convergen uniformemente a γ^* casi seguramente.

Demostración. Referimos a la demostración original de [7], donde es el Teorema 2.3. □

Comentario 2.0.7 (Implementación)**.** La distancia de Fermat discreta (2.1) dos a dos entre los puntos de $Q = \{q_i\}_{i=1}^n$ es una matriz que contiene las longitudes de los caminos mínimos de un vértice a otro en el grafo completo dado por los vértices q_i donde consideramos la longitud de las aristas pero elevadas a la α. El algoritmo de camino mínimo de Floyd-Warshall nos permite obtener todas las distancias en una cantidad de operaciones $\mathcal{O}(n^3)$. Esta complejidad puede ser muy costosa para grandes conjuntos de datos y se puede estimar considerando el grafo de k-vecinos más cercanos en vez del grafo completo. Al correr el algoritmo de Dijkstra sobre cada uno de los vértices, repetiremos n veces un algoritmo que implementado con una "priority queue" es $\mathcal{O}(kn + nlog(n))$ ya que la cantidad de aristas del grafo de k-vecinos más cercanos es kn, reduciendo la complejidad a $\mathcal{O}(n^2 log(n))$ si la cantidad de vecinos se toma como $k \propto log(n)$. Los autores mostraron que también en este caso se obtiene un teorema de consistencia (ver Proposición 2.12 de [7]). Por último, proponen como método más rápido correr Dijkstra no para los n punto sino sólo para $m << n$ "landmarks" y a partir de esos caminos acotar por arriba y por abajo las distancias usando la desigualdad triangular; esto da una complejidad de $\mathcal{O}(mnklog(n))$.

Observación 2.0.8. Si consideramos la distancia definida por considerar los k-vecinos más cercanos como se comentó en la anterior observación y tomamos $\alpha = 1$, se recupera la distancia geodésica de Isomap [9].

Esta distancia presenta numerosas aplicaciones potenciales en casos donde los datos están soportados en una superficie que expresa mucho mejor su geometría que el espacio

ambiente. Esto puede ser muy útil por ejemplo para el problema de clusterización, que consiste en encontrar grupos ("clusters") en un conjunto de puntos. Podemos ver en la siguiente figura que al resolver "k-means" con esta distancia se pueden encontrar clusters donde con la euclídea no.

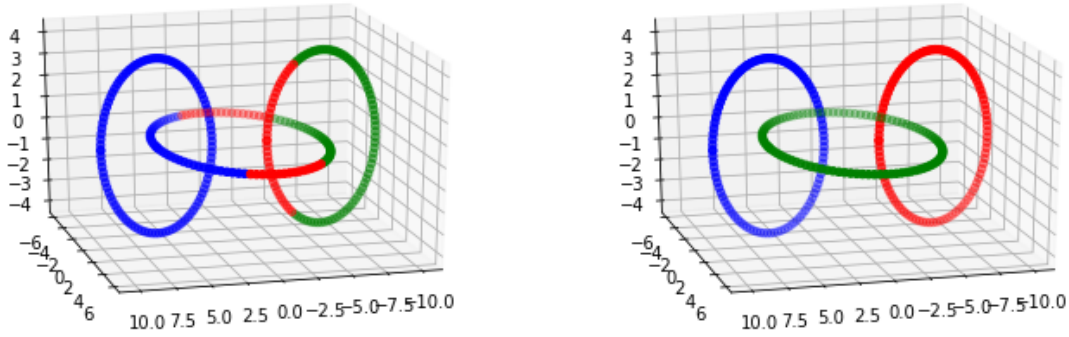


Figura 2.2: Clusters obtenidos al resolver k-means utilizando la distancia euclídea (izquierda) y la distancia de Fermat con $\alpha = 2$ (derecha).

En la Figura 2.2 vemos la ventaja de que sea una distancia geodésica, pero no solo esa es una gran ventaja sino el hecho de que considere también la densidad. Así, con la distancia de Fermat, un punto $\overline{x}$ puede estar más cerca de un x_2 que de un x_1, por más que x_1 esté a menor distancia geodésica, como ejemplifica la siguiente figura.

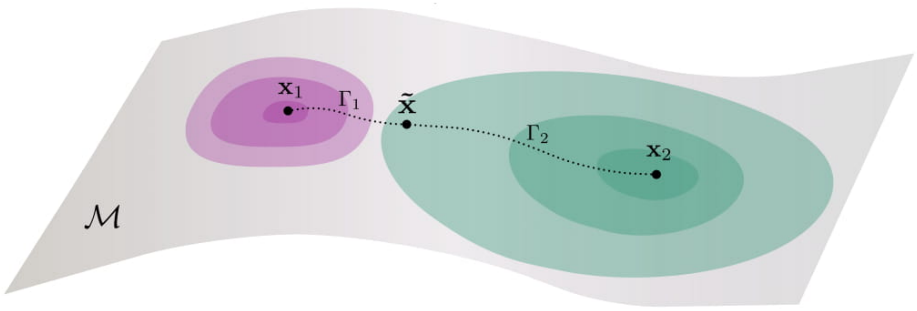


Figura 2.3: Vemos en la imagen una superficie $\mathcal{M}$, que podemos imaginar de dimensión mucho menor que la del espacio ambiente. Los colores representan una densidad que está soportada en $\mathcal{M}$ y que tiene dos modas alrededor de x_1 y x_2. Las zonas en gris son de muy baja densidad. Si bien $\overline{x}$ está mas cerca en distancia geodésica de x_1 que de x_2 (es decir que $long(\Gamma_1) < long(\Gamma_2)$), en distancia de Fermat se encuentra mas cerca de x_2 que de x_1. Esto es porque cualquier curva que una a $\overline{x}$ con x_1 debe atravesar una zona donde f es muy pequeña por lo que integrar su inversa será muy grande, mientras que hay curvas que unen a $\overline{x}$ con x_2 y no atraviesan zonas de muy baja densidad. Imagen de Facundo Sapienza [24].

Los resultados teóricos que conciernen a la distancia de Fermat se han generalizado a variedades que no son isométricas a un abierto conexo de $\mathbb{R}^d$ en [31]. Allí se muestran resultados en aplicaciones de topología computacional, particularmente en homología persistente.

Capítulo 3

Transporte Óptimo y Baricentro con distancia de Fermat

En el primer capítulo definimos el problema de transporte óptimo, baricentro de Wasserstein y algunos resultados de la teoría desarrollada en torno a estos. En el segundo capítulo hemos definido la distancia de Fermat y su estimador discreto. Teniendo en cuenta las ventajas mencionadas en el capítulo anterior de utilizar una distancia que considere la superficie en la que se encuentran los datos y su densidad, se trata de un gran candidato para reemplazar al costo euclídeo en el problema de Transporte Óptimo. Este capítulo da formulaciones de la discretización de los problemas teóricos de transporte óptimo y baricentro de Wasserstein en general y con la distancia de Fermat como costo en particular, define algoritmos para resolverlos computacionalmente y muestra sus resultados. Dichos algoritmos pueden encontrarse implementados en `https://github.com/nicocheh/FermatOT`.

Si en la formulación de Monge del problema de transporte óptimo

$$\min_{T_{\#}\mu=\nu} \int c(x, T(x)) d\mu, \tag{3.1}$$

utilizamos la distancia de Fermat como función de costo, podemos naturalmente obtener transportes y muestras en la superficie, y no por restricciones adicionales impuestas. Esto sucederá porque la misma función de costo nos llevará a poner puntos en la variedad, puesto que estar lejos de la superficie aumentará el costo. La interpolación de McCann (1.12) utiliza la naturaleza euclídea (o de la distancia de base) al realizar una combinación convexa moviendo las distribuciones por líneas rectas al variar el t además de deformarlas;

si quisiéramos generalizarlo a alguna superficie, deberíamos imponer que esa combinación se haga sobre las geodésicas. Si bien no podemos generalizar fácilmente la interpolación de McCann (1.12) a la distancia de Fermat, utilizar esta va a forzar que los puntos "viajen" por la superficie ya que los caminos que minimicen la distancia estarán contenidos en ella, si está densamente poblada.

Más allá del costo que utilicemos, cuando queremos llevar el problema de transporte óptimo (3.1) a una formulación muestral, tenemos un conjunto de puntos $\{x_i\}_{i=1}^{i=N_x}$ sampleados como muestras i.i.d de la distribución μ que queremos transportar en puntos $\{y_j\}_{j=1}^{j=N_y}$ sampleados como muestras i.i.d de la distribución ν. Se pueden tomar dos enfoques: el primero es buscar qué asignación de x_i a $y_{\sigma(i)}$ minimiza el costo de transporte como en (1.6) y el segundo es pensar que los puntos transportados $T(x_i)$ no tienen por qué ser y_j. Esta última formulación permite no solo obtener muestras distintas de puntos distribuidos como los y_j, sino que además no impone la restricción de tener igual cantidad de puntos x_i que puntos y_j. Por estos motivos nos centraremos en la segunda, aunque una breve descripción y exploración con la primera puede encontrarse en la Sección 3.3.

En el caso de dar total libertad a los $T(x_i)$ (deseamos que caigan en la superficie, pero para eso ayudará usar el costo de Fermat), reformular la integral del problema poblacional es tan simple como pasarlo a sumatorias, donde llamaremos $\tilde{x_i} = T(x_i)$ para simplificar. Así, la función objetivo a minimizar pasa a ser:

$$\sum_{i=1}^{N_x} c(x_i, \tilde{x_i}),$$

donde nos queda imponer la condición de que $T_{\#}\mu = \nu$, cuya versión muestral es que $\{\tilde{x_i}\}_{i=1}^{i=N_x}$ "tenga la misma distribución" que $\{y_i\}_{i=1}^{i=N_y}$. Esta condición muestral puede enfocarse para expresar esto a partir de la divergencia de Kullback-Leibler (que se puede formular como un problema de maximización [26] [27]) o a través de "feature functions".

Para este último enfoque, podemos recordar que X e Y tienen la misma distribución sii $\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$ para toda función f en el conjunto de funciones continuas y acotadas (podría ser en otra clase que lo garantice). Podemos entonces tomar un subconjunto $\mathcal{F}$ de esas funciones e imponer para estas dicha condición en su versión muestral, o sea

$$\frac{1}{N_x}\sum_{i=1}^{N_x} f(\tilde{x_i}) = \frac{1}{N_y}\sum_{i=1}^{N_x} f(y_i), \quad \forall f \in \mathcal{F}.$$

Estas funciones $f \in \mathcal{F}$ serán las que pedirán que la distribuciones se parezcan en ciertas "features" o características.

Algunos ejemplos de estas f podrían ser:

- Funciones lineales, que igualarán primer momento.
- Funciones cuadráticas, que igualarán el segundo momento.
- Funciones kernel, que tratarán de que la densidad alrededor de ciertos centros sea la misma. Estas se describirán en detalle más adelante en (3.4).

Así, esto puede formularse como un problema de minimización con restricciones de igualdad (o relajaciones a desigualdad con una tolerancia ε) o puede formularse como un problema de minimización por parte de la función de transporte y maximización por parte de la "feature function" (o divergencia Kullback-Leibler). También puede considerarse un enfoque en el que lo que se busca encontrar es la función T de transporte, parametrizándola por ejemplo, y no los $\tilde{x}_i$. Si bien esto es más difícil, permitiría poder transportar puntos que no son los x_i directamente.

Estas ideas y formulaciones para la resolución del problema discreto aplican para la resolución de transporte óptimo en general, más allá del costo que se use. Resulta muy útil conocer el gradiente de la función de costo ya que nos permitirá utilizar técnicas como descenso por el gradiente para minimizar la función objetivo.

En principio, un problema inicial sería considerar restricciones de igualdad tanto de primer como segundo momento para una función objetivo con un costo que sea la distancia de Fermat. Para abordar este problema hay aún varias decisiones a tomar con esta distancia, que detallamos en la siguiente sección.

3.1. Definiendo Detalles de Fermat

Para definir la distancia de Fermat discreta de (2.1), uno podría utilizar tanto los puntos x_i como los puntos y_i, considerando que pueden estar soportados en la misma superficie y nos dan conocimiento sobre esta. También uno puede tener previo conocimiento

de esta superficie, con más puntos que llamaremos s_i. Para tratar de poner en contexto de minimización con restricciones a la distancia de Fermat, debemos definir rigurosamente:

- con qué puntos se definirá el grafo de la distancia de Fermat. discreta (2.1)
- cómo se definirá para puntos fuera de la muestra.

Y, fundamentalmente:

- cuál es su gradiente, lo que es necesario para casi cualquier técnica de minimización.

En principio, precomputaremos la distancia de Fermat en algún conjunto S de puntos s_i que típicamente contendrá también a los x_i y a los y_i, pero podría no hacerlo.

Para definir la distancia de Fermat entre $s_0 \in S$ y z con los puntos S, precomputamos la distancia de Fermat en S y luego definimos la distancia como la distancia de Fermat con la superficie representada obtenida al agregar z a S, es decir $c_{S\bigcup\{z\}}(s_0, z)$ que por simplicidad denotaremos $c_S(s_0, z)$. El problema de esto último es que recomputar toda la distancia de Fermat es computacionalmente muy caro y más aún si S consta de muchos puntos. Podríamos tomar la Definición 2.0.4 que es computacionalmente barata y considerar al punto y como su vecino inmediatamente más cercano. Esto es una buena aproximación si tenemos muchos puntos s_i que dan lugar a una superficie muy poblada, que no tiene por qué ser el caso de los datos.

Una propuesta intermedia es considerar tan solo una cantidad k de vecinos más cercanos de z que se encuentren en S. Podemos pensar que lo que hacemos es utilizar la aproximación de considerar los k-vecinos más cercanos de z según se indicó en la Observación 2.0.7 que, como se menciona allí, es también una estimación consistente de la distancia de Fermat continua (2.2). Si llamamos a estos vecinos $s_1, \ldots, s_k$, extendemos la definición como:

$$c_S(s_0, z) = \min_{i=1,\ldots,k} c_S(s_0, s_i) + |s_i - z|^\alpha,$$

donde c_S del lado derecho es la distancia de Fermat discreta (2.1) con $Q = S$ y con un α que omitimos en la notación. En principio en caso que los x_i sean parte de S, tenemos definida la función de costo que precisamos para transporte óptimo que es $c(x_i, \tilde{x_i})$. Si no son parte de S, habría que redefinirla en caso de que los dos puntos no pertenezcan a S, lo que se puede hacer de forma totalmente análoga.

Podríamos repetir lo mismo para ambos y buscar el par de k vecinos más cercanos (un vecino del punto inicial y otro del final) que minimicen la distancia de Fermat. Esto aumenta el costo en complejidad de computarla, ya que una vez elegidos los k vecinos de cada uno, hay que computar la mejor pareja que es $\mathcal{O}(k^2)$ en vez de $\mathcal{O}(k)$. Otra opción sería considerar el vecino más cercano del punto inicial y aplicar lo antes comentado donde un solo punto no pertenece a S. Esta es quizás la más esperanzadora ya que no requiere mucho cómputo y porque típicamente esperamos que las muestras x_i de la distribución fuente estén en la superficie S. Si n es la cantidad de puntos de S, podemos obtener en tiempo logarítmico $\mathcal{O}(klog(n))$ los k-vecinos más cercanos habiendo construido previamente un KD-Tree [11] [12] con los puntos de S, lo que se hace en tiempo lineal $\mathcal{O}(n)$. Es una importante propiedad que el cómputo de los costos sea veloz una vez realizado todo el precomputo de las distancias de Fermat y armado del árbol, ya que después optimizaremos sobre esta función llamándola varias veces. Si bien los KD-Tree pierden eficiencia en altas dimensiones con pocos datos debido a la maldición de la dimensionalidad, hay algoritmos llamados ANN, ("Approximate Nearest Neighbours") [13] que implementan una búsqueda de vecinos aproximada pero más rápida, que podrían utilizarse en un contexto de altas dimensiones y pocos datos.

Estas decisiones son quizás bastante razonables a partir de conocer la distancia de Fermat discreta (2.1) y querer generalizarla. Lo que quizás no es fácil de ver es cómo definir una estimación del gradiente para esta distancia que aunque sea tenga cualidades básicas deseables en el contexto de optimización.

3.1.1. El gradiente de Fermat

Antes que nada, cabe destacar que diferenciar la distancia de Fermat continua (2.2) en una superficie $\mathcal{M}$ no es inmediato. En principio deberíamos imponer alguna condición sobre $\mathcal{M}$ para asegurar que pequeños cambios en el punto final no produzcan grandes cambios en las geodésicas. En una variedad con una métrica, las geodésicas pesadas de Fermat son geodésicas si cambiamos la métrica multiplicándola por $1/f^\beta$. Bajo esta métrica adecuada, podemos pensar que la distancia de Fermat es la distancia geodésica de la variedad. En general, al diferenciar la distancia geodésica al cuadrado $d_p(x) = d(p, x)^2$ desde un punto

fijo p, obtendremos $-2exp_x^{-1}(p)$, donde exp_x es la función exponencial en el punto x, que conecta localmente el espacio tangente con la variedad. Es decir, dado un vector v del plano tangente en x, $exp_x(v)$ nos dará un punto de la única geodésica que pasa por x y tiene velocidad v. Lo que nos interesa es que $-2exp_x^{-1}(p)$ apunta en la dirección de la geodésica de p a x y con norma que, salvo el 2, va como la velocidad de dicha curva. Así, la dirección en la que apunta el gradiente de la distancia será también la misma.

Al calcular el gradiente discreto, pondremos principal foco en que la dirección sea la correcta, ya que la norma puede ser corregida por el "learning rate" de un algoritmo de descenso por el gradiente. Para computar el gradiente comenzaremos con el caso en que el punto inicial está fijo y pertenece a S y el punto final es un punto que no pertenece a S: esto es lo necesario para definir el gradiente utilizado en transporte óptimo en caso que la distribución source (los x_i) pertenezca a S. En caso que no, consideramos su vecino más cercano y el computo del gradiente será análogo pero a partir de ese punto.

Definir el gradiente en la versión poblacional es una tarea bastante complicada y que sin duda requiere conocimiento de la superficie como hemos visto, ya que precisa la aplicación exponencial. En cambio, si nos basamos en la distancia muestral, esta labor se torna más accesible. En principio si el camino no cambiara, es muy fácil derivar y obtener que el gradiente en un punto z será:

$$\alpha|q-z|^{\alpha-2}(q-z), \tag{3.2}$$

donde q es el anteúltimo punto en el camino mínimo de Fermat al incluir a z en S. Si bien esta parece ser la elección más directa, sensata y simple del gradiente, conlleva varios problemas:

- Cuando z está muy cerca de un punto de la muestra, el gradiente se acerca a 0. Más aún, el gradiente vale 0 en los puntos de la muestra. Esto es un claro problema ya que querríamos que la dirección opuesta al gradiente del costo de ir de w_0 a z nos vaya llevando a través del camino hacia w_0 y que valga 0 al llegar a w_0 que es el mínimo global de la función (y en muchas superficies el único mínimo).

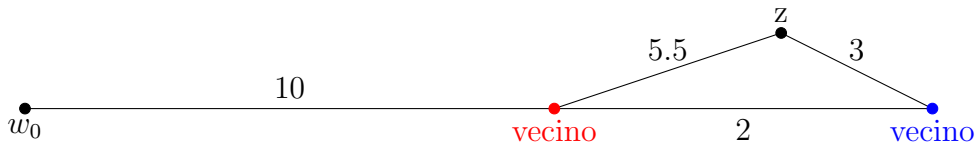
- Este gradiente se basa fuertemente en el hecho de que el camino óptimo no cambia. Si al mover z un poco el camino óptimo sigue siendo el mismo, este gradiente es el

acertado, pero no lo es en el caso que cambie el camino. Típicamente, uno espera que al optimizar e ir moviendo los puntos decida tomar nuevos caminos ya que no es claro de antemano que tras mover un poco el punto de su geodésica (cosa que numéricamente sucederá) el camino siga siendo el mismo.

Proponemos como gradiente uno que busque la dirección de máximo crecimiento en un entorno. La "localidad" del gradiente estará dada justamente por ese entorno. Para elegir el gradiente, nos centraremos en que su opuesto nos de una dirección de descenso. Para ello, podemos considerar los k vecinos más cercanos y quedarnos con el que tenga menor costo de Fermat de ellos. Una vez elegido ese vecino, computamos el gradiente anteriormente mencionado, donde en (3.2) ponemos a tal vecino en lugar de q. Luego, a partir de esta definición del gradiente, sabemos que $-\nabla$ nos dará la dirección de máximo decrecimiento entre los k vecinos más cercanos. Ahora este gradiente logra resolver algunos de estos problemas:

- Al acercarse a un punto de S, el gradiente no se anula ya que apuntará hacia otro vecino.

- En el siguiente dibujo, este gradiente apuntará hacia el vecino rojo, que tiene costo 10, en vez de hacia el vecino azul, que tiene costo 12. Notemos que el camino que pasa por el punto azul y llega a z tiene menor costo total que el que pasa por el punto rojo y llega a z. Podemos ver que el gradiente propuesto apunta en dirección opuesta a un vecino que no necesariamente es un punto del actual mejor camino pero que sí va a hacer descender la función de costo si nos movemos en el sentido opuesto al gradiente. De hecho, tras un paso lo suficientemente largo, la hará descender más que si nos moviésemos hacia el vecino azul. Notemos que si se trata de una superficie con muchos puntos, tomar una cantidad fija (no muy grande) de vecinos sigue siendo moverse en un entorno local. En resumen, estamos definiendo el gradiente como la dirección de máximo crecimiento en un entorno, donde ese entorno es mirar los vecinos más cercanos del punto en donde queremos diferenciar. Si la superficie está poblada esperamos que, al mirar varios vecinos, el gradiente nos vaya llevando hacia el origen del camino w_0.



3.2. Optimización con Restricciones

Sea cual sea el costo, podemos plantear un problema de minimización con restricciones

$$\begin{aligned} \min_{\tilde{x}_i} \quad & \sum_{i=1}^{N_x} c(x_i, \tilde{x}_i) \\ \text{s.a.} \quad & \frac{1}{N_x} \sum_{i=1}^{N_x} f(\tilde{x}_i) = \frac{1}{N_y} \sum_{i=1}^{N_x} f(y_i), \quad \forall f \in \mathcal{F}. \end{aligned} \tag{3.3}$$

Aquí $\mathcal{F}$ es el conjunto de las "feature functions" y s.a. significa "sujeto a", que indica cuales son las restricciones que imponemos a la minimización. Siempre que conozcamos el gradiente de la función de costo $c(x, y)$ podremos resolver con mayor facilidad este problema utilizando alguna técnica de minimización que aproveche el gradiente, como por ejemplo descenso por el gradiente. En el caso euclídeo computar dicho gradiente es muy simple ya que este será $2(y - x)$ y en el caso de Fermat ya hemos definido el gradiente; podemos entonces resolver el problema (3.3) para ambos costos.

El problema (3.3) se puede resolver con un solver no lineal; en este caso se usó [21]) para hallar el resultado del trasporte óptimo con el costo euclídeo y con el costo de Fermat. Se dejan correr varias iteraciones (a elegir por el usuario) hasta no haber una clara mejoría o cambio en la solución propuesta. El criterio de parada no es tan claro ya que la elección que hacemos del gradiente hace que este no se anule, salvo en el mínimo trivial que generalmente no cumplirá las restricciones. Es también importante mencionar que se puede dar cierta tolerancia en cumplir las restricciones ("constraints") ya que al tratarse de una muestra nunca esperamos que coincidan exactamente las $\mathbb{E}[f(\cdot)]$. Dependiendo del problema y aplicación puede no ser fundamental encontrar un mínimo local sino aproximarse a este como también puede no ser tan importante que las restricciones se cumplan exactamente.

Utilizando solo como "feature functions" los primeros dos momentos, se realizaron varios experimentos para ver el comportamiento de las soluciones y del gradiente. En todos los

casos se utilizó $\alpha = 2$ para la distancia de Fermat discreta y se consideraron $k = 15$ vecinos para el cálculo del gradiente.

Doble puente

Los puntos s_i son los generados por dos normales y dos puentes que los unen, como se puede ver en la Figura 3.1.

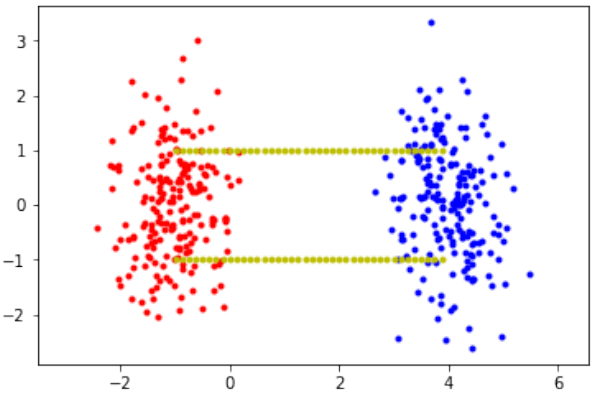


Figura 3.1: s_i del doble puente, en rojo la distribución source y en azul la objetivo (target).

Primero se resolvió el problema de transporte óptimo desde la distribución roja hacia la azul, los puntos rojos serán x_i y los azules los y_i. Inicialmente, utilizaremos sólo "feature functions" lineales y cuadráticas. Podemos obtener diversas soluciones según cual solución inicial o factible demos al algoritmo como se observa en la siguiente figura.

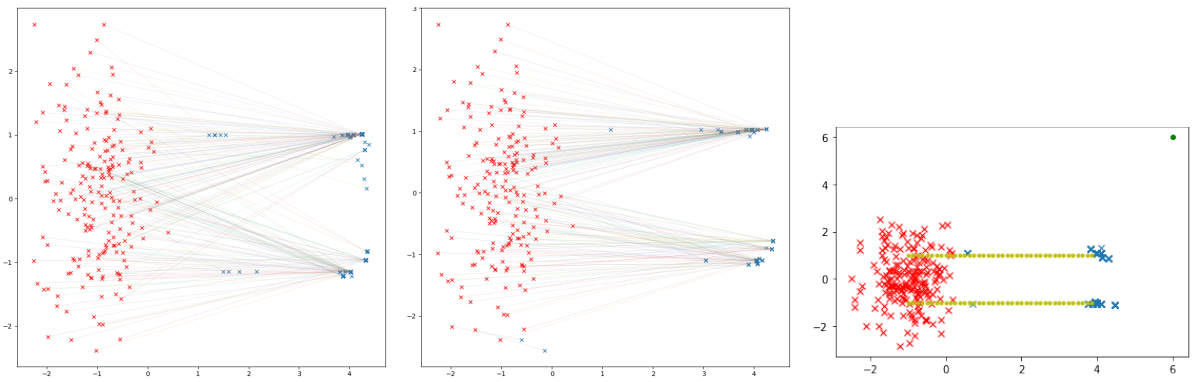


Figura 3.2: En azul el resultado del transporte óptimo con distancia de Fermat con $\alpha = 2$ tras 500 iteraciones. Cada línea une a un x_i -puntos rojos- con su correspondiente $\tilde{x_i}$ -cruces azules-. Izquierda: con solución inicial y_i. Centro: con solución inicial x_i. Derecha: con una solución inicial que son todos puntos iguales (todos en (6,6)) que se refleja en verde.

Podemos ver en la izquierda de la Figura 3.2 cierto problema: hay puntos de abajo a la izquierda que se asignan a los de arriba a la derecha, o sea que están pasando puntos de abajo por el puente de arriba, que no es lo esperado. Esto es porque la asignación de puntos iniciales fue totalmente aleatoria y la configuración obtenida se trata de un mínimo local. Si bien cambiar asignaciones puede mejorar la función, nuestro gradiente no permite que el método vea que hay del otro lado de la colina local. Si bien podrían aumentarse la cantidad de vecinos que tiene en consideración el gradiente para poder ver del otro lado de la colina, esto a su vez quitaría la localidad del gradiente haciendo que pequeños pasos no hagan descender la función.

Podemos ver en la segunda imagen que el óptimo sí parece tener buenas asignaciones; esto es porque al proponer una solución inicial igual a la distribución inicial se logra que los puntos "viajen" hacia el objetivo. Sin embargo, las tres imágenes no recuperan para nada una distribución normal ya que las restricciones impuestas no son suficientes para hacerlo: solo se busca tener la misma media y varianza y concentrar los puntos en la llegada del puente permite lograrlo. Para buscar mejorar esto se podrían agregar más "feature functions" que recuperen mejor la forma de la distribución objetivo. Una buena manera de hacer esto es utilizar funciones que estimen la densidad de dicha distribución. Para eso, se seleccionaron centros y se pusieron en cada uno de ellos núcleos ("kernels") centrados allí; utilizando también como solución inicial los x_i se obtuvo la solución de la Figura 3.3, donde se recupera satisfactoriamente la distribución objetivo.

Los kernels utilizados fueron de la forma

$$F_{z_0}(x) = \frac{1}{h^d} K\left(\frac{x - z_0}{h}\right) = K_h(x - z_0). \tag{3.4}$$

En el caso de los kernels gaussianos, K es la función de densidad de una normal con media 0 y matriz de covarianza $\mathbb{I}$. Esto implica que K_h es la función de densidad de una normal con media 0 y matriz de covarianza $h\mathbb{I}$, es decir:

$$K_h(x) = \frac{1}{\left(\sqrt{2\pi}h\right)^d} e^{\frac{-||x||^2}{2h^2}}, \tag{3.5}$$

donde d es la dimensión (de x) y h es un parámetro a elegir, comúnmente llamado ventana ("bandwidth") ya que justamente determina el radio alrededor del z_0 al cual damos más

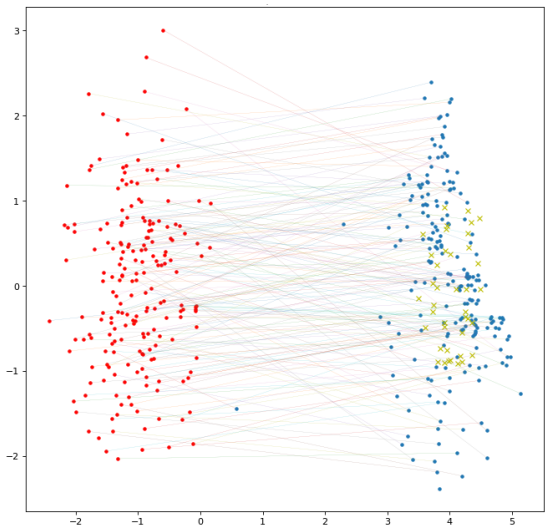


Figura 3.3: En azul el resultado del transporte óptimo con distancia de Fermat tras 500 iteraciones en el doble puente, con solución inicial x_i. Las cruces amarillas representan los centros de los kernels y los puntos rojos la distribución source.

importancia. A menor h, la función se concentra cerca de z_0 y a mayor h pesan más las colas. Cabe destacar que se trata de una función de decaimiento exponencial, si bien aumentar el parámetro h nos permite dar más importancia a zonas lejanas la función en sí considerará entornos locales.

Arrollado de dulce de leche

Ahora consideramos puntos en una superficie que se asemeja a un arrollado o pionono de dulce de leche (también conocida en la literatura como "Swiss Roll") , como se muestra en la siguiente figura.

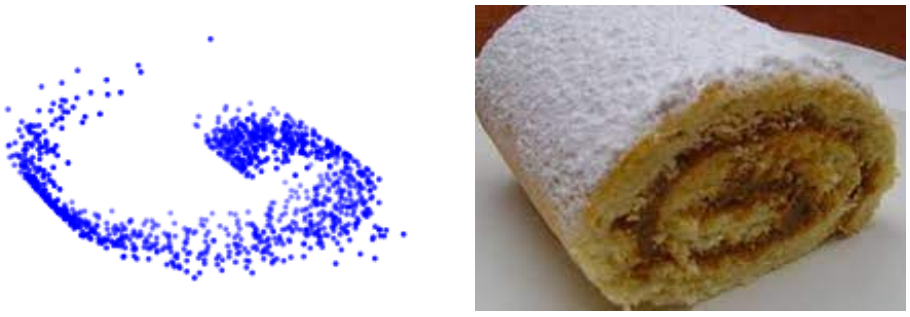


Figura 3.4: Izquierda: Puntos s_i que definen nuestra superficie. Los puntos se obtienen de sortear normales bivariadas y luego aplicarles la función $g(x, y) = (x\cos(ax), y, x\sin(ax))$, con $a = 6$ en este caso. Derecha: un arrollado de dulce de leche.

Si aplicamos el mismo algoritmo de minimización con restricciones de primer y segundo momento obtenemos resultados como se ve en la siguiente figura.

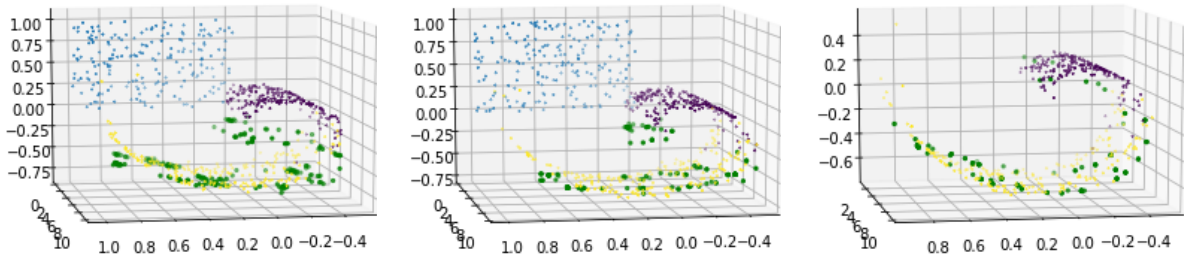


Figura 3.5: Soluciones de transporte óptimo desde la distribución violeta (source) hacia la amarilla (target) utilizando la distancia de Fermat. La superficie S es la determinada por la unión de los puntos de ambas distribuciones. Izquierda: en azul la solución inicial provista y en verde la dada por el transporte de Fermat tras 1000 iteraciones, con $\alpha = 2$. Centro: ídem con $\alpha = 3$, los puntos quedan más cerca de la superficie. Derecha: con dato inicial y_i (o sea como la distribución amarilla) tras 5000 iteraciones.

Podemos ver en la Figura 3.5 que se obtiene una solución satisfactoria, que mejora a partir de una solución inicial muy mala. Al comparar la imagen de la izquierda con la del centro, vemos claramente en la solución que aumentar el α hace que el costo le de más importancia a que los puntos estén cerca de la superficie. Notemos que las soluciones sólo cumplen la restricción de primer y segundo momento, si se deseara obtener una distribución que se parezca aún más a la objetivo habría que incluir más restricciones, por ejemplo kernels. La imagen de la derecha ilustra que al empezar con una solución inicial buena (de hecho es la distribución objetivo) y buscar minimizar la función de costo, los puntos tienden a concentrarse en puntos de la superficie.

Notemos que esto puede ser una buena forma de obtener una nueva muestra de una distribución (la objetivo). Si quisiéramos obtener más datos con dicha distribución podríamos tomar datos de alguna otra distribución (que será la source) y transportarlos a la objetivo. Esto suena más esperanzador aún si la distribución source está soportada en la misma superficie que la objetivo, lo que puede darse si tomamos distribuciones (o datos) que presenten rasgos en común o que provengan de un mismo ámbito. Podríamos no sólo obtener nuevos datos sino también saber como transportar los datos source, que podrían generarse sintéticamente, en los datos fuente, que podrían ser datos reales.

Vemos en la Figura 3.6 lo que Fermat aporta, donde resolvemos el mismo problema pero

en lugar del costo de Fermat ponemos el costo euclídeo. Utilizar costo de Fermat codifica información de la superficie en la misma función de costo, y no solo en las feature functions.

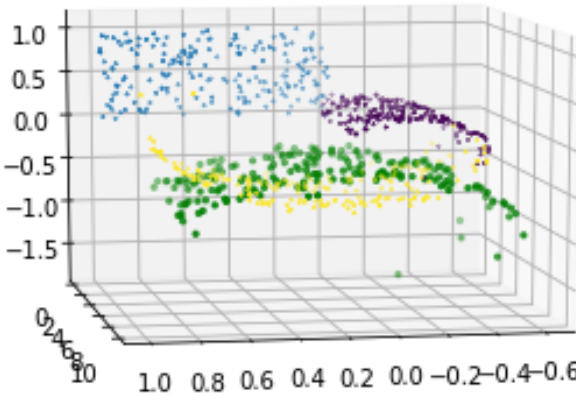


Figura 3.6: En azul la solución inicial provista y en verde la dada por transporte óptimo euclídeo con restricciones de primer y segundo momento. El solver terminaba satisfactoriamente en 74 iteraciones pues la función de costo es convexa y simple. Se obtenía la misma solución con distintos datos iniciales.

3.3. Optimización Combinatoria

En la sección anterior hemos resuelto el problema de transporte óptimo optimizando sobre las variables $T(x_i)$, imponiendo que estos puntos tengan la misma distribución que los y_i. Otra forma de abordar esto podría ser encontrar qué asignación de x_i a $y_{\sigma(i)}$ minimiza el costo. Notemos que esta sólo será útil cuando la cantidad de muestras sea la misma en ambos casos y además no nos permitirá obtener nuevas muestras de la distribución objetivo. Con este enfoque, podemos pensar al problema de transporte óptimo como uno de optimización combinatoria si ambas distribuciones tienen la misma cantidad de muestras. Básicamente estamos resolviendo el problema (1.6), que se trata de un problema bastante estudiado. Uno podría resolver esto para cualquier costo utilizando algún algoritmo como simplex por ejemplo. La matriz de costo en nuestro caso será la que contiene los pares de distancias de Fermat entre puntos de source y target. Para resolver esto se utilizó la función ot.emd de la librería POT (Python Optimal Transport).

Doble puente invertido

Inicialmente veamos un caso similar al doble puente, donde la distribución fuente y la objetivo están compuestas por dos normales, pero con distintas conexiones como muestra la siguiente figura.

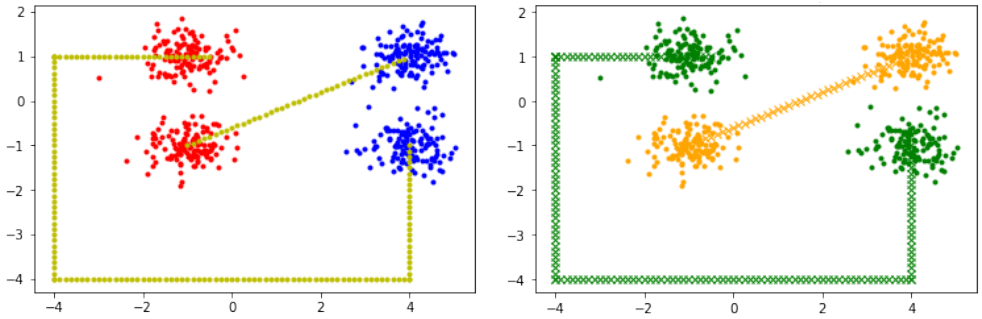


Figura 3.7: Izquierda: Puntos s_i de la superficie S (rojos, azules y amarillos), en rojo el source y en azul el target. Derecha: en igual color los puntos que fueron asignados unos con otros al resolver el problema de transporte óptimo con el costo de Fermat con $\alpha = 3$, en cruces los puntos s_i que no son ni de source ni target.

Usar aquí el costo de Fermat da una asignación bastante distinta a la del costo euclídeo, que sería similar a una traslación. Vemos efectivamente que en la Figura 3.7 los puntos de arriba a la izquierda se asignan con los de abajo a la derecha y los de abajo a la derecha con los de arriba a la izquierda ya que "viajan" por los puentes amarillos.

Podemos utilizar este enfoque para resolver también el problema del baricentro. Ahora precisamos brindar mas información porque el baricentro será definido en un nuevo lugar. Uno puede pensar que una distribución se puede representar computacionalmente como un conjunto de muestras, todas con igual importancia, es decir a las cuales le asignamos igual peso como hemos hecho hasta ahora. Sin embargo, uno también podría elegir un conjunto de puntos fijo que actúe como soporte de la distribución, que podría ser por ejemplo una grilla del espacio, y asignarle pesos a cada uno de esos puntos, como si se definiera una distribución discreta; los pesos dicen que constante acompaña a cada delta de Dirac. Si tomamos este enfoque, la distribución source y target serán un conjunto de pesos y el baricentro que estamos buscando también lo será; estas tres distribuciones estarán definidas en un mismo soporte. En nuestro contexto podemos realizar el análisis en la superficie S, es decir, hallar

el baricentro en S.

Parábola

En este caso, usando la función ot.bregman.barycenter de la librería POT (que resuelve una regularización del problema del baricentro -no utiliza exclusivamente optimización combinatoria-), se obtienen resultados para el baricentro de dos normales en una superficie como una parábola, como se ve en la Figura 3.8.

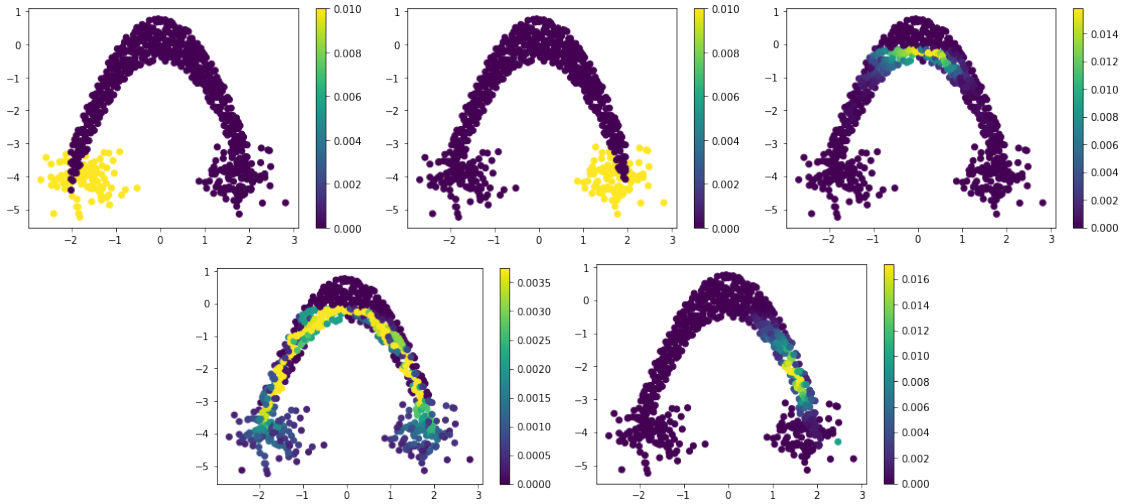


Figura 3.8: En todos los gráficos los puntos son el soporte que coincide con la superficie S, representada por los s_i; la escala de colores representa los pesos que se le asignan a la delta de Dirac en cada punto. (i) Una distribución a la que tomamos baricentro (ii) La otra distribución (iii) Baricentro con el costo de Fermat con pesos $w_i = 1/2$ y $\alpha = 2$ (iv) Ídem pero con el costo de Fermat sin elevar al cuadrado (v) Baricentro con pesos 0,8 y 0,2 con el costo de Fermat y $\alpha = 2$.

Vemos en la imagen (iii) de la Figura 3.8 que la solución del problema del baricentro se encuentra sobre la geodésica, ya que allí se minimiza el costo y que se encuentra en un lugar intermedio de la geodésica que une los centros de las dos nubes de puntos. En dicha figura vemos también la importancia de tomar las distancias al cuadrado en el baricentro: si no hacemos esto, como se trata de una distancia, es indistinto dónde poner el baricentro siempre que caiga en la geodésica (revisar Comentario 1.3.7). Esto es lo que sucede en (iv), donde obtenemos un baricentro no deseado, aunque nos muestra claramente la "geodésica dada por Fermat entre ambas distribuciones".

Por su parte, el enfoque de programación lineal puede ser muy ventajoso para obtener

transportes entre distribuciones que son sumas de deltas de Dirac, pero tiene sus limitaciones para el baricentro ya que precisa un soporte definido si uno no utiliza la distancia euclídea (problema bastante estudiado y con muy buenas soluciones, ver librería POT [22])). Si quisiéramos dar mayor libertad dando un soporte con muchos puntos, se dificultaría mucho el cálculo de las distancias de Fermat dos a dos (algoritmo $\mathcal{O}(n^3)$ -que puede aproximarse con menor complejidad-). Más aún, los enfoques de programación lineal son lentos para muchas aplicaciones prácticas ya que son $\mathcal{O}(n^3)$. Utilizar programación lineal requiere las distancias precalculadas, mientras que si hacemos una optimización no lineal podría pensarse alguna forma de ir aproximando la distancia de Fermat y calculando a la vez la solución que uno quiere, sea de transporte óptimo o baricentro.

Además, en la optimización no lineal no hay un soporte fijo y justamente se pueden obtener nuevos puntos fuera de la superficie y muestras. Sin embargo, el algoritmo propuesto no puede llevarse a resolver el problema del baricentro. Esto es porque si bien se trata de minimizar la suma de los costos de problemas de transporte óptimo, deberíamos calcular a cada paso el transporte óptimo desde el baricentro hacia cada distribución y, en base a eso, ir modificando el baricentro para mejorar la función objetivo. Esto sería muy lento ya que cada iteración consiste en resolver el problema de transporte óptimo antes descrito tantas veces como distribuciones haya, haciendo que el cómputo sea demasiado lento. La sección 3.4 busca mostrar como podemos resolver el problema del baricentro con un enfoque de optimización no lineal y basado en muestras. Antes de eso veremos una aplicación de transporte óptimo con costo de Fermat, que brinda resultados muy distintos a los obtenidos por el costo euclídeo.

3.3.1. Color Transfer con Fermat

Color Transfer es una aplicación de transporte óptimo que ya hemos comentado (ver Figura 1.4). Utilizaremos imágenes que se componen de vectores en $\mathbb{R}^3$, de tamaño 178×283, es decir de 50374 pixels. Querremos transportar una imagen en otra, es decir 50374 puntos en otros; esto sería computacionalmente muy caro sin utilizar alguna regularización que facilite el problema y permita utilizar un algoritmo como el de Sinkhorn-Knopp [33] que es mucho más veloz que simplex. Para estudiar puramente el efecto del costo de transporte

resolveremos el problema sin ninguna regularización. Computaremos el transporte con una cantidad acotada de puntos, tomando 2000 de cada imagen elegidos aleatoriamente. Luego, para transportar otro punto x_0 utilizaremos "transport map-up sampling" (ver fórmula (4.1) de [30]). Si el vecino más cercano (en distancia euclídea) a x_0 de quien conocemos el transporte es x_1 e y_1 es el transportado de este x_1, esto consiste en proponer al transportado de x_0 como $y_0 = y_1 + (x_0 - x_1)$.

Figura 3.9: Izquierda y centro: Dos imágenes, transferiremos color de una a otra. Derecha: se grafican las coordenadas rojo y azul de los 2000 pixels elegidos aleatoriamente de cada imagen.

Transferiremos color de una imagen a la otra y viceversa, con las imágenes que se ven en la Figura 3.9. Aplicaremos a todos los pixels el transporte previamente aprendido con 2000 de cada imagen, utilizando ot.emd de la librería POT. En la siguiente figura vemos los resultados para la distancia euclídea y la de Fermat con $\alpha = 2$ como costo.

Figura 3.10: Color Transfer en ambos sentidos de las imágenes de la Figura 3.9. Arriba: costo euclídeo al cuadrado. Abajo: costo de Fermat al cuadrado con $\alpha = 2$.

Vemos en la Figura 3.10 que los colores se transfieren de forma opuesta, lo que se puede entender al observar la imagen de la derecha de la Figura 3.9. Si allí consideramos la distancia euclídea, tendríamos un transporte similar a una traslación, que asigna a los valores con alto rojo y bajo azul (abajo a la derecha) valores con alto azul y rojo (arriba a la derecha). En cambio, si consideramos la distancia de Fermat, tendríamos un transporte que asigne a los valores con alto rojo y bajo azul (abajo a la derecha) los de bajo azul y rojo (abajo a la izquierda).

3.4. Minimax: Baricentro "Sample Based"

Para aplicar el algoritmo propuesto en la Sección 3.2 y calcular el baricentro μ de $\mu_1, \ldots, \mu_k$ con pesos w_i tendríamos que resolver muchos problemas de transporte óptimo a cada paso: los que van desde la distribución baricentro μ (que estamos optimizando) hacia cada una de las μ_i. En el caso euclídeo es claro que actualizar un punto de μ como el promedio pesado de los puntos a los que fue transportado hacia cada una de las μ_i es una buena elección, sin embargo al utilizar el costo de Fermat u otro costo no hay una forma tan clara de actualizar ese baricentro y ver cómo irlo mejorando. Podríamos pensar quizás en sumar (pesando con w_i) los gradientes de los costos de cada transporte desde μ hacia μ_i, sin embargo esto es costoso e ineficiente ya que requiere resolver k problemas de transporte óptimo a cada paso. Más aún, habría que definir "feature functions" de antemano para cada transporte. Además, sumar pesadamente gradientes que están contenidos en la superficie puede dar un resultado que se salga de esta lo que no es para nada deseable porque estamos buscando un baricentro soportado allí.

Más allá del costo que usemos, hay también una desventaja que presenta utilizar "feature functions": la distribución se parecerá solamente en esas funciones que le imponemos. Para evitar esto podemos pensar en un enfoque minimax, en el cual queremos minimizar el costo entre todos los posibles transportes a la vez que queremos expresar como una maximización el cumplir la restricción de que las distribuciones "se parezcan" (sea el transportado de source con target en transporte óptimo o los transportados de todas las distribuciones μ_i entre sí en baricentro). Esto se puede formular como una maximización si a la función objetivo le sumamos un término no negativo multiplicado por λ que valga 0 cuando se cumple

la restricción. Si la restricción no se cumple, podemos tomar λ tan grande como querramos para maximizar. Tenemos una minimización y maximización a la vez, cosa que puede pensarse como un juego de dos rivales: los transportes jugarán a minimizar el lagrangiano mientras que las "feature functions" jugarán a maximizarlo. Para formalizar esto, generalicemos el problema del baricentro, donde en vez de tener finitas $\mu_1(x), \ldots, \mu_k(x)$ pasamos al caso continuo donde tenemos una conjunta $\rho(x, z)$. Ahora tomamos baricentro sobre las $\rho(\cdot|z)$. Esto funciona no sólo en el caso de que z sea categórico (que es el baricentro entre finitas distribuciones) sino también en el caso que sea continuo. El transporte T dependerá ahora también de z. Luego, el problema es:

$$\min_T \int\int c(x, T(x, z))\rho(x, z)dxdz, \quad \text{s.a.} \quad T_\#\rho(\cdot|z) = \mu \quad \forall z. \tag{3.6}$$

Notemos que aquí ponemos como restricción que, independientemente del z, todas las distribuciones se transporten a la misma distribución μ, que será el baricentro. Como hemos dicho, vamos incluir esta última restricción en el funcional a minimizar, siguiendo la propuesta de [23] que es aplicable para cualquier costo en general. Para ello, denotemos $y = T(x, z)$ a los puntos transportados que representarán la distribución baricentro; podemos pensar que la distribución representada es la empírica, es decir, la suma de δ_{y_i} con pesos uniformes (como en (1.4) con $b_j = 1/n$). Notemos ahora que $Y = T(X, Z)$ debería ser independiente de Z pensándolos como variables aleatorias, ya que todos los transportados de las distribuciones van a parar al mismo baricentro μ independientemente de qué z provengan. Si son independientes sabemos que $T_\#\rho(y, z) = \nu(y)\eta(z)$, de donde

$$\int\int F(T(x, z), z)\rho(x, z)dxdz = \int\int F(y, z)T_\#\rho(y, z)dydz = \int\int F(y, z)\nu(y)\eta(z)dydz. \tag{3.7}$$

Luego, si pedimos que para todo y suceda que $\mathbb{E}_z[F(y, \cdot)] = 0$ (donde $\mathbb{E}_z$ denota que tomamos esperanza sobre la variable z), tendremos que (3.7) se anula, es decir que $0 =$
$= \int\int F(T(x, z), z)\rho(x, z)dxdz$. Tenemos entonces que (3.7) se anula si y solo si y es independiente de z que es justamente lo que queremos imponer al calcular el baricentro. Podemos, en vez de imponer que la $\mathbb{E}_z[F(y, \cdot)] = 0$, restar directamente la esperanza para garantizarnos esto. Así, el término que queremos agregar será $\int\int (F(y, z) - \mathbb{E}_z[F])\, T_\#\rho(y, z)dydz$,

dando lugar al siguiente problema de minimax (minimización y maximización):

$$\min_{T} \max_{F} \int\int c(x, T(x,z))\rho(x,z) + (F(y,z) - \mathbb{E}_z[F])\, T_{\#}\rho(y,z)dydz. \tag{3.8}$$

El funcional al que le aplicamos minimax en (3.8) es el lagrangiano L que lo podemos separar en dos partes: L_c que se refiere al costo de transporte y L_F que codifica si las distribuciones transportadas se parecen o no. Ahora, la función objetivo se puede llevar fácilmente al caso discreto: las integrales pasan a ser sumas y las esperanzas promedios. Otro aspecto importante es que aquí no aparece el baricentro μ en el problema a resolver. Es por esto que tomaremos la formulación de (3.8) en vez de la de (3.6). En un enfoque de minimax uno tiene además la ventaja de no fijarse en ciertas F determinadas, sino que la F se irá adaptando según T; podemos interpretarlo como un juego entre dos jugadores que compiten.

En [23] se propone evitar la maximización sobre todas las funciones, llevándola a una maximización en un parámetro. Esto se logra eligiendo una F en particular que codifique bien dicha maximización y que sea dependiente de T. Para ello dan dos elecciones de F como propuesta, veremos y utilizaremos la primera que es tomar $F(y,z) = T_{\#}\rho(y|z)$. Esta función es adecuada ya que logra que el término que agrega en el lagrangiano (el L_F) sea siempre mayor a 0 y se anule solo si las distribuciones transportadas se parecen. Esto es lo que nos dice la siguiente proposición (Proposition 1 de [23]).

Proposición 3.4.1. *Si $F(y,z) = T_{\#}\rho(y|z)$, $L_F = \int\int (F(y,z) - \mathbb{E}_z[F])\, T_{\#}\rho(y,z)dydz$ es estrictamente positivo salvo que $T_{\#}\rho(y|z)$ sea independiente de z.*

Demostración. Recordando que $T_{\#}\rho(y,z) = \nu(y)\eta(z)$, tenemos que

$$L_F = \int \left(\int F(y,z) T_{\#}\rho(y|z)\eta(z)dz - \int F(y,z)\mathbb{E}_z[F]\eta(z)dz \right) dy.$$

Si aquí reemplazamos $F(y,z)$ por $T_{\#}\rho(y|z)$ obtenemos

$$L_F = \int \left(\mathbb{E}_z[T_{\#}\rho^2(y|\cdot)] - \mathbb{E}_z[T_{\#}\rho(y|\cdot)]^2 \right) dy.$$

Basta aplicar la desigualdad de Jensen que nos da que la integral será estrictamente positiva salvo que (fijado el y) $T_{\#}\rho(y|\cdot)$ sea constante, es decir que $T_{\#}\rho(y|z)$ sea independiente de z. Podemos sino pensar que la varianza es siempre no negativa y sólo es 0 cuando se trata de una constante. □

Así, el término agregado será nulo cuando se cumple la restricción de (3.6) para alguna μ, que es justamente lo que queremos: que las distribuciones transportadas sean la misma independientemente del z. Esta F en particular penaliza la dependencia en z de $T_{\#}\rho(y|z)$. Luego, si tomamos la maximización en F de (3.8) solamente sobre las funciones de la forma $F(y,z) = \lambda T_{\#}\rho(y|z)$ obtenemos el siguiente problema a resolver:

$$\min_{T} \max_{\lambda} \int c(x, T(x,z))\rho(x,z)dxdz + \lambda \int \left(T_{\#}\rho(y|z) - \int T_{\#}\rho(y|w)\eta(w)dw \right) T_{\#}\rho(y,z)dydz. \tag{3.9}$$

Ahora sí estamos en condiciones de expresar este problema a partir de las muestras. No parametrizaremos los transportes T sino que optimizaremos directamente la variable $y = T(x,z)$. De esta manera, lo único que queda determinar es la versión discreta de $T_{\#}\rho(y|z)$ fijado el z. Esto se puede hacer mediante cualquier estimación de densidad condicional, en particular utilizaremos la de Nadaraya-Watson [25] que nos da

$$T_{\#}\rho(y|z_k) \approx \frac{\sum_i K_a(y, y_i)K_b(z_k, z_i)}{\sum_j K_b(z_k, z_j)}, \tag{3.10}$$

donde K_a y K_b son núcleos a elección que en nuestro caso fueron tomados como núcleos gaussianos (ver (3.5)). Notemos que esto se puede expresar como $\sum_i K_a(y, y_i)Z_{ik}$ donde $Z_{ik} = \dfrac{K_b(z_k, z_i)}{\sum_j K_b(z_k, z_j)}$. Esto es muy ventajoso ya que los valores z no cambiarán, por lo que es una matriz precomputable. Luego, tenemos que

$$T_{\#}\rho(y_l|z) - \int T_{\#}\rho(y_l|w)\eta(w)dw = T_{\#}\rho(y_l|z) - \mathbb{E}_z[T_{\#}\rho] \approx \sum_i K_a(y_l, y_i)\left(Z_{il} - \frac{\sum_k Z_{ik}}{N} \right), \tag{3.11}$$

con N el tamaño de la muestra. De aquí podemos definir $C_{il} = Z_{il} - \dfrac{1}{N}\sum_k Z_{ik}$ y aproximar (3.11) como $\sum_i K_a(y_l, y_i)C_{il}$ con C una matriz precomputable.

En el caso en que las variables de z sean simplemente etiquetas (o sea que z sea una variable categórica, que nos separe en grupos, sin ninguna noción de distancia entre las z) tendremos que $Z_{ij} = \dfrac{1}{N_i}\mathbb{1}_{z_i = z_j}$ donde $N_i = \#\{z_k : z_k = z_i\}$. Podemos pensar esto como un caso límite en que tomamos la ventana b del núcleo muy pequeña. Una vez definida la matriz C ya no serán relevantes los kernels en z y solo trabajaremos con los K_a, que notaremos K_h también, es decir, llamaremos h en vez de a al "bandwitdth" de los kernels en y.

En [23], la Proposición 4 muestra con una simple cuenta que en el caso discreto L_F es siempre positivo, como comentamos en la siguiente proposición.

Proposición 3.4.2. *Si las matrices K_a y K_b (con entradas $K_a(y_k, y_i)$ $K_b(z_k, z_i)$) son semidefinidas positivas, la discretización de L_f es no negativa, es decir $\sum_{i,l} K_a(y_i, y_l)C_{il} \geq 0$*

Demostración. En la discretización de L_F estamos haciendo el producto punto a punto de la matriz C y la matriz K_a, es decir su producto interno en el espacio de matrices. Notemos que si ambas matrices son semidefinidas positivas, $C : K_a = \sum_{i,l} K_a(y_i, y_l)C_{il} \geq 0$. Para ver esto, se puede observar que $C : K_a = \mathrm{tr}(CK_a^T)$ y usar que K_a es simétrica para ver que equivale a $\mathrm{tr}(CK_a)$. Luego, al descomponer K_a en una base de autovectores $v_1, \ldots, v_n$ con respectivos autovalores $\lambda_1, \ldots, \lambda_n \geq 0$, obtenemos que $C : K_a = \sum_i \lambda_i \langle Cx_i, x_i\rangle$ será no negativo si C es semidefinida positiva.

Veamos que C es semidefinida positiva para completar la demostración.

$$x^T Cx = \sum_{i,j} x_i C_{i,j} x_j = \sum_{i,j} x_i Z_{i,j} x_j - \frac{1}{N} \sum_{i,j,k} x_i Z_{i,k} x_j =$$

$$= \sum_{i,j} x_i Z_{i,j} x_j - \frac{1}{N} \sum_{i,j,k} x_i Z_{i,k}(x_j - x_k + x_k) = -\frac{1}{N} \sum_{i,j,k} x_i Z_{i,k}(x_j - x_k) =$$

$$= \sum_{i,k} x_i Z_{i,k}(x_k - \overline{x}) = \sum_{i,k} (x_i - \overline{x}) Z_{i,k}(x_k - \overline{x})$$

Como Z es semidefinida positiva por construcción, pues K_b lo es, se sigue que la última sumatoria es positiva y por ende C es semidefinida positiva. □

Comentario 3.4.3. Los núcleos gaussianos son definidos positivos y nos aseguran que se cumpla la hipótesis de la Proposición 3.4.2

Estamos ahora en condiciones de proponer un algoritmo que resuelva el problema discreto, que es

$$\min_y \max_\lambda \sum_i c(x_i, y_i) + \lambda \sum_{i,l} K_a(y_i, y_l)C_{il}. \tag{3.12}$$

Para resolver esto uno podría optar por un algoritmo de minimax, pero siendo que la maximización se da solo sobre un escalar es una posibilidad utilizar un método de penalización

donde vamos minimizando en y mientras que vamos aumentando λ progresivamente, como se propone en [23]. En esencia, la forma de aumentar λ es tal que el gradiente de L tenga componente positiva en L_F, es decir que siempre mejoremos, aunque sea un poco, en la condición de que las distribuciones se parezcan. Cuánto exigimos mejorar en esa dirección se verá dado por un parámetro β (llamado α en [23]), de forma que:

$$\langle \nabla L, \nabla L_F \rangle \geq \beta \langle \nabla L_F, \nabla L_F \rangle. \tag{3.13}$$

En particular, se sugiere que se tome β de forma adaptativa ya que en caso de haber un óptimo local para un valor de λ, obtendremos que para actualizar se propone $\beta + \lambda$, donde suena razonable que β crezca como λ para que la actualización sea significativa. Es decir, si λ_n es el valor de λ en el paso n-ésimo, para obtener λ_{n+1} tomaremos $\beta = \omega\lambda_n$ con $\omega \in (0, 1)$.

Además de la elección de β que refleja la velocidad con la que crece λ hay otro parámetro muy sensible del algoritmo que es h, el "bandwidth" de los kernels. Notemos que un h muy grande no tendrá ninguna noción de localidad y por ende buscará hacer que las medias coincidan, mientras que un h muy chico verá mas detalle pero puede ser tan local que si las distribuciones se encuentran inicialmente lejos nunca se acerquen. Podemos ver en la siguiente imagen la importancia de h y la sensibilidad a este valor en un problema de baricentro con el costo euclídeo.

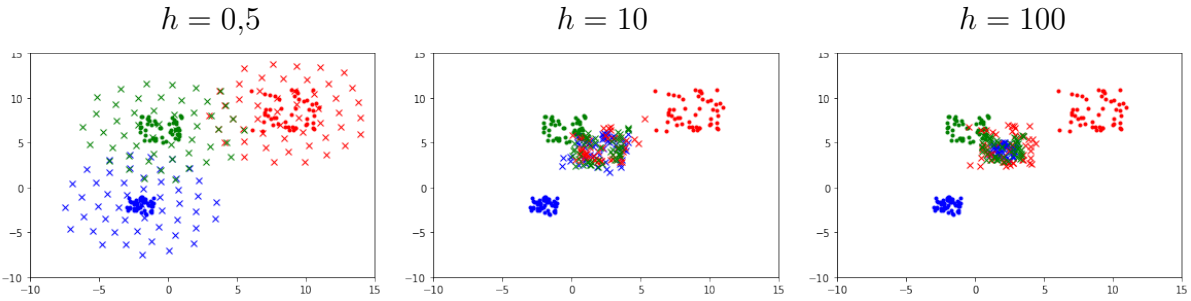


Figura 3.11: Primeras 100 iteraciones del algoritmo propuesto para el problema de baricentro con costo euclídeo entre tres normales con distintos valores de h fijos. Las muestras de las distribuciones x_i se ven representadas con puntos de distintos colores (según su valor z_i). El baricentro $y_i = T(x_i, z_i)$ se ve con cruces que tienen el mismo color que su x_i respectivo. En todos los casos se ejecuto el algoritmo con $\omega = 0{,}6$ y $\lambda_0 = 0{,}1$.

La Figura 3.11 muestra claramente la desventaja de tomar tanto un h muy grande como uno muy chico. Notemos que la matriz C tiene entradas positivas para los y_i que tienen un

mismo z_i y entrada negativa para los que tienen distinto z_i; es decir que al minimizar lo que buscamos hacer es alejarnos de los puntos de igual z y acercarnos a los de distinto valor de z. Al tomar h pequeño, los kernels no llegan a tener en cuenta puntos lejanos y en el caso de la Figura 3.11 solo tienen en cuenta a puntos de su misma distribución, por lo que la única forma de minimizar el lagrangiano es alejándose entre sí. Es por esto que con $h = 0{,}5$ se observa que las distribuciones agrandan su varianza. Esto se puede ver teóricamente si tomáramos un costo $c \equiv 0$ y buscásemos maximizar el término L_F (ver (3.8)), cosa que se logra maximizando la varianza, como hemos visto en la Proposición 3.4.1. A su vez, en la imagen de la derecha vemos que tomar un h muy grande logra acercar las medias pero no ve más detalle ya que no distingue entre puntos cercanos, lo que da como resultado que las tres distribuciones se trasladen pero sus varianzas no se ajusten. Vemos en la imagen del centro que con un h intermedio esto se logra correctamente. Uno podría fijar el h, pero sería muy dependiente de cada caso encontrar un h que funcione correctamente y poco general. Una posible propuesta para solucionar esto es ir modificando el h paso a paso: inicialmente un h alto permitirá que las distribuciones se acerquen pareciéndose en media y luego un h pequeño permitirá mas detalle y que las distribuciones se parezcan en otras características. Dos opciones posibles con buenos resultados son tomar un h muy grande e ir achicándolo paso a paso, o tomar a cada paso un h que dependa del desvío estándar de los y actuales quizás con alguna constante que asegure que al principio las medias se acerquen, por lo que al estar muy separadas las distribuciones el h será grande y al acercarse será más pequeño.

Otro parámetro sensible es el "learning rate", es decir cuán grande es el paso en la dirección del gradiente. Típicamente en técnicas de descenso por el gradiente se sigue alguna regla (Armijo, Wolfe, etc.) que asegure descenso con un paso lo mas grande posible, que es lo que haremos en nuestro caso como se sugiere en [23] (cambiando quizás la constante 2,01 por la que se multiplica el learning rate tras cada iteración). Seguiremos lo allí propuesto pero con un pequeño cambio para no quedarnos estancados en algún λ: cuando el gradiente se anule (i.e. muy pequeño) resetearemos el learning rate actual al learning rate inicial (o lo aumentaremos considerablemente) y multiplicaremos λ por una constante γ (que tomaremos como 1,1). En caso de que $\lambda \geq \lambda_{max}$ devolveremos el valor actual ya que hemos llegado al valor máximo de λ y tenemos un gradiente nulo, es decir, un óptimo local.

Hasta aquí hemos revisado la propuesta de [23] para resolver algorítmicamente el problema baricentro de Wasserstein con un costo en general para el cual conozcamos su gradiente. Con todo lo necesario definido, ya podemos resolver el problema de baricentro con costo de Fermat. Resolviéndolo en una superficie donde ya hemos visto el resultado esperado (ver la Figura 3.8), obtenemos el baricentro para distintos valores de h fijos como se ve en la Figura 3.12.

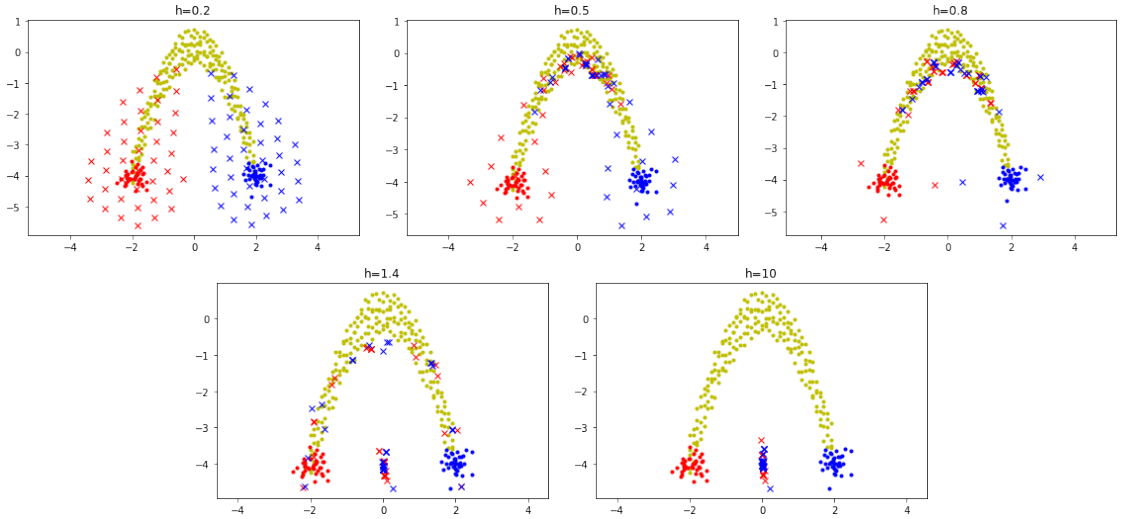


Figura 3.12: Primeras 300 iteraciones del problema de baricentro con costo de Fermat con $\alpha = 2$ en la superficie de puntos amarillos con distintos valores de h fijos. Las muestras de las distribuciones x_i se ven representadas con puntos de distintos colores (según su valor z_i), el baricentro $y_i = T(x_i, z_i)$ se ve con cruces que tienen el mismo color que su x_i respectivo. En todos los casos se ejecuto el algoritmo con $\omega = 0{,}05$ y $\lambda_0 = 1$.

En la Figura 3.12 vemos resultados no satisfactorios salvo quizás para $h = 0{,}8$. Nuevamente vemos el mismo efecto que tiene h sobre la forma en que inicialmente se acercan ambas distribuciones. El problema ahora es de una distinta naturaleza: los kernels están evaluados en distancia euclídea; por lo que al tomar un h grande las distribuciones tienden a igualar sus medias pero lo hacen sin tener en cuenta la superficie, porque la distancia euclídea no lo hace. Esto se ve claramente en $h = 1{,}4, 10$ de la Figura 3.12. Allí también se ve que en caso de tomar h pequeño la varianza se agranda llevando nuevamente a varios puntos lejos de la superficie. Es por esto que la estimación de densidad que estamos utilizando es incorrecta en este contexto: si bien el costo de Fermat lleva a que los puntos caigan sobre la superficie, no es suficiente ya que los kernels euclídeos no lo hacen.

3.4.1. Kernels de Fermat

Para resolver transporte óptimo en la Sección 3.2, utilizábamos funciones test al comparar que coincidan las $\mathbb{E}[f(\cdot)]$ de los transportados y la distribución objetivo; si esto lo hiciéramos sobre todas las f obtendríamos que la distribución es la misma ya que si la distribución coincide en el espacio ambiente, lo hará también en la superficie. Sin embargo, por la contribución de [23] de proponer una f particular que este atada al transporte T que simplifica notablemente la resolución computacional del problema, nos estamos limitando a una única f. Si mantenemos los kernels euclídeos esta única f no tendría en cuenta a la superficie y trataría de acercar las distribuciones logrando que estas se parezcan en el espacio euclídeo y no en el de la variedad que es lo que buscamos. La ventaja de los kernels gaussianos (ver (3.5)) es que están definidos a partir de una distancia, por lo que podemos cambiar la distancia euclídea $||\cdot||$ por la distancia de Fermat. Podemos también conocer el gradiente de estos kernels haciendo regla de la cadena, utilizando la ya dada definición del gradiente de Fermat. Tenemos todos los ingredientes necesarios para correr el algoritmo en este contexto. Ahora al tener un h grande, esperamos que las medias se acerquen pero que lo hagan moviéndose por la superficie. Los puntos se moverán por la geodésica pesada de Fermat para unirse, esto es exactamente lo que queremos y que recrea lo que sucede en el espacio euclídeo, donde los puntos se acercan en la linea recta que los une. Si bien el núcleo da más peso a un entorno local, con una ventana grande se tendrá en cuenta a puntos lejanos y la gran diferencia se va a dar en que los gradientes que apuntan hacia esos puntos van a ser geodésicos y por ende va a ser una buena decisión seguir la dirección que estos proponen.

Ahora entonces los núcleos K_a utilizados en el espacio de las y han cambiado, siendo kernels gaussianos con la distancia de Fermat en lugar de la euclídea. Para poder realizar el mismo algoritmo debemos ver que la estimación hecha en (3.10) siga siendo buena, o corregirla para que lo sea. Los núcleos en z no cobran relevancia ya que simplemente son la parte condicional y como vimos puede considerarse que estimamos con $\sum_i K_a(y, y_i)Z_{ik}$ donde Z es una matriz de coeficientes que no cambia, que sin perdida de generalidad podemos obviar para estudiar la estimación de densidad en y. Abordaremos entonces el problema de estimar una densidad en una variedad con kernels de Fermat. La estimación

por núcleos en variedades riemannianas fue estudiada por Pelletier en [28]. Allí se define en una variedad $\mathcal{M}$ de dimensión d (con su métrica g y distancia d_g) un estimador de una densidad f a partir de $X_i \sim f$ i.i.d. como

$$f_n(p) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\theta_{X_i}(p) r^d} K\left(\frac{d_g(p, X_i)}{r}\right), \tag{3.14}$$

donde θ_q es la función que refleja la densidad de volumen en puntos cercanos a q, es decir que compara áreas entre el plano tangente y la superficie. No ahondaremos en θ y lo supondremos constante (cosa que sucede en superficies que tienen curvatura constante, o si su curvatura no varía mucho θ tampoco lo hará). El núcleo K utilizado debe ser un kernel isotrópico, como el gaussiano que utilizamos. El siguiente teorema muestra la consistencia de dicho estimador (ver más detalle sobre condiciones en el Teorema 3.1 de [28], como también la demostración del mismo) .

Teorema 3.4.4. *Bajo condiciones de regularidad de f y una ventana r lo suficientemente pequeña, se tiene que*

$$\mathbb{E}_f ||f_n - f||_{L^2(\mathcal{M})} = \mathcal{O}\left(n^{-4/(d+4)}\right),$$

es decir, que el estimador es consistente con convergencia en sentido L^2.

Para obtener la consistencia de los kernels de Fermat nos apoyaremos fuertemente en este resultado, debiendo modificar únicamente la distancia geodésica d_g utilizada en la definición del estimador de (3.14) por la distancia de Fermat $d_{\tilde{f}}$, donde $\tilde{f}$ es la densidad con la que se define la distancia de Fermat (ver (2.2)). Notemos que si $\tilde{f}$ es constante $d_g(p, \cdot) = d_{\tilde{f}}(p, \cdot)\tilde{f}^{\beta}(p)$. Bajo hipótesis de suavidad de f, podemos suponer que esto valdrá localmente, que es el contexto que miran los núcleos con r pequeño. Luego, podemos considerar un nuevo estimador f_n^*

$$f_n^*(p) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\theta_{X_i}(p) r^d} K\left(\frac{d_{\tilde{f}}(p, X_i)}{r}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\theta_{X_i}(p) r^d} K\left(\frac{d_g(p, X_i)}{r\tilde{f}^{\beta}(p)}\right) \tag{3.15}$$

donde si llamamos $\tilde{r} = r\tilde{f}^{\beta}(p)$, obtenemos que

$$f_n^*(p) = \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{f}^{\beta d}(p)}{\theta_{X_i}(p)\tilde{r}^d} K\left(\frac{d_g(p, X_i)}{\tilde{r}}\right) = \tilde{f}^{\alpha-1}(p)\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\theta_{X_i}(p)\tilde{r}^d} K\left(\frac{d_g(p, X_i)}{\tilde{r}}\right)\right), \tag{3.16}$$

donde en la última igualdad simplemente usamos la relación $\beta = \dfrac{\alpha - 1}{d}$. Luego, utilizando el Teorema 3.4.4, podemos ver que $f_n^*(p) \longrightarrow f(p)\tilde{f}^{\alpha-1}$. Esto nos da que simplemente reemplazar la distancia euclídea por la de Fermat no es un estimador correcto sino que debemos dividir por $\tilde{f}^{\alpha-1}$. A su vez esta $\tilde{f}$ no es conocida, por lo que debemos estimarla. Podemos hacerlo nuevamente vía estimación de densidad por núcleos con centros en s_j que son los puntos que definen la superficie S de Fermat. Aquí se podría pensar naturalmente en dos caminos: usar kernels euclídeos o de Fermat. Utilizar kernels euclídeos sería arruinar toda la estimación basada en la variedad $\mathcal{M}$ por lo que tomaremos el segundo camino. Al estimar $\tilde{f}$ con kernels de Fermat, obtendremos que la convergencia es a $\tilde{f}\tilde{f}^{\alpha-1} = \tilde{f}^{\alpha}$, por lo que si queremos estimar $\tilde{f}^{\alpha-1}$ debemos corregir elevando a la $\dfrac{\alpha - 1}{\alpha}$.

Así, en vez de estimar con (3.11), la estimación que realizamos será

$$\frac{\sum_i K_a(y_l, y_i) C_{il}}{\dfrac{1}{m}\left(\sum_j K_a(y_l, s_j)\right)^{(\alpha-1)/\alpha}}, \tag{3.17}$$

donde m es la cantidad de puntos s_j en la superficie de Fermat. Para mayor eficiencia computacional podríamos considerar solo alguna cantidad de vecinos s_j a la hora de estimar f en el denominador.

Ahora sí, podemos utilizar el algoritmo con kernels de Fermat ya que esta estimación es correcta. Dividir por la densidad (elevada a la $\alpha - 1$) aporta a la hora de calcular el gradiente: donde haya baja densidad de Fermat, el gradiente será muy grande y donde haya alta densidad de Fermat será chico. Esto ayuda a que los puntos se mantengan cerca de la superficie. Hay también que destacar que si θ_{X_i} fuese una constante distinta de 1 no es un problema ya que esto sería absorbido por el parámetro λ que vamos incrementando (ver (3.12)). También cabe destacar que no conocemos la dimensión d de la variedad, por lo que es imposible normalizar correctamente con $1/h^d$, pero nuevamente es algo que λ absorbe. En caso de conocer la dimensión de la variedad en la que viven los datos normalizaremos correctamente, si no simplemente no lo pondremos. No es una buena idea normalizar con D la dimensión del espacio ambiente ya que de ser muy grande los núcleos serían muy pequeños, trayendo problemas numéricos.

En el algoritmo, hemos llamado h a la ventana a de los kernels. La actualización de h ahora debe hacerse con cuidado: utilizar el desvío estándar no es una buena idea ya que

nuevamente esto subyace en la distancia euclídea, una versión simple de modificar esto es considerar un δ-cuantil de las distancias de Fermat dos a dos de los transportados y_i (que aprovechamos ya hemos calculado, pues la necesitamos al calcular la matriz de los kernels dos a dos). En el caso z categórico, esperamos que en general haya que tomar un δ lo suficientemente grande para tener en cuenta a los puntos de las otras distribuciones, con el recaudo de que un δ demasiado grande puede dar poca precisión en que las distribuciones se parezcan, como hemos visto con h grande. Veamos entonces el resultado que se obtiene en la misma superficie al poner los kernels de Fermat.

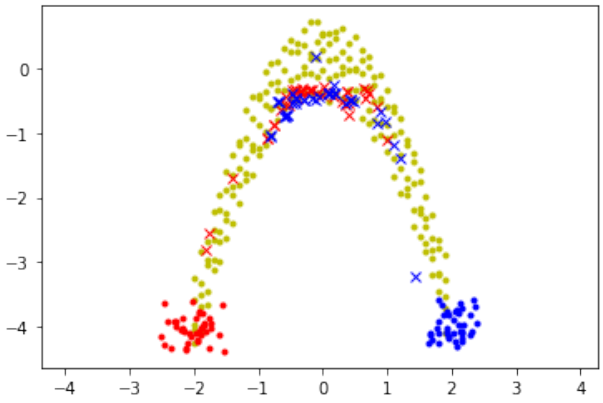


Figura 3.13: Solución del algoritmo al problema de baricentro con costo de Fermat con $\alpha = 2$ en la superficie de puntos amarillos. Las muestras de las distribuciones x_i se ven representadas con puntos de colores azul y rojo (según su valor z_i), el baricentro $y_i = T(x_i, z_i)$ se ve con cruces que tienen el mismo color que su x_i respectivo. Vemos el resultado tras 150 iteraciones. Se ejecutó el algoritmo con $\omega = 0{,}01$, $\lambda_0 = 10$ y $\delta = 0{,}65$

Herradura con ruido

Consideramos ahora una superficie en que la geodésica que une las dos distribuciones apunte en un sentido opuesto a la línea recta que las une. Además hemos agregado ruido a la superficie, cosa que se parece más a situaciones reales. Si utilizáramos kernels euclídeos, el gradiente de L_F llevaría a ambas distribuciones a juntarse en el centro, lo que es inadecuado. La siguiente figura muestra que los kernels de Fermat son los que hacen que los puntos viajen por la herradura abriéndose inicialmente para luego juntarse.

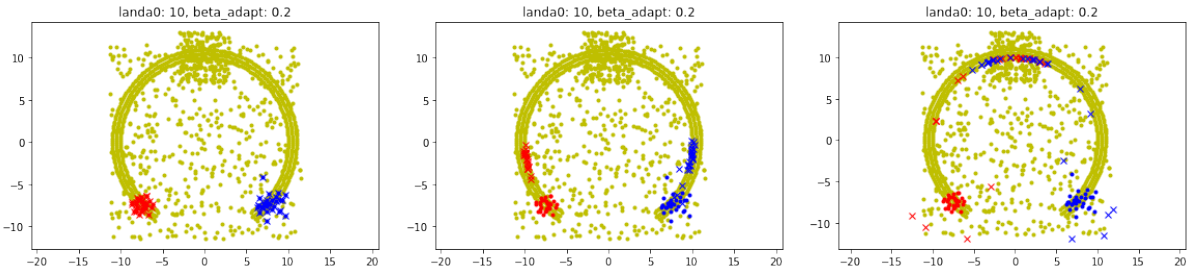


Figura 3.14: Solución del algoritmo al problema del baricentro con costo de Fermat con $\alpha = 2$ en la superficie de puntos amarillos. Vemos al comenzar el algoritmo (iteración 0) a la izquierda, iteración 100 en el medio e iteración 201 a la derecha (cuando se alcanza el $\lambda_{max} = 10000$ con gradiente de norma $< 10^{-7}$). Se ejecutó el algoritmo con $\omega = 0{,}2$ y $\lambda_0 = 10$ y $\delta = 0{,}65$.

Vemos en la Figura 3.14 que se alcanza un baricentro con éxito. En la imagen del centro los puntos viajan por la superficie en busca de mejorar el término L_F, es decir que las distribuciones se parezcan, cosa que no sucedía con kernels euclídeos. En la imagen de la derecha vemos la presencia de puntos "que se quedan en el camino', lo que se debe a que el h de los kernels se achicó y esos puntos quedaron fuera de la ventana de quienes están cerca del baricentro. Si bien esto no es para nada grave, se podría evitar quizás considerando un h muy grande cada determinadas iteraciones para no dejar puntos atrás.

En resumen, el algoritmo ahora tiene como parámetros a definir por el usuario:

- $\delta \in (0, 1)$, que indica el cuantil que tomamos, es decir cuan grande tomamos h.
- $\omega \in (0, 1)$, que nos indica cuan rápido aumentara λ.
- γ, que nos dice cuanto aumentar λ en caso de quedarnos trabados en un óptimo local para ese λ.
- λ_0 que es el λ inicial, λ_{max} que es el λ máximo que permitimos, el learning rate inicial y la constante por la que lo multiplicamos en cada paso. Estos son todos parámetros también libres en [23], donde los últimos dos se refieren más al descenso por el gradiente como técnica que a este algoritmo en sí.
- k, que es la cantidad de vecinos a considerar en el calculo de la distancia de Fermat y su gradiente para puntos que no son s_i, es decir que no están en la superficie.

De estos parámetros, δ es muy relevante porque, como hemos visto, el algoritmo es muy sensible al "bandwidth" de los kernels y δ controla eso. El resto de los parámetros no son tan importantes si se toman en escalas correctas, aunque sí tendrán sus consecuencias. Los parámetros ω ó γ muy grandes pueden dar poca importancia a que el costo se minimice, por crecer λ muy rápido, mientras que si se toma muy pequeño se harán muchas iteraciones. El parámetro λ_0 afectará exactamente de manera inversa: si es muy pequeño se necesitarán muchas iteraciones y si se toma muy grande puede que no se le de importancia al costo. El parámetro λ_{max} nos da una condición de corte si no limitamos la cantidad de iteraciones, pero en esencia esto es la versión computacional del infinito al que hacemos tender λ. Los parámetros del learning rate simplemente nos indican cuan conservadores o arriesgados trataremos de ser en los pasos que siguen al gradiente, mientras estén en el orden correcto esto no afectará demasiado más que eventualmente hacernos dar muchas iteraciones por ser muy pequeños. Por su parte, k cobra importancia como lo hacía al considerar Fermat en el grafo de k-vecinos más cercanos que muy similar a lo que estamos haciendo al calcular la distancia y el gradiente para puntos fuera de la superficie. Si pensamos en la convergencia del caso discreto al caso poblacional, necesitamos que k crezca hacia ∞ a medida que consideramos más puntos, por ejemplo de forma $k \propto log(n)$ (ver Comentario 2.0.7).

Otro algoritmo quizás más sencillo y que mostró mejores resultados fue el de hacer un método de barrera clásico. A diferencia del algoritmo propuesto en [23] que va aumentando λ a medida que minimiza en y, se podría resolver para un valor de λ el problema hasta conseguir un óptimo local y luego aumentar λ, por ejemplo multiplicándolo por un factor fijo mayor que 1. En este algoritmo γ juega ese rol y no hay parámetro ω. La actualización de λ propuesta en [23] (ver (3.13)) no es adecuada para un contexto no euclídeo: que el gradiente de L tenga proyección positiva en el gradiente de L_F no parece ser una buena condición ya que podría apuntar hacia afuera de la superficie (por ejemplo en un círculo, si L_c y L_F lo hacen en dirección tangencial con sentidos opuestos y el gradiente de L apunta hacia el centro, este puede tener proyección positiva en L_F pero es sin duda una mala elección, dando una dirección en la que no se descenderá y estancando al algoritmo).

Numéricamente hay problemas al alejarse de la superficie, ya que si un punto se encuentra lejos de esta, su distancia de Fermat a cualquier otro será muy grande. Esto puede

suceder para valores altos de λ, cuando el costo no tiene mucha importancia. Un punto que se alejó de la superficie sólo se moverá si la ventana h de todos los demás puntos es suficientemente grande. Esto motiva a tener un h grande, pero que sea grande a su vez no permitirá resolver detalles ni que las distribuciones se parezcan realmente (ver la imagen de la derecha de la Figura 3.11), o hasta podría empeorar detalles ya resueltos. Para evitar que los puntos salgan de la superficie podemos proponer que a cada paso un ε-cuantil no se agrande además de que el lagrangiano disminuya para efectivamente descender, con $\varepsilon \leq \delta$. De esta forma, la distancia de Fermat entre los puntos no se puede agrandar mucho y por ende los puntos no podrán "salirse" de la superficie. Podríamos a la vez proponer un enfriamiento del h, es decir, en vez de tomar siempre un δ-cuantil con δ fijo, podríamos considerar un δ que disminuya, empezando en δ_0 en las primeras iteraciones y terminando en ϵ en las últimas. La forma propuesta para actualizar δ es $\varepsilon + (\delta - \epsilon) * \eta$, con $\eta \in (0, 1)$ una tasa de enfriamiento fija cercana a 1. Con estas modificaciones, se obtuvieron buenos resultados para más superficies además de para las ya mencionadas con el algoritmo anterior.

Arrollado de dulce de leche

Volvemos al caso del arrollado de dulce de leche, donde ahora tomamos baricentro entre dos grupos de puntos en los extremos de la superficie, como muestra la siguiente figura.

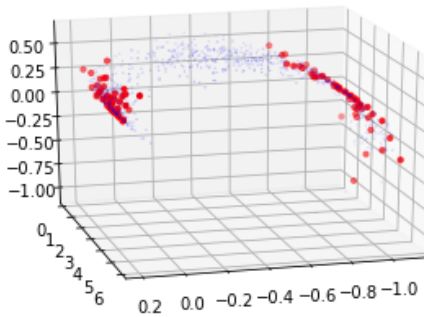


Figura 3.15: 1000 puntos en la superficie del arrollado, en rojo dos grupos a los que tomaremos baricentro.

Al aplicar el algoritmo de baricentro esperamos ver una distribución que se encuentre en un punto intermedio de la superficie y que tenga una varianza intermedia, ni tan pequeña como la del grupo del centro ni tan grande como la de la cola del arrollado. Esto se puede ver en la siguiente figura.

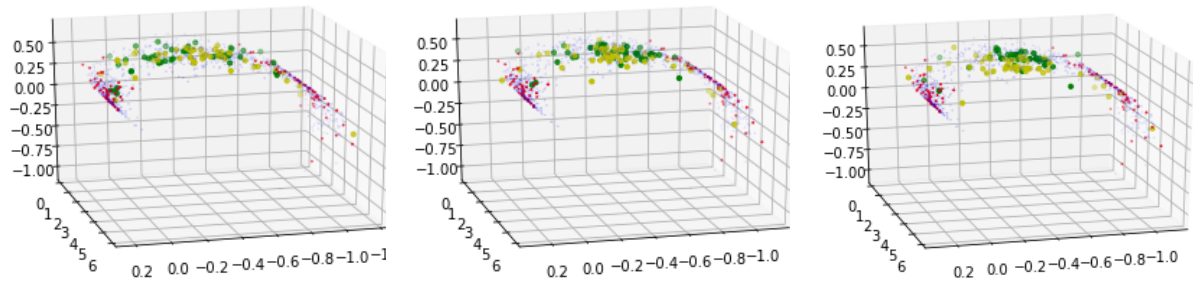


Figura 3.16: En amarillo los transportados del grupo de la derecha y en verde los del grupo de la izquierda. El algoritmo se corrió con parámetros $\lambda_0 = 0{,}01$, $\gamma = 1{,}5$, $\alpha = 2$, $k = 30$, $\lambda_{max} = 100$, $\eta = 0{,}95$. Izquierda: $\delta_0 = 0{,}9$, $\varepsilon = 0{,}8$. Centro: $\delta_0 = 0{,}9$, $\varepsilon = 0{,}8$. Derecha: $\delta_0 = 0{,}95$, $\varepsilon = 0{,}8$.

Un h grande es necesario para que se acerquen al principio las distribuciones y puedan parecerse, aunque sea, en media; por eso se tomo δ_0 como 0,9 o 0,95. Sin embargo, es importante que el "bandwidth" se achique luego para poder resolver detalles y que las distribuciones se parezcan en otras características. Aquí entra en juego la tasa de enfriamiento η y el valor ε que pondrá el tope inferior de enfriamiento. Podemos ver en la Figura 3.16 la importancia de los parámetros δ_0 y ε, que reflejan la importancia de como ir actualizando los valores de h, la ventana del núcleo. En la imagen de izquierda jamás achicamos el δ (pues $\delta_0 = \varepsilon$), por lo que si bien el h se achica al juntarse los puntos, no lo hace lo suficiente ya que siempre es muy cercano al diámetro de los transportados, lo que no permitirá resolver detalles. En la imagen del centro, δ se va enfriando entre 0,9 hasta 0,8 y las distribuciones se parecen más y se resuelven más detalles. Finalmente en la imagen de la derecha vemos más claramente este efecto.

Altas dimensiones

Si bien encontrar vecinos exactos (cosa que hacemos al utilizar un KD-Tree) no es muy eficiente para altas dimensiones, se busca ver en este ejemplo si la alta dimensionalidad afecta al algoritmo en general más allá del tiempo. Para esto, se realizó un ejemplo en 10 dimensiones de espacio ambiente pero con una superficie de dimensión 2, donde las dos primeras sean las significativas. Se muestra a continuación la distribución de los puntos y la superficie. La superficie de Fermat a considerar fue la unión de los puntos de las distribuciones y la superficie. El resultado obtenido fue el siguiente:

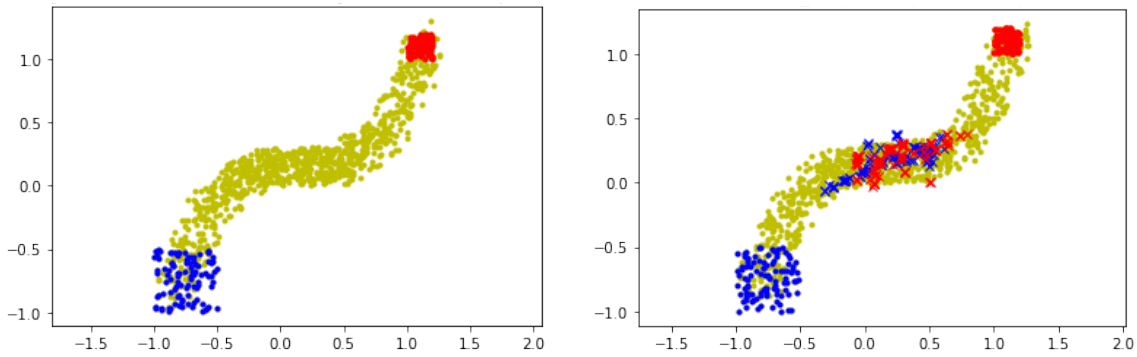


Figura 3.17: En puntos azules y rojos dos normales con distinta varianza, los puntos x_i. En amarillo la superficie de Fermat dada por puntos de la función x^3 con $x \in [-1,1{,}5]$ y con ruido uniforme dos dimensional en $[0,0{,}3] \times [0,0{,}3]$. Se grafican sólo dos de las 10 dimensiones de los puntos, las restantes son con ruido uniforme en $[0,0{,}3]$ para todos los puntos. Las cruces representan los transportados y_i. Se ejecutó el algoritmo con parámetros $\lambda_0 = 0{,}0005, \gamma = 1{,}5, \delta_0 = 0{,}95, \varepsilon = 0{,}7$.

z continuo

Tener valores de z continuos puede también ser frecuente. Una posible aplicación sería una donde la variable z sea irrelevante, aportando variabilidad que no deseamos y de la cual nos queremos independizar encontrando el baricentro. Aquí no tenemos grupos de datos que se distinguen por una etiqueta z como utilizamos en los ejemplos hasta aquí, sino que z es continuo y su valor conlleva una noción de distancia. Al permitir que z sea una variable continua y deje de ser categórica, se pone en juego la importancia de los núcleos K_b utilizados en la estimación de densidad condicional (3.10). Estos núcleos conllevan utilizar una distancia en el espacio de las z. Podemos tomar la distancia euclídea en este espacio inicialmente, para ver que el algoritmo se adapta sin ningún tipo de problema. Para esto, se tomó la misma superficie del caso anterior y se sortearon valores de z_i que darían lugar a muestras x_i, como muestra la imagen de la izquierda de la siguiente Figura 3.18. Al aplicar el algoritmo se encontró el baricentro en un caso de z continuo como muestra la imagen de la derecha de dicha figura. Todos los algoritmos utilizados se encuentran implementados en el repositorio `https://github.com/nicocheh/FermatOT`, donde se generan también los ejemplos sintéticos mostrados en este trabajo.

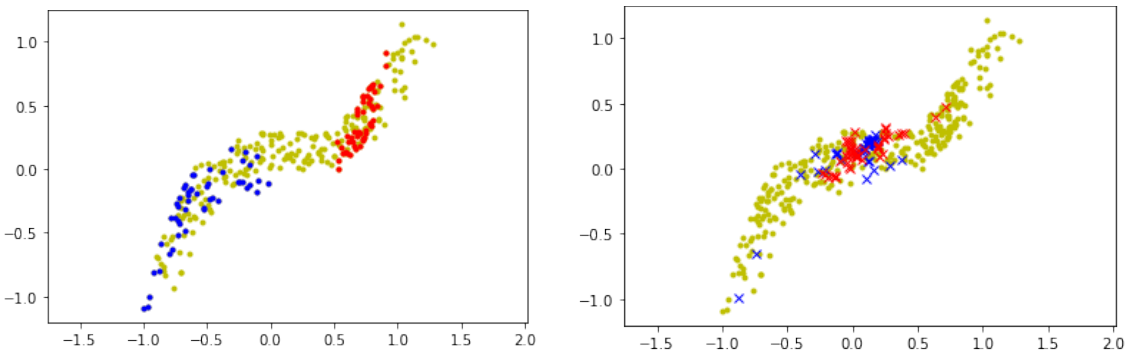


Figura 3.18: Los z_i fueron generados con dos normales de media $-0{,}5$ y $0{,}7$ y varianza $0{,}25$ y $0{,}1$ respectivamente. Los x_i se generaron como $(z_i, z_i^3 + \varepsilon_i)$ donde $\varepsilon_i \sim \mathcal{U}_{[-0{,}2,0{,}2]}$. Izquierda: se grafican en azul y rojo los x_i y en amarillo la superficie. Derecha: se grafican los transportados de los puntos iniciales (x_i, z_i). Se ejecutó el algoritmo con parámetros $\eta = 0{,}8, \lambda_0 = 0{,}01, \gamma = 1{,}5, \delta_0 = 1, \varepsilon = 0{,}7$.

Conclusiones

A lo largo de esta tesis hemos visto como combinar el problema de transporte óptimo y baricentro de Wasserstein con la distancia de Fermat, previamente introducidos en los capítulos 1 y 2 respectivamente. Considerando que hay muchos casos donde ciertos datos se ven muy bien explicados en dimensiones mucho menores a las de su espacio ambiente, es deseable poder transportarlos adecuadamente en una superficie de menor dimensión, donde realmente se encuentran y explican, en vez de en el espacio ambiente. Así, uno consideraría distancias geodésicas en la variedad subyacente aunque desconocida; más aún, la distancia de Fermat tiene en cuenta la densidad de los puntos también, dando más "confianza" a zonas donde hay más datos. Con esta motivación, la introdujimos como función de costo en el problema de transporte óptimo, comprobando que es adecuada para transportar dentro de la variedad. Se definió una versión discreta del gradiente de la distancia de Fermat para poder realizar descenso por el gradiente. La definición de dicho gradiente y los algoritmos propuestos no requerían recalcular la distancia de Fermat, por lo que esta podría ser precomputable con un conjunto de puntos s_i que definen la superficie por la cual queremos transportar los datos.

Para abordar el problema de transporte óptimo y baricentro con costo de Fermat se estudiaron varios enfoques: optimización combinatoria, con restricciones y minimax. La primera con dificultad de escalar a gran cantidad de datos y con la imposibilidad de generar nuevas muestras de la distribución objetivo; la segunda muy dependiente de las características impuestas como restricciones y la tercera como una buena solución a esto último aunque conllevando resolver un problema mucho más difícil y sin técnicas generales robustas. A la hora de resolver el problema del baricentro con este último enfoque, se siguió un método que facilitaba la maximización reduciéndola a solo un parámetro pero a cambio requería realizar estimaciones de densidad en dicha variedad. Las estimaciones se realizaron por medio de núcleos, que utilizan una distancia entre los puntos. Utilizar la distancia euclídea para esto no es adecuado tampoco ya que si bien localmente es similar a una distancia en la variedad, difiere mucho de una distancia deseable al utilizar ventanas grandes en los núcleos; en cambio, la distancia de Fermat sí resulta adecuada.

Trabajo Futuro

Esta sección resume líneas e ideas en las que se está trabajando y/o aún no se ha profundizado, que podrían ser interesantes para continuar y ahondar el trabajo realizado.

- Demostrar teóricamente que la propuesta de gradiente empírico de la distancia de Fermat converge al gradiente de la distancia en su versión poblacional. Un buen primer paso sería probarlo en superficies isométricas a un conexo de $\mathbb{R}^n$.
- Estimar la dimensión de la variedad, para poder escalar adecuadamente los núcleos, siguiendo [29] en vez de [28] a la hora de estimar densidades sobre la variedad.
- Encontrar ejemplos más allá de los datos sintéticos y aplicaciones a datos reales [34].
- Aplicar la versión empírica del gradiente de Fermat en otro contexto. El gradiente resulta útil en cualquier contexto de optimización y su cómputo es veloz dados los vecinos y habiendo precomputado las distancias de Fermat.
- Estudiar las características de la estimación de densidad por kernels de Fermat y su eficiencia respecto de otras formas de estimación de densidad en variedades desconocidas.
- Realizar una implementación que utilice ANN en vez de un KD-Tree, para poder ser eficiente en altas dimensiones a la hora de encontrar vecinos.
- Agregar alguna penalización por irse de la superficie, en vez de proponer que un cuantil no se reduzca.
- Utilizar la distancia de Fermat en la estimación de densidad de los z en el problema del baricentro (ver (3.10)). Esto podría combinarse con resolver el baricentro euclídeo o de Fermat, es decir usando costo y estimación euclídeos o de Fermat en y.

Bibliografía

[1] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. Foundations and Trends in Machine Learning, Vol. 11, Issue 5-6, Pages 355-607, 2019.

[2] Cedric Villani. Topics in Optimal Transportation. American Mathemathical Society - Graduate Studies in Mathematics 58, 2003.

[3] Glenn Hurlbert. A short proof of the Birkhoff-Von Neumann theorem. 2012.

[4] Ralph Tyrell Rockafellar. Convex analysis. Princeton Landmarks in Mathematics and Physics, 1996.

[5] Filippo Santambrogio. Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling. Birkhäuser, 2015.

[6] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis, 43(2):904–924, 2011.

[7] Pablo Groisman, Matthieu Jonckheere and Facundo Sapienza. Nonhomogeneous euclidean first-passage percolation and distance learning. arXiv: 1810.09398. 2018.

[8] Facundo Sapienza, Pablo Groisman and Matthieu Jonckheere. Weighted geodesic distance following Fermat's principle. International Conference on Learning Representation, 2018.

[9] Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319–2323, 2000.

[10] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. 2012b.

[11] `scikit-learn.org/stable/modules/generated/sklearn.neighbors.KDTree.html`

[12] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. Communications of the ACM, Vol. 18, No. 9, 1975.

[13] towardsdatascience.com/comprehensive-guide-to-approximate-nearest-neighbors-algorithms-8b94f057d6b6

[14] Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. Journal of Mathematical Analysis and Applications, 441(2):744–762, 2016.

[15] Yoav Zemel and Victor M Panaretos. Fréchet means and procrustes analysis in wasserstein space. Bernoulli, 2018.

[16] M. Kuang, E. G. Tabak. Sample-based optimal transport and barycenter problems. Communications on Pure and Applied Mathematics 72 (8), 1581–1630, 2019.

[17] Julien Rabin, Sira Ferradans and Nicolas Papadakis. Adaptive Color Transfer With Relaxed Optimal Transport. IEEE international Conference on Image Processing. Oct 2014.

[18] Gabriel Peyré. Numerical Optimal Transport and its Applications. mathematical-tours.github.io/book-basics-sources/ot-sources/TransportEN.pdf

[19] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning, 70:214–223. 2017.

[20] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha and Kilian Q Weinberger. Supervised word mover's distance. Advances in Neural Information Processing Systems, pages 4862– 4870, 2016.

[21] docs.scipy.org/doc/scipy/reference/optimize.minimize-trustconstr.html

[22] pythonot.github.io

[23] Esteban G. Tabak, Giulio Trigila, Wenjun Zhao. Distributional barycenter problem through data-driven flows. arXiv: 2104.14329. 2021.

[24] Facundo Sapienza. Distancia de Fermat y geodésicas en percolación euclídea: teoría y aplicaciones en Machine Learning. Tesis de Licenciatura, Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 2018.

[25] J. G. De Gooijer and D. Zerom. On conditional density estimation. Statistica, Neerlandica 57 (2), 159–176. 2003.

[26] M. Essid, D. F. Laefer, and E. G. Tabak, "Adaptive optimal transport". Information and Inference :A Journal of the IMA, vol. 8, no. 4, pp. 789–816. 2019.

[27] E. G. Tabak, G. Trigila, and W. Zhao. "Data driven conditional optimal transport". arXiv: 1910.11422. 2019.

[28] Bruno Pelletier. Kernel density estimation on Riemannian manifolds. Statistics & Probability Letters, Vol. 73, 3, 297-304. 2005

[29] Clément Berenfeld and Marc Hoffmann. Density estimation on an unknown submanifold. arXiv: 1910.08477. 2020

[30] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré and Jean-François Aujol. SIAM Journal on Imaging Sciences, 7(3), 1853-1882. 2014

[31] Eugenio Borghini, Ximena Fernández, Pablo Groisman and Gabriel Mindlin. Intrinsic persistent homology via density-based metric learning. arXiv: 2012.07621. 2020

[32] Nicolas Courty, Rémi Flamary, Devis Tuia and Alain Rakotomamonjy. Optimal Transport for Domain Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence , vol.PP, no.99, pp.1-1. 2016

[33] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems (pp. 2292-2300). 2013

[34] V. Peterson, N. Nieto, D. Wyser, O. Lambercy, R. Gassert, D.H Milone and R. Spies. Transfer Learning based on Optimal Transport for Motor Imagery Brain-Computer Interfaces. IEEE Transactions on Biomedical Engineering, Under Review. 2021.