UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Distancia de Fermat y geodésicas en percolación euclídea: teoría y aplicaciones en Machine Learning

Facundo Sapienza

Director: Dr. Pablo Groisman

Fecha de Presentación: Agosto de 2018

Índice general

Resumen

En la presente tesis se introduce la *distancia de Fermat* junto con su estimador. Dado un conjunto de puntos con densidad f soportada sobre una variedad $\mathcal{M}$, la distancia de Fermat contempla tanto f como $\mathcal{M}$, captando la estructura intrínseca de los puntos y haciéndola una excelente candidata para muchos problemas de estadística y Machine Learning. Mas aún, la convergencia del estimador de la distancia de Fermat se contextualiza dentro de la teoría de percolación euclídea de primera pasada. A lo largo de la tesis veremos aplicaciones así como demostraciones rigurosas pertinentes a la distancia de Fermat.

El presente trabajo está basado en las siguientes publicaciones:

- *Weighted Geodesic Distance Following Fermat's Principle* (2018); F. Sapienza, P. Groisman, M. Jonckheere; 6th International Conference on Learing Representations.
- *Geodesics in First Passage Percolation and Distance Learning* (2018); P. Groisman, M. Jonckeheere, F. Sapienza; en preparación.

Agradecimientos

Primero me gustaría agradecer a mis dos directores, Pablo Groisman y Matthieu Jonckheere. Desde un principio Patu me incentivo a ser creativo y buscar un tema de tesis que nos gustara a ambos y sobre el cual pudiésemos hacer algún avance. Arrancamos estudiando modelos de opinión, pasando por redes neuronales y pruebas de consistencia de clustering. Finalmente, una tarde como cualquiera otra en Aristas, junto con Matt surgió la idea de la ahora adoptada *Distancia de Fermat*. Apenas lo compartimos con Patu nos entusiasmamos todos y ahí arranca la historia de esta tesis.

Como toda investigación, fue una historia de muchas satisfacciones y un par de disgustos. Algoritmos que dan buenos resultados; demostraciones que parecen estar bien pero se caen nuevamente; noches de desvelo; trabajos aceptados en congresos internacionales en una comunidad completamente nueva para todos nosotros; resultados ya publicados que de repente empiezan a hacer cosas similares a las nuestras. Al final, como todo en la vida, siempre se trata de un camino con subidas y con bajadas, pero que vale la pena recorrer una y otra vez. Tal como nos enseñaron en las olimpiadas de matemática, es el placer de resolver problemas lo que nos hace elegir esto todos los días.

Me gustaría agradecer a la familia y a los amigos. A todos los que sin saber qué es lo que hace un físico o matemático, me apoyaron porque sabían que esto es lo que quiero. Sobre todo Mamá y Papá. Los títulos son para ustedes. A los que compartieron conmigo estos años maravillosos años en la universidad.

A todos los tutores y directores que tuve la suerte de tener a lo largo de la carrera. Augusto, Carlos, Leo, Matt, Patu. De todos ustedes aprendí y sigo aprendiendo cosas. De cada unos de ustedes me llevo algo y por eso gracias.

Un agradecimiento especial a todo Aristas, donde tuve la suerte de pasar los últimos 3 años aprendiendo, investigando, resolviendo problemas. Lo que aprendí trabajando en Aristas es invaluable y hoy día no sólo constituye parte de mi formación sino también de mi persona.

También me gustaría agradecer a la educación pública de nuestro país, en particular a la Universidad de Buenos Aires. No puedo dejar de lado el hecho de haber podido asistir a una universidad de excelencia y gratuita, oportunidad que sé que no todos tienen en el mundo, inclusive en nuestro país.

Para finalizar me gustaría agradecer a los magníficos jurados de esta tesis que tan amablemente aceptaron formar parte de esta historia, los profesores Pablo Ferrari y Esteban Tabak.

Muchas son las personas a la que agradecer y corta es cualquier cosa que pueda escribir para ellos. Simplemente, muchas gracias a todos.

Introducción

En muchas tareas de aprendizaje tales como clustering, clasificación, recomendación y reducción de dimensión, una noción de similaridad o distancia entre puntos no sólo es crucial para el problema en cuestión, sino que típicamente no es inmediata de definir. Tareas como la de agrupar puntos en clusters pueden depender mucho más de la medida de distancia con la cual se trabaja que del algortimo utilizado para realizar el agrupamiento. Algoritmos de aprendizaje basados en estimación de similitudes han tenido éxito en muchas aplicaciones: series temporales (Morse & Patel (2007)), clasificación de compuestos químicos (Barnard & Downs (1992)), datos genéticos (Lawson & Falush (2012)), texto (Wang et al. (2011)). Sin embargo, la dificultad de definir una buena métrica entre puntos se debe a dos problemas principales: la maldición de la dimensión y el hecho de que los datos típicamente suelen vivir en una superficie de dimensión mucho menor que la del espacio ambiente.

La maldición de la dimensión es un efecto que sufren todas las distancias cuanto el espacio donde se encuentran los puntos tiene alta dimensión y que tiene que ver con el hecho de que la resolución entre las distancias más pequeñas (puntos que están muy cerca) y las distancias más grandes (puntos más alejados) comienza a perderse a medida que la dimensión aumenta. De está manera, todos los puntos pasan a estar igual de cerca que de lejos. Consideramos el ejemplo mostrado en Bishop (2006). Sea en $\mathbb{R}^D$ la bola de radio unitario y nos preguntamos cuál es la fracción de volumen que se encuentra entre las franja de radio $r = 1 - \varepsilon$ y $r = 1$. El volumen de la bola de radio r en dimensión D está dado por $V_D(r) = \omega_D r^D$, donde ω_D es el volumen de la bola unitaria. Luego, dicha fracción está dada por

$$\frac{V_D(1) - V_D(1-\varepsilon)}{V_D(1)} = 1 - (1-\varepsilon)^D,$$

la cual es una cantidad que converge a 1 cuando $D \to \infty$. Es decir, en espacios de dimension grande la mayoría del volumen esta concentrada en la cáscara de la esfera. Por lo tanto, si sobre la esfera unitaria sampleamos puntos con alguna distribución, veríamos que la distancia de todos los puntos al origen se mueve en una pequeña franja alrededor de 1. Una explicación más precisa de este fenómeno la podemos encontrar en el siguiente resultado.

Teorema (Aggarwal et al. (2001), Teorema 2)**.** *Sean $\mathbf{x_1}$ y $\mathbf{x}_2$ puntos sampleados independientemente a partir de una distribución uniforme en $[0,1]^D$ y notemos por $|\cdot|_p$ la norma p en $\mathbb{R}^D$. Luego*

$$\lim_{D\to\infty} \mathbb{E}\left[\left(\frac{\text{máx}\{|\mathbf{x}_1|_p, |\mathbf{x}_2|_p\} - \text{mín}\{|\mathbf{x}_1|_p, |\mathbf{x}_2|_p\}}{\text{mín}\{|\mathbf{x}_1|_p, |\mathbf{x}_2|_p\}}\right) \cdot \sqrt{D}\right] = C\sqrt{\frac{1}{2p+1}}$$

donde C es alguna constante.

Este fenómeno dificulta la mayoría de las tareas en las cuales es necesario trabajar con distancias entre datos en espacios de alta dimensión. En tales casos es necesario recurrir a técnicas que permitan encontrar

representaciones en espacios de menor dimensión de los datos o que logren definir distancias que eviten este problema.

El otro punto clave es el de entender la geometría intrínseca y la dimensión en la que los puntos realmente se encuentran. Este es el caso en el cual los datos viven en una superficie de dimensión mucho menor que la del espacio ambiente, el cual es típicamente la situación en muchas aplicaciones ((Bengio et al., 2013)). A esta tarea se la conoce como *nonlinear dimensionality reduction* (NLDR). Consideremos por ejemplo un conjunto de fotografías donde se muestra un mismo rostro en distintas posiciones y con distinta luz, de Silva & Tenenbaum (2002). El objetivo es identificar variables intrínsecas o grados de libertad, como la orientación de la cámara o la intensidad de la luz, que parametrizan la superficie en la cual están contenidas las imágenes. Esta situación se modela a partir de la hipótesis de que el conjunto de datos proviene de una distribución de probabilidad $f : \mathcal{M} \subset \mathbb{R}^D \mapsto \mathbb{R}_{\geq 0}$, donde $\mathcal{M}$ es una superficie de dimensión d, es decir, $\mathcal{M}$ es localmente equivalente a $\mathbb{R}^d$. Típicamente se tiene $d \ll D$. El siguiente lema refleja el hecho de que si la cantidad de puntos n no es suficientemente grande, siempre existe una superficie de dimensión mucho más chica donde los puntos se encuentran, salvo un pequeño error.

Lema (Johnson-Lindenstrauss)**.** *Sean $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^D$ puntos arbitrarios y sea $\varepsilon > 0$. Luego, para algún $d = \mathcal{O}(\log(N)/\varepsilon^2)$ existen puntos $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \in \mathbb{R}^d$ tales que*

$$(1-\varepsilon)|\mathbf{x}_i| \leq |\mathbf{y}_i| \leq (1+\varepsilon)|\mathbf{x}_i| \quad \forall i$$
$$(1-\varepsilon)|\mathbf{x}_i - \mathbf{x}_j| \leq |\mathbf{y}_i - \mathbf{y}_j| \leq (1+\varepsilon)|\mathbf{x}_i - \mathbf{x}_j| \quad \forall i, j.$$

Más aún, en tiempo polinomial es posible encontrar una transformación lineal $L : \mathbb{R}^D \mapsto \mathbb{R}^d$ tal que $L(\mathbf{x}_i) = \mathbf{y}_i$ y que ambas condiciones se satisfagan con probabilidad mayor a $1 - 2/n$, Matoušek (2002).

El problema de encontrar una representación de menor dimensión que refleje la estructura de los datos (*manifold learning*) es un problema bien estudiado en los últimos años y que está intrínsecamente relacionado con aprender una distancia (*metric learning*). Ejemplos de dichas técnicas incluyen *multidimensional scaling* (Borg & Groenen (2003)), *t-distributed stochastic distance embedding* (van der Maaten & Hinton (2008)), *Spectral embedding* (Belkin & Niyogi (2003)), *Isometric mapping* (Isomap) y *C-Isomap* (Tenenbaum et al. (2000); de Silva & Tenenbaum (2002)).

Es importante remarcar que de todos estos métodos, solo *Isomap* y *C-Isomap* tienen la particularidad de estimar distancias por medio de geodésicas contenidas en la superficie $\mathcal{M}$. El trabajo donde se introduce *Isomap*, (Tenenbaum et al., 2000), remarca la mejora que se obtiene al definir una distancia que mida geodésicas sobre $\mathcal{M}$, en particular sobre conjuntos de datos formados por imágenes. Sin embargo, ninguno de estos métodos considera los valores que toma la densidad f sobre la misma para definir la distancia y por lo tanto la no-homogeneidad de los datos no se ve reflejada en la distancia.

El aporte de esta tesis es el de introducir la *la distancia de Fermat* junto a su estimador, una nueva métrica para espacios de alta dimension y típicamente no homogéneos. A diferencia de trabajos anteriores, nosotros no estamos estimando ni la distancia euclídea del espacio ni la geodésica, sino una distancia pesada por una potencia inversa de la densidad f. De esta manera, dos puntos van a estar cerca si y sólo si existe un camino corto que las conecte y esté contenido en una región de densidad alta. Esta distancia puede ser usada como input de algoritmos de reducción de dimensión y clustering.

La tesis se encuentra organizada en tres capítulos. En el primer capítulo se hace una revisión de algunas de las técnicas anteriormente mencionadas, así como se introducen los algorítmos de clustering e indicadores de performance que posteriormente vamos a utilizar. En el segundo capítulo se definen la distancia de Fermat y su estimador, se enumeran sus propiedades y se muestra su performance en datos sintéticos. Para finalizar, en el tercer capítulo se exhiben las demostraciones de consistencia del estimador, probando la convergencia del mismo en el régimen macroscópico.

Capítulo 1

Reducción de dimensión y clustering

El objetivo de este primer capítulo es el de introducir algunas técnicas para reducir la dimensión de un conjunto de puntos respectando la estructura intrínseca de los mismos lo más fehacientemente posible. Encontrar una representación de los datos en menor dimensión está relacionado con el problema de definir una métrica o distancia dentro de los mismos. A su vez, haremos un pequeño repaso por las ideas esenciales de clustering. Algunos ejemplos de métodos de reducción de dimensión aplicados al *MNIST dataset* se encuentran disponibles en *github.com/facusapienza21/dimensionality-reduction*.

Consideremos un conjunto de puntos $\mathbb{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^D$, donde n es el número total de puntos y D la dimensión del espacio ambiente. Si bien ninguno de los métodos requiere que la siguiente hipótesis sea cierta, siempre vamos a estar pensando que los puntos en $\mathbb{X}_n$ son una muestra i.i.d con alguna distribución de probabilidad con soporte en una superficie $\mathcal{M}$ de dimensión d, con $d \leq D$, y densidad $f : \mathcal{M} \mapsto \mathbb{R}_{\geq 0}$. A su vez, vamos a notar por $\mathbf{y}_i$ a la proyección del punto $\mathbf{x}_i$ en un espacio de dimensión menor.

1.1. Aprendizaje de distancias y variedades

1.1.1. Análisis de componentes principales

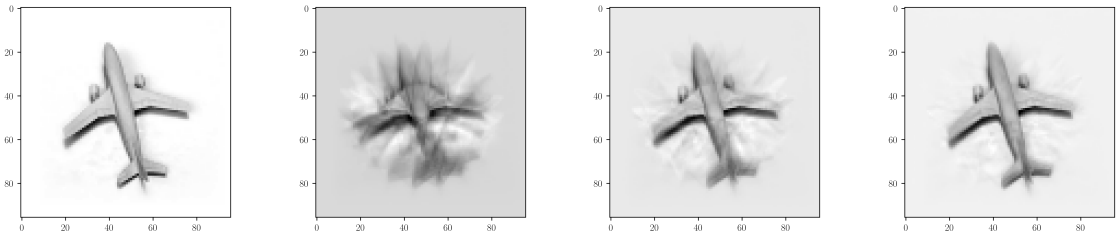
Análisis de componentes principales (*Principal Component Analysis* o simplemente PCA), Friedman et al. (2001), es un método que busca proyectar los datos en un hiperplano de dimensión menor y luego quedarse con la representación de los datos sobre esta variedad. Buscamos una transformación lineal $L : \mathbb{R}^d \mapsto \mathbb{R}^D$ de la forma

$$L(\mathbf{y}) = \mathbf{b} + A_d \mathbf{y},$$

donde $\mathbf{b} \in \mathbb{R}^D$, y $A_d \in \mathbb{R}^{D \times d}$ y proyecciones $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} \subset \mathbb{R}^d$ de manera tal que

$$\mathbf{b}, \{\mathbf{y}_i\}_{i=1,2,\ldots,n}, A_d = \text{argmín} \sum_{i=1}^{n} |\, \mathbf{x}_i - L(\mathbf{y}_i)|^2. \tag{1.1}$$

De esta manera, la minimización hace que cada punto $\mathbf{x}_i$ pase a estar asociado con su proyección ortogonal $L(\mathbf{y}_i)$ sobre la variedad lineal definida por la imagen de la transformación L. La proyección de todos los puntos $\mathbb{X}_n$ es tal que la varianza de los puntos $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ sea máxima. A su vez, la transformación L permite recuperar una aproximación $L(\mathbf{y}_i) \approx \mathbf{x}_i$, aunque la representación de los puntos pasa a vivir en un espacio de dimensión menor.



(a) Imagen original. (b) PCA con $d = 10$ (c) PCA con $d = 30$ (d) PCA con $d = 60$

Figura 1.1: ¿Cómo funciona PCA?. *El NORB dataset es un conjunto de datos formado por fotografías de 86x86 píxeles de distintos juguetes tomadas en distintos ángulos y con distintas condiciones de iluminación, LeCun et al. (2004). Consideremos un subconjunto de 680 imágenes correspondientes a la imagen original 1.1(a), calculamos la proyección con PCA para distintos valores de d y reconstruimos el vector en el espacio original por medio de la transformación $L : \mathbb{R}^d \mapsto \mathbb{R}^D$. Observamos con sólo $d = 60$ ya es posible representar una imagen perfectamente definida de la imagen original.*

La optimización (1.1) se realiza de manera eficiente mediante una descomposición en valores singulares (SVD). Si bien PCA busca una representación lineal de los datos (el cual no suele ser el caso en muchos ejemplos, como veremos más adelante), puede ser un muy buen primer paso cuando se trabaja con datos reales y permite hacer un preprocesamiento de los datos. Suponiendo que los datos viven en una variedad de dimensión $d \ll D$, primero se puede buscar una representación de los datos en $\mathbb{R}^{d_2}$, con $d \ll d_2 \ll D$, antes de efectuar algún otro método. Notemos que el Lema de Johnson-Linderstrauss da una pista de que esta puede ser una muy buena estrategia. A su vez, se reduce el tiempo de corrida del algoritmo que vaya a efectuarse posteriormente y reduce el ruido y el efecto de la maldición de la dimensión. En la Figura 1.1 se visualizan distintas proyecciones utilizando PCA para el mismo subconjunto de datos del NORB dataset, LeCun et al. (2004).

1.1.2. Escalamiento multidimensional

Escalamiento multidimensional (*multidimensional scaling* o MDS) es un método que busca encontrar una representación en baja dimensión de los puntos, pero en vez de minimizar el error que se comete al proyectar los datos en una superficie de menor dimensión (ecuación (1.1)) se busca minimizar la diferencia entre la distancia real y la distancia proyectada de los datos (Kruskal (1964); Borg & Groenen (2003); Friedman et al. (2001)). Si llamamos d_{ij} a la distancia entre los puntos $\mathbf{x}_i$ y $\mathbf{x}_j$ (por ejemplo, d_{ij} podría ser la distancia euclídea dada por $|\,\mathbf{x}_i - \mathbf{x}_j\,|$), buscamos $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \in \mathbb{R}^d$, con $d < D$, tales que minimicen la *stress function*:

$$\{\mathbf{y}_i\}_{i=1,2,\ldots,n} = \text{argmín} \sum_{i \neq j} \left(d_{ij} - |\mathbf{y}_i - \mathbf{y}_j|\right)^2 .$$

A diferencia de PCA, donde la minimización se realiza fácilmente mediante métodos lineales, en MDS la minimización se realiza por algún otro método como descenso por el gradiente, lo cual se ve reflejado en tiempos de corrida más largos. Entre las variantes de MDS destacamos:

- *Sammon mapping*: Busca minimizar

$$\sum_{i \neq j} \frac{\left(d_{ij} - |\mathbf{y}_i - \mathbf{y}_j|\right)^2}{d_{ij}}.$$

  De esta manera, se penaliza más a puntos cercanos que lejanos, permitiendo recuperar mejor la estructura local de los datos.

- *Local MDS*: Fijado un valor $k \in \mathbb{N}$, sea $(\mathbb{X}_n, \mathcal{N})$ el grafo de k-vecinos más cercanos simetrizado, donde $(i,j) \in \mathcal{N}$ si $\mathbf{x}_i$ es uno de los k vecinos más cercanos de $\mathbf{x}_j$ o viceversa. Luego se busca minimizar

$$\sum_{(i,j)\in\mathcal{N}} \left(d_{ij} - |\mathbf{y}_i - \mathbf{y}_j|\right)^2 - \tau \sum_{(i,j)\notin\mathcal{N}} |\mathbf{y}_i - \mathbf{y}_j|$$

donde τ es algún parámetro positivo. El primer término busca acercar cosas que están cerca mientras que el segundo busca alejar cosas lejanas.

1.1.3. Isomap

Dados dos puntos sobre una variedad $\mathcal{M}$, la geodésica entre dos puntos se define como la curva contenida en la variedad con menor longitud que las conecta. Notemos que la longitud de la geodésica define una distancia entre puntos que contempla la estructura de la variedad $\mathcal{M}$, independientemente de cómo sea la distancia euclídea entre pares de puntos.

Isomap, Tenenbaum et al. (2000), es una técnica que estima la longitud de las geodésicas sobre la variedad $\mathcal{M}$ donde están soportados los datos y luego realiza una proyección basada en dicha distancia. El algoritmo de *Isomap* funciona de la siguiente manera:

1. Dado $k_{\text{Isomap}} \in \mathbb{N}$, se construye el grafo $(\mathbb{X}_n, E)$ de k_{Isomap} vecinos más cercanos, donde $(\mathbf{x}_i, \mathbf{x}_j) \in E$ si $\mathbf{x}_i$ es un k_{Isomap} vecino más cercano del $\mathbf{x}_j$ o viceversa.
2. Para cada par de puntos, se calcula el camino mínimo que los conecta donde el peso de cada arista del grafo es $|\,\mathbf{x}_i - \mathbf{x}_j\,|$. Es decir, dados $\mathbf{p}, \mathbf{q} \in \mathbb{X}_n$, buscamos

$$d_{\text{graph}}(\mathbf{p},\mathbf{q}) = \min_{\substack{(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K) \subset \mathbb{X}_n^K \\ (\mathbf{y}_i, \mathbf{y}_{i+1}) \in E}} \sum_{i=1}^{K-1} |\mathbf{y}_{i+1} - \mathbf{y}_i|. \tag{1.2}$$

3. Con la longitud del camino mínimo, se realiza una proyección en un espacio de dimensión menor por medio de MDS.

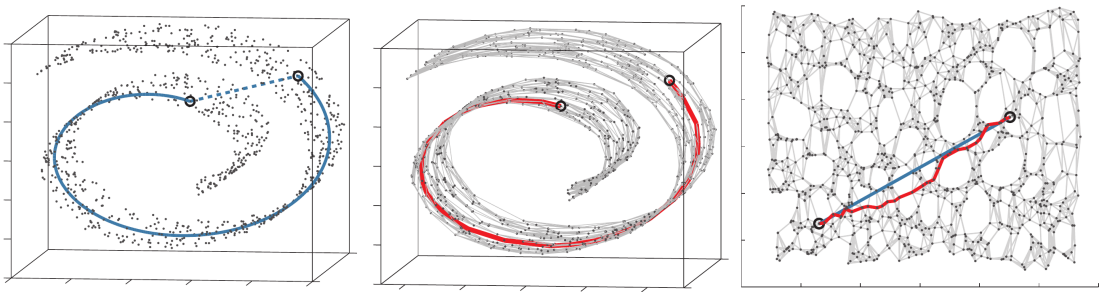
Sea $d_{\text{geodesic}}(\mathbf{p},\mathbf{q})$ la longitud de la geodésica contenida en $\mathcal{M}$ que conecta a los puntos $\mathbf{p}, \mathbf{q} \in \mathbb{X}_{n_0}$ para algún n_0. Luego se puede probar que, dados $\lambda_1, \lambda_2, \mu > 0$, se tiene

$$1 - \lambda_1 \leq \frac{d_{\text{geodesic}}(\mathbf{p},\mathbf{q})}{d_{\text{graph}}(\mathbf{p},\mathbf{q})} \leq 1 + \lambda_2 \tag{1.3}$$

sucede con probabilidad al menos $1 - \mu$ para n suficientemente grande, Bernstein et al. (2000).

Una buena manera de comprender cómo funciona el algoritmo de Isomap es mediante el conocido *Swiss Roll*. Consideremos en tres dimensiones un conjunto de puntos soportados sobre una superficie de dimensión dos enrollada sobre sigo misma (1.2(a)). La idea es buscar una manera de desenrollar el *Swiss Roll*, encontrar una buena representación en dos dimensiones de la misma y medir la distancia sobre esta proyección (lo cual es equivalente a medir la longitud de las geodésicas). Para una elección adecuada del parámetro k_{Isomap} se puede observar cómo el camino mínimo que conecta dos puntos va pegado a la superficie, de manera tal que su longitud es un buen estimador de la geodésica (1.2(b)). Por último, si efectuamos una proyección por medio de MDS en dos dimensiones podemos recuperar una representación de los datos en dimensión menor (1.2(c)). Por otro lado, la Figura 1.3 muestra el resultado que se obtiene cuando los datos están formados por fotografías reales de una mano en distintas posiciones.

Si bien *Isomap* da muy buenos resultados y permite definir una distancia que refleja la estructura intrínseca de los datos, no considera la densidad de probabilidad subyacente. Independiente de como sean sam-



(a) Geodesica del *Swiss Roll*. (b) Estimador de la geodésica que resulta de *Isomap*. (c) Proyección con MDS.

Figura 1.2: ¿Cómo funciona Isomap?. *El algoritmo de Isomap es capaz de encontrar la estructura intrínseca de los datos y construir una representación en dimensión menor que represente adecuadamente las distancia. Notemos que es deseable que la distancia entre los puntos en 1.2(a) sea medida por medio de geodésicas y no por medio de la distancia euclídea. De esta manera, el algoritmo de Isomap calcula la distancia pero moviéndose siempre localmente entre pares de puntos que realmente sean parecidos, evitando así la maldición de la dimensión. Imagen extraida de Tenenbaum et al. (2000).*

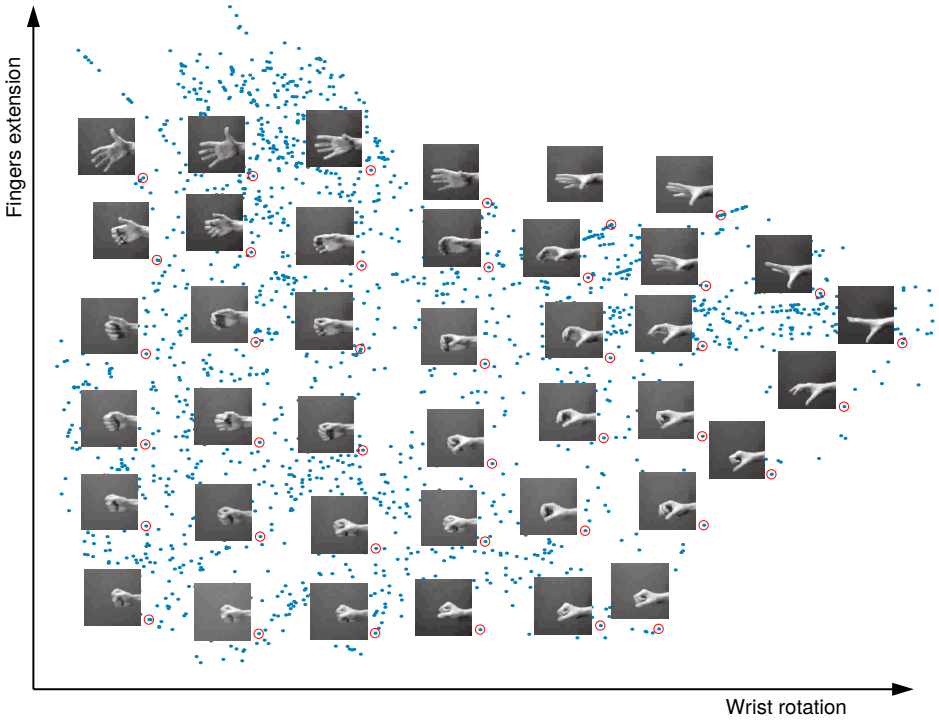


Figura 1.3: Isomap sobre datos reales*. Consideremos un conjunto de datos formado por imágenes de una misma mano en distintas posiciones, donde la mano se mueve con dos grados de libertad: puede girar en torno a su eje o puede cerrarse y abrirse. Es natural pensar que la dimensión intrínseca de estos datos es dos, aunque las imágenes estén formadas miles de píxeles. Sin embargo, Isomap permite proyectar los datos en dos dimensiones de manera consistente con estos dos grados de libertad. Imagen obtenida del sitio* `http://web.mit.edu/cocosci/isomap/isomap.html`.

pleados los puntos en la variedad $\mathcal{M}$, el estimador de Isomap converge a la geodésica en el sentido (1.3). El estimador de la distancia de Fermat que definiremos en el próximo capítulo va a contemplar tanto la estructura de la variedad $\mathcal{M}$ como la densidad de puntos sobre la misma.

C-Isomap

C-Isomap es una generalización de *Isomap*. Tal como está presentado en el trabajo introductorio, de Silva & Tenenbaum (2002), el problema de encontrar una parametrización (y a su vez, una distancia) sobre la superficie $\mathcal{M}$ donde están soportados los datos puede entenderse de la siguiente manera. Dado un conjunto de puntos $\mathbb{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n\} \subset \mathcal{M}$ buscamos una función $h : \mathcal{Y} \subset \mathbb{R}^d \mapsto \mathcal{M} \subset \mathbb{R}^D$ y puntos $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\} \subset \mathcal{Y}$ tales que $\mathbf{x}_i = h(\mathbf{y}_i)$. Luego, entendemos la distancia entre los puntos $\mathbf{y}_i$ y $\mathbf{y}_j$ como la distancia intrínseca entre los puntos $\mathbf{x}_i$ y $\mathbf{x}_j$. Dependiendo de las hipótesis que impongamos sobre la transformación h, vamos a obtener distintas representaciones. En el caso donde h es una isometría (es decir, que localmente preserva las longitudes y los ángulos) se recupera *Isomap*: la curva de menor longitud que conecta los puntos $\mathbf{y}_i$ y $\mathbf{y}_j$ (es decir, la recta) coincide con la geodésica entre $\mathbf{x}_i$ y $\mathbf{x}_j$ cuando se aplica la transformación h.

Asumiendo que la transformación h es conforme, es decir, preserva localmente los ángulos, *C-Isomap* define un estimador para recuperar la distancia en la preimagen $\mathcal{Y}$. El estimador se obtiene a partir de los mismos tres pasos que definen *Isomap* pero remplazando el segundo paso por

2. Para cada par de puntos, se calcula el camino mínimo que los conecta donde el peso de cada arista del grafo es $|\,\mathbf{x}_i - \mathbf{x}_j\,|/\sqrt{M_i M_j}$ y M_i es la distancia media del punto $\mathbf{x}_i$ a sus k_{Isomap} vecinos más cercanos.

Esta generalización permite trabajar con una familia más grande de transformaciones h. Notemos que en este caso el estimador que devuelve el algoritmo no coincide con la longitud de la geodésica, sino que es un enfoque distinto donde se busca recuperar la geometría de la preimagen $\mathcal{Y}$.

1.1.4. *t*-SNE

t-Stochastic Neighbor Embedding, o simplemente t-SNE, es un algoritmo de reducción de dimensión introducido en van der Maaten & Hinton (2008) que surge a partir de una pequeña (pero sumamente importante) variación del método SNE. El enfoque de SNE es definir distribuciones de probabilidad a partir de la distancia original y proyectada de los puntos. Concretamente, dados dos puntos $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{X}_n$ se define

$$p_{j|i} = \frac{\exp(-|\,\mathbf{x}_i - \mathbf{x}_j\,|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-|\,\mathbf{x}_i - \mathbf{x}_k\,|^2/2\sigma_i^2)}, \tag{1.4}$$

donde σ_i es un parámetro y ponemos $p_{i|i} = 0$. Notemos que $p_{j|i}$ puede ser interpretado como la probabilidad de elegir al punto $\mathbf{x}_j$ como vecino de $\mathbf{x}_i$ cuando las probabilidades alrededor de cada punto vecino son distribuciones normales. Para los puntos proyectados $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ se define una probabilidad de la misma manera:

$$q_{j|i} = \frac{\exp(-|\mathbf{y}_i - \mathbf{y}_j|^2)}{\sum_{k\neq i}\exp(-|\mathbf{y}_i - \mathbf{y}_k|^2)}. \tag{1.5}$$

Luego, los puntos proyectados son elegidos de tal manera que minimicen la siguiente función de costo dada por la divergencia de Kullback-Leibler $D_{KL}(\cdot|\cdot)$ entre las distribuciones de probabilidad inducidas por (1.4) y (1.5):

$$\{\mathbf{y}_i\}_{i\leq n} = \text{argmín}\sum_i D_{KL}(P_i|Q_i) = \text{argmín}\sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}, \tag{1.6}$$

donde $P_i = \sum_j p_{i|j}$ y $Q_i = \sum_j q_{i|j}$. SNE realiza una búsqueda sobre todos los posibles valores de σ_i de manera que el parámetro de *perplexity* sea el mismo para todos los puntos $\mathbf{x}_i$. El parámetro de *perplexity* esta definido como $2^{H(P_i)}$, siendo $H(\cdot)$ la entropía de Shannon dada por

$$H(P_i) = -\sum_j p_{i|j}\log p_{i|j}, \tag{1.7}$$

la cual representa la cantidad efectiva de vecinos que la distribución P_i llega a observar.

Si bien SNE es una idea muy elegante, en la práctica presenta dos principales problemas. El primero de ellos es la dificultad que presenta minimizar la función de costo. El segundo problema es conocido como *clowding problem* y sucede cuando se desea proyectar un conjunto de puntos con dimensión mayor en uno de dimensión menor. Por ejemplo, es fácil observar que es imposible efectuar una proyección en dos dimensiones de tres puntos equidistantes entre sí y que refleje correctamente el vecindario de cada punto. El efecto que esto tiene sobre la minimización (1.6) es el de colapsar varios puntos en una misma coordenada en el espacio proyectado. Hay varias maneras de evitar esto. Una es introduciendo un término repulsivo entre los pares de puntos que evite el colapso de puntos.

Las modificaciones que introduce t-SNE son:

1. La función de costo es remplazada por la divergencia de Kullback-Leibler entre la distribuciones globales P y Q:
$$C = D_{KL}(P|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{1.8}$$

2. Define la probabilidad p_{ij} en el espacio de dimensión alta como la probabilidad condicional simetrizada, es decir, $p_{ij} = (p_{i|j} + p_{j|i})/2n$. Esto asegura que todos los puntos contribuyan a la función de costo de manera significativa. Esto asegura que $\sum_j p_{ij} > 1/2n$.

3. Modifica la distribución $q_{j|i}$ cambiando la distribución normal alrededor de los puntos por una distribución t de Student con un grado de libertad (o distribución Cauchy), es decir,
$$q_{j|i} = \frac{(1 + |\mathbf{y}_i - \mathbf{y}_j|^2)^{-1}}{\sum_{k \neq l}(|\mathbf{y}_k - \mathbf{y}_l|^2)^{-1}}.$$
Dado que la distribución t de Student tiene una cola pesada respecto que la distribución normal, puntos que están muy cercanos en el espacio original y no pueden ser proyectados adecuadamente a un espacio de dimensión menor respetando la distancia mediante SNE pueden ser representados a partir de una distancia fija mediante t-SNE, evitando el *clowding problem*. Otra manera de entender esto es observando que el volumen que ocupa una determinada población proveniente de un distribución t de Student es mayor que la de una normal, de manera tal que hay más espacio para acomodar el volumen de puntos proveniente de las distribuciones normales del espacio original.

En la Figura 1.4 una proyección en dos dimensiones del MNIST dataset mediante t-SNE. El MNIST dataset consiste en imágenes de 28×28 píxeles en escala de grises de los diez dígitos (del 0 al 9) escritos por personas. Se puede ver como la proyección permite identificar los clusters correspondientes a cada uno de los dígitos.

1.2. Clustering

El problema de encontrar grupos de datos que compartan propiedades comunes (clustering de aquí en adelante) es uno de los problemas clásicos más estudiados y con múltiples aplicaciones en Machine Learning y estadística. A grandes rasgos, podemos organizar la tarea de clustering en tres pasos. La distinción entre ellos puede ser más o menos difusa dependiendo del problema. Las tres instancias son:

1. **Representación de los datos.** Dependiendo de la naturaleza de los datos, cada punto del conjunto de datos va a estar representado por variables que pueden ser cuantitativas, ordinales o categóricas. La tarea de representar la información en una estructura de datos adecuada es una tarea delicada que depende mucho del problema a tratar. A su vez, si los datos viven en

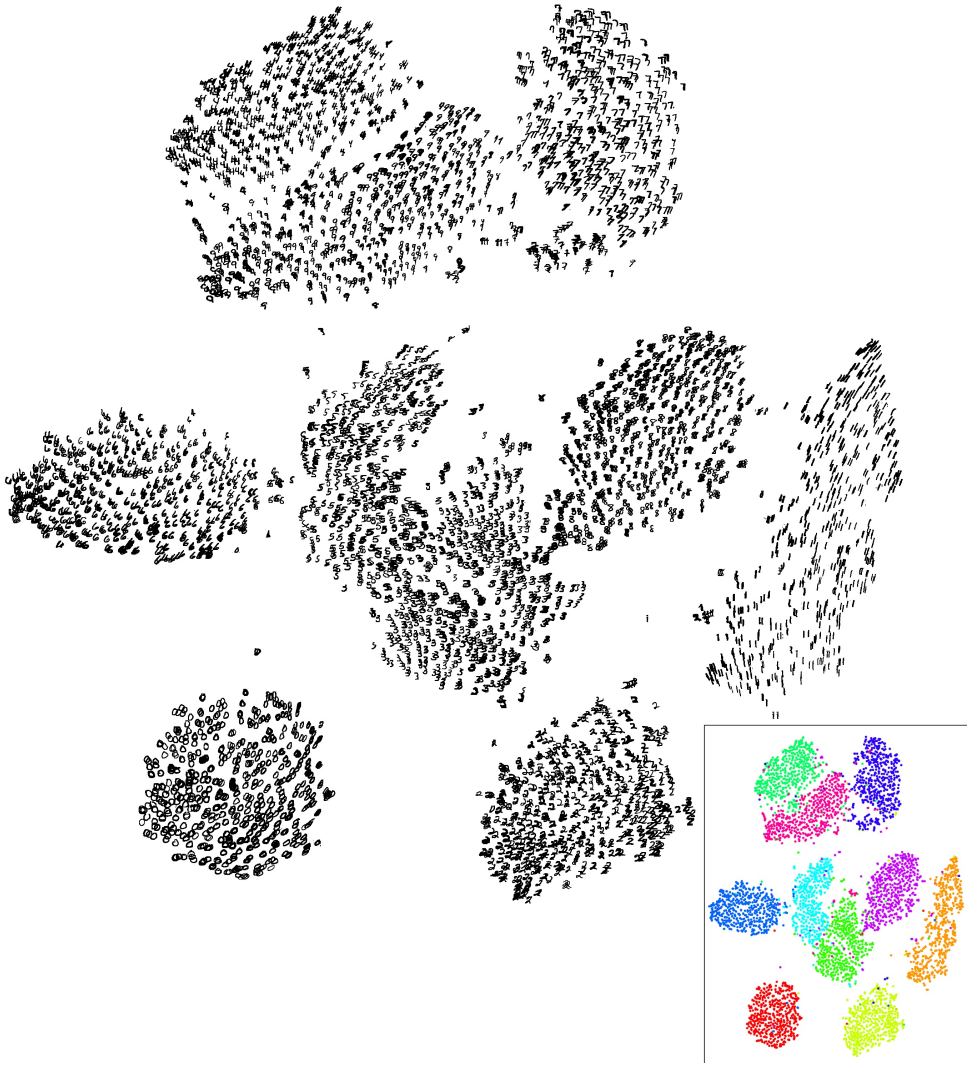


***Figura 1.4: Proyección mediante t-SNE de MNIST**. Se puede observar que el algoritmo de t-SNE es capaz de encontrar una representación de los puntos que refleja la estructura de clusters que se espera encontrar dada la naturaleza de los datos. Observar que cada punto representado es en realidad una miniatura de la imagen proyectada. Imagen obtenida de van der Maaten & Hinton (2008).*

un espacio de dimensión muy grande, puede ser deseable realizar primero una representación en menor dimensión y trabajar con la representación en dicho espacio.

2. **Distancia entre datos.** Una vez que se tiene una representación de los datos, es necesario definir una métrica o similaridad que trate de representar lo más fehacientemente posible la noción de semejanza que se desea reflejar en los datos. Por ejemplo, si los datos están representados por variables cuantitativas, una distancia posible sería la euclídea. Si las variables fueran categóricas, podría ser algo que refleje que tan parecidas son las categorías que se representan. Si los datos fueran imágenes, podríamos representar los datos como una o varias matrices donde en cada entrada se representa la escala de grises o un código RGB y luego utilizar la norma Frobenius (error cuadrático) como distancia. Sin embargo, en este caso tenemos el problema de que la distancia no capta si dos píxeles distintos de la imagen están cerca o no. Por lo tanto, puede ser deseable utilizar otra estructura para representar los datos o directamente modificar la distancia.

3. **El Algoritmo.** Una vez que tenemos una representación de los datos y definida una distancia entre ellos, el último paso consiste en efectuar un algoritmo de clustering que encuentre los grupos de puntos que son más parecidos entre sí que con el resto de los puntos.

Existen muchos algoritmos de clustering, entre ellos destacamos *K-means, hierarchical clustering, DBSCAN, spectral clustering* y *mean shift*. Cada uno de ellos tiene distintas ventajas y desventajas cuyo análisis escapa al objetivo de está tesis. Dado que el objetivo del presente trabajo es el de evaluar la performance del estimador de la distancia de Fermat, cantidad que introduciremos en el próximo capítulo, vamos a elegir uno de los algoritmos de clustering para trabajar. Vamos a trabajar con el algoritmo de K-medoids, una variante del algoritmo de K-means que permite trabajar con conjuntos de datos donde sólo se conoce la distancia entre ellos (sin necesidad de tener una representación de los mismos).

1.2.1. K-MEANS

Dado un conjunto de puntos $\mathbb{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, un agrupamiento o clustering está dado por una partición $\mathcal{C} = \{U_i\}_{i \leq K}$ que cumple

$$\mathbb{X}_n = \bigcup_{i=1}^{K} U_i \qquad U_i \cap U_j = \emptyset \quad i \neq j$$

donde K es la cantidad de clusters, Friedman et al. (2001). Luego, una manera de formular el problema de clustering es a partir de un problema de optimización donde se busca minimizar la distancia entre los puntos dentro de un cluster. Por ejemplo, dada una distancia $\ell(\cdot, \cdot)$ sobre $\mathbb{X}_n$, podemos buscar minimizar

$$W(\mathcal{C}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in U_k^2} \ell(\mathbf{x}_i, \mathbf{x}_j). \tag{1.9}$$

Notemos que así formulado, el problema sólo tiene sentido cuando el número de clusters K está fijo, pues la función de costo disminuye trivialmente cuando permitimos mayor número de particiones.

Dentro de este marco es que se encuentra el algoritmo de K-means. K-means es el algortimo clásico de clustering (Jain (2010)) y está diseñado para el caso en el cual los datos están descriptos mediante variables cuantitativas y la distancia entre datos está dada por la distancia euclídea. La función de costo es

$$W_K(\mathcal{C}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in U_k^2} |\, \mathbf{x}_i - \mathbf{x}_j \,|^2 = \sum_{k=1}^{K} |U_k| \sum_{\mathbf{x} \in C_k} |\, \mathbf{x} - \bar{\mathbf{x}}_k|^2,$$

donde $|U_k|$ es la cantidad de puntos del cluster U_k y $\bar{\mathbf{x}}_k$ es el punto medio de los puntos del cluster U_k dado por

$$\bar{\mathbf{x}}_k = \frac{1}{|U_k|} \sum_{\mathbf{x} \in U_k} \mathbf{x}. \tag{1.10}$$

Por lo tanto, cada cluster queda completamente caracterizado por su centro $\bar{\mathbf{x}}_k$ y cada punto en $\mathbb{X}_n$ pasa a formar parte del cluster asociado al centro más cercano.

Dado que la cantidad de particiones posibles de $\mathbb{X}_n$ en K clusters crece exponencialmente en n, es imposible encontrar una solución exacta. Sin embargo, remplazando los puntos medios $\bar{\mathbf{x}}_k$ por un punto libre $\mathbf{m}_k$ podemos reescribir el problema de optimización como

$$\min_{\mathcal{C}, \{\mathbf{m}_k\}_{k \leq K}} \sum_{k=1}^{K} |U_k| \sum_{\mathbf{x} \in C_k} |\, \mathbf{x} - \mathbf{m}_k|^2 \tag{1.11}$$

El algoritmo de K-means va minimizando alternadamente entre las particiones $\mathcal{C}$ y los centros $\{\mathbf{m}_k\}_{k \leq K}$ de la siguiente manera:

1. Dada un agrupamiento $\mathcal{C}$, se buscan $\{\mathbf{m}_k\}_{k \leq K}$ de manera de minimizar (1.11). Dicha minimización es inmediata y se obtiene a partir de (1.10).

2. Dados los centros $\{\mathbf{m}_k\}_{k\leq K}$ se define una nueva partición que minimice (1.11) a partir de

$$U_k = \left\{ \mathbf{x} \in \mathbb{X}_n : |\,\mathbf{x} - \mathbf{m}_k| = \underset{j\leq K}{\text{argmín}}\, |\,\mathbf{x} - \mathbf{m}_j| \right\} \tag{1.12}$$

3. Se repiten 1 y 2 hasta que no se actualicen los clusters.

El algoritmo asegura llegar a un mínimo local de la función de costo pero no a un mínimo absoluto. Típicamente se realizan varias iteraciones del algoritmo con distintas asignaciones iniciales y se elije aquella que minimice la función de costo (1.11).

1.2.2. K-medoids

K-medoids es una modificación del algoritmo de K-means, Friedman et al. (2001). A diferencia de K-means, no necesita que los datos estén representados por medio de variables cuantitativas ni necesita medir una función de error de la forma (1.11). El único input del algoritmo es la distancia $\ell(\cdot,\cdot)$ entre todos los pares de puntos. El objetivo es minimizar una función de costo como (1.9). Lo que se hace es modificar el paso 1 del algoritmo y remplazarlo por

$$\mathbf{m}_k = \underset{\mathbf{m}\in U_k}{\text{argmín}} \sum_{\mathbf{x}\in U_k} \ell(\mathbf{m}, \mathbf{x}),$$

es decir, en vez de actualizar $\mathbf{m}_k$ como el centro geométrico de los puntos del cluster U_k, lo remplazamos por el punto del cluster más cercano a todos los demás. Esto no solamente da una versión más robusta del algoritmo de K-means, sino que permite trabajar con datos que no necesariamente están soportados en un espacio con estructura geométrica y donde lo único que se conoce es la distancia entre puntos.

1.2.3. Performance

Existen distintos indicadores para evaluar la performance de un algoritmo de clustering. Dentro de los problemas de clasificación, el cual engloba muchas técnicas y algoritmos distintos, podemos diferenciar aquellos que son *supervisados* de los que son *no supervisados*. Como su nombre indica, en un problema supervisado conocemos la verdadera clasificación de los datos y dicha información es utilizada para buscar criterios que permita clasificar correctamente otra nueva familia de datos. El ejemplo clásico de un problema de clasificación supervisado seria una regresión lineal. Por el contrario, en un problema no supervisado no se conoce la naturaleza de los datos y se buscan algoritmos que aprendan la estructura interna de los datos. Un ejemplo de clasificación no supervisada es un algoritmo de clustering, como los que mencionamos anteriormente.

Dependiendo de si el problema es supervisado o no, el criterio con el cual se evalúa la performance del clasificador varía. En el caso supervisado, típicamente se definen medidas que contrasten el resultado del clasificador con la verdadera clasificación de los datos. Por otro lado, para los problemas que son no supervisados se suele definir una función de costo que se busca minimizar. A lo largo de la tesis nos concentraremos en desarrollar herramientas para tratar problemas no supervisados, si bien la técnica puede extenderse al caso supervisado. Sin embargo, dado que vamos a trabajar con datos de los cuales se conoce su verdadera clasificación, vamos a usar indicadores que evalúen la performance de estás técnicas como si fuera un problema supervisado (si bien la verdadera clasificación no es utilizada por el algoritmo y sólo se usa para evaluar la performance al final).

Existen distintas maneras de cuantificar qué tan parecidas son dos clasificaciones $\mathcal{C} = \{U_i\}_{i\leq K}$ y $\tilde{\mathcal{C}} = \{V_j\}_{j\leq \tilde{K}}$.

Partición	V_1	V_2	$\dots$	$V_{\tilde{K}}$	Suma
U_1	n_{11}	n_{12}	$\dots$	$n_{1\tilde{K}}$	a_1
U_2	n_{21}	n_{22}	$\dots$	$n_{2\tilde{K}}$	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_K	n_{K1}	n_{K2}	$\dots$	$n_{K\tilde{K}}$	a_K
Suma	b_1	b_2	$\dots$	$b_{\tilde{K}}$	N

Cuadro 1.1: Tabla de contingencias. *Dadas dos particiones $\mathcal{C} = \{U_i\}_{i\leq K}$ y $\tilde{\mathcal{C}} = \{V_j\}_{j\leq \tilde{K}}$ se define la tabla de contingencias a partir de $n_{ij} = |U_i \cap V_j|$.*

- **Adjusted mutual information.** Es un indicador basado en conceptos de la teoría de la información y cuantifica que tanta información se gana o se pierde al pasar de una clasificación a otra. Dadas $\mathcal{C}$ y $\tilde{\mathcal{C}}$ se define la tabla de contingencias (Tabla 1.1) como $n_{ij} = |U_i \cap V_j|$, donde $i = 1, 2, \dots, K$ y $j = 1, 2, \dots, \tilde{K}$. Notemos que $P(i, j) = n_{ij}/n$ define una distribución de probabilidad conjunta con marginales $p_i = |U_i|/n$ y $q_j = |V_j|/n$. Luego se define la *mutual information* entre $\mathcal{C}$ y $\tilde{\mathcal{C}}$ como la entropía mutua de las distribuciones marginales, Meilă (2007),

$$MI(\mathcal{C}, \tilde{\mathcal{C}}) = \sum_{i=1}^{K}\sum_{j=1}^{\tilde{K}} P(i, j) \log \frac{P(i, j)}{p_i q_j}.$$

  Notemos que en el caso donde ambas distribuciones son independientes y vale $P(i, j) = p_i q_j$ la información mutua es idénticamente cero. Es interesante remarcar que a partir de la información mutua es posible construir una métrica dentro del espacio de clasificaciones. Dada la entropía de Shannon definida como

$$H(\mathcal{C}) = -\sum_{i=1}^{K} p_i \log p_i$$

  se define la *variation of information* entre las particiones como

$$VI(\mathcal{C}, \tilde{\mathcal{C}}) = H(\mathcal{C}) + H(\tilde{\mathcal{C}}) - 2 \cdot MI(\mathcal{C}, \tilde{\mathcal{C}})$$

  La *variation of information* cumple las propiedades de ser positiva para todo par $\mathcal{C}$ y $\tilde{\mathcal{C}}$ y ser igual a cero si y sólo si ambas particiones coinciden; es simétrica; y cumple la desigualdad triangular. Si bien la mutual information cuantifica qué tanto se parecen dos particiones, no cumple la propiedad de ser *corrected for chance*, es decir, que toma un determinado valor (cero) cuando las dos particiones fueron elegidas bajo alguna hipótesis nula. De esta manera, no queda muy claro como interpretar su valor, lo que representa y cómo se compara con otros índices. Para ello, en Hubert & Arabie (1985) proponen una fórmula general para corregir cualquier índice dada por

$$\text{Adusted\_Index} = \frac{\text{index} - \text{expected\_Index}}{\text{max\_Index} - \text{expected\_Index}} \tag{1.13}$$

  donde expected_index representa el valor medio del índice cuando se elige un modelo hipergeométrico de aleatoriedad[1]; y max_Index es el máximo valor que toma el índice. Con esta corrección el índice pasa a ser un parámetro entre que es igual a 0 cuando se está bajo la hipótesis nula de aleatoriedad y 1 cuando las particiones coindicen. De esta manera, el *adjusted mutual information* queda definido como, Vinh et al. (2010),

$$AMI(\mathcal{C}, \tilde{\mathcal{C}}) = \frac{MI(\mathcal{C}, \tilde{\mathcal{C}}) - \mathbb{E}_{\text{perm}}\left[MI(\mathcal{C}, \tilde{\mathcal{C}})\right]}{\text{máx}\left\{H(\mathcal{C}), H(\tilde{\mathcal{C}})\right\} - \mathbb{E}_{\text{perm}}\left[MI(\mathcal{C}, \tilde{\mathcal{C}})\right]}$$

[1] Dadas las dos particiones, se toma esperanza del índice para todas los pares de particiones elegidos al azar pero sujeto a que el tamano de los clusters es el mismo que para las particiones originales $\mathcal{C}$ y $\tilde{\mathcal{C}}$.

- **Adjusteed Rand index.** Es un índice combinatorio que cuantifica cuantos son los pares de puntos que aparecen en el mismo cluster en ambas particiones. Sean:

$$A = \#\text{pares de puntos que aparecen en el mismo cluster en ambas particiones,}$$
$$D = \#\text{pares de puntos que aperecen en distintas clusters en ambas particiones.}$$

Luego se define el *Rand index* como:

$$RI(\mathcal{C}, \tilde{\mathcal{C}}) = \frac{A + D}{\binom{n}{2}}.$$

El *Rand index* es un índice acotado por 1 y que alcanza dicho valor únicamente cuando ambas particiones coinciden. Nuevamente, no incorpora correciones por aleatoriedad. A partir de la fórmula (1.13) se define el *adjusted Rand index* al igual que como hicimos con la *adjusted mutual information*, Hubert & Arabie (1985). A partir de la tabla de contingencias es posible calcular el *adjusted Rand index* a partir de

$$ARI(\mathcal{C}, \tilde{\mathcal{C}}) = \frac{\sum_{ij} \binom{n_{ij}}{2} - \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} / \binom{n}{2}}.$$

- **Accuracy.** Primero realizamos una asignación entre los clusters de ambas particiones $\mathcal{C}$ y $\tilde{\mathcal{C}}$ de acuerdo a la tabla de contingencias: consideramos el mayor de los n_{ij} y asociamos el cluster U_i con el V_j; luego buscamos el siguiente valor más grande de $n_{i'j'}$ con $i' \neq i$, $j \neq j'$ y asociamos el cluster $U_{i'}$ con el $V_{j'}$; y así sucesivamente. En el caso de que ambas particiones tengan la misma cantidad de elementos, esto devuelve una relación uno a uno entre particiones. En el caso donde ambas particiones tienen distintos elementos, quedan clusters sin asociarse. Luego, el *accuracy* se define como la fracción de puntos clasificados correctamente.

- F_1 **score.** Supongamos que $\mathcal{C}$ y $\tilde{\mathcal{C}}$ están formados por sólo dos particiones y las interpretamos como resultados positivos y negativos de un determinado síntoma y diagnóstico, respectivamente. Se definen:

$$\text{precission} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad , \quad \text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}.$$

Luego el F_1 *score* se define como la media armónica entre ambas cantidades, Powers (2011),

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}.$$

En el caso de tener varios clusters, el F_1*-score* se define de la siguiente manera. Para cada cluster consideramos su cluster asociado de la otra partición, tal como se hace con el *accuracy*; les asignamos una etiqueta 1 y al resto de los puntos los clasificamos como 0 y calculamos el F_1 score en ese caso; por último calculamos el promedio de todos resultados obtenidos.

Capítulo 2

Distancia de Fermat: propuesta, método y resultados

En este capítulo introducimos la *distancia de Fermat* junto con su estimador. La misma define una distancia sobre el soporte de una determinada distribución de probabilidad f y cuantifica qué tan parecidos son dos puntos a partir de la estructura del soporte y de la densidad f. Nos concentraremos en definir el estimador y exhibir sus propiedades, mostrando por qué es de gran utilidad para muchas tareas. Vamos a mostrar cómo se realiza su implementación algorítmica y evaluaremos su performance como input de un problema de clustering con datos sintéticos.

2.1. Distancia de Fermat

Sea $\mathcal{M} \subset \mathbb{R}^D$ una variedad de dimensión d, es decir, una superficie que localmente es equivalente a $\mathbb{R}^d$. Típicamente vamos a tener que $d \ll D$, aunque no es una hipótesis necesaria. Consideremos un conjunto de n puntos $\mathbb{X}_n \subset \mathcal{M}$ sampleados a partir de una determinada distribución con densidad $f : \mathcal{M} \mapsto \mathbb{R}_{\geq 0}$. Por otro lado, consideremos sobre $\mathbb{R}^D$ la distancia inducida por la norma euclídea $|\cdot|$. Luego, dado un parámetro $\alpha \geq 1$ y dos puntos $\mathbf{p}, \mathbf{q} \in \mathcal{M}$ definimos el *estimador de la distancia de Fermat* como

$$\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) = \min_{(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K) \in \mathbb{X}_n^K} \sum_{i=1}^{K-1} |\, \mathbf{x}_i - \mathbf{x}_{i+1} \,|^{\alpha} \tag{2.1}$$

donde la minimización se realiza sobre todos los $K \geq 2$ y todos los caminos de puntos contenidos en $\mathbb{X}_n$ con

$$\mathbf{x}_1 = \text{argmin}_{\mathbf{x} \in \mathbb{X}_n} |\mathbf{x} - \mathbf{p}| \quad , \quad \mathbf{x}_K = \text{argmin}_{\mathbf{x} \in \mathbb{X}_n} |\mathbf{x} - \mathbf{q}|. \tag{2.2}$$

Notemos que para $\alpha = 1$ el estimador de la distancia de Fermat coincide con la distancia euclídea, mientras que para $\alpha > 1$ vamos a ver que la distancia tiende a cuantificar qué tan cerca están $\mathbf{p}$ y $\mathbf{q}$ cuando se mide la longitud de la geodésica contenida en $\mathcal{M}$ pesada por una función de la densidad f.

Primero, observemos que $\mathcal{D}_{\mathbb{X}_n}(\cdot, \cdot)$ define una distancia sobre $\mathbb{X}_n$ y una pseudo-distancia sobre $\mathcal{M}$. Dados $\mathbf{p}, \mathbf{q} \in \mathbb{X}_n$, $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) = 0$ si y sólo si $\mathbf{p} = \mathbf{q}$ y $\mathcal{D}_{\mathbb{X}_n}(\cdot \cdot)$ es simétrico. Por otro lado, dados tres puntos $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathcal{M}$ es claro que a partir los caminos que realizan el mínimo en (2.1) que conectan a $\mathbf{p}$ con $\mathbf{r}$ y a $\mathbf{r}$ con $\mathbf{q}$ se puede construir un nuevo camino que conecte a $\mathbf{p}$ con $\mathbf{q}$, de manera tal que se cumple:

$$\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \leq \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{r}) + \mathcal{D}_{\mathbb{X}_n}(\mathbf{r}, \mathbf{q}) \quad \forall\, \mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathcal{M}.$$

Dado que puntos consecutivos $\mathbf{x}_i$ y $\mathbf{x}_{i+1}$ de un camino que se encuentren alejados contribuyen negativamente a la minimización en (2.1), se tiene que el camino de puntos $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)$ que realice el mínimo va a estar formado por puntos consecutivos que se encuentren a corta distancia. Debido a los efectos que los espacios de alta dimensión tienen sobre las distancias usuales (*curse of dimensionality*), es deseable trabajar sólo con la distancia entre puntos que realmente se encuentran cerca, dado que realmente reflejan la estructura del espacio, Aggarwal et al. (2001). Notemos que la definición de (2.1) y las propiedades que acabamos de enunciar no se ven modificadas si en vez de usar la distancia euclídea entre puntos usamos cualquier otra métrica sobre $\mathbb{R}^D$. Desde un punto de vista práctico, se puede usar cualquiera de estas métricas, sin embargo los resultados que siguen a continuación son demostrados únicamente para el caso euclídeo.

Consideremos las siguientes hipótesis sobre la variedad y la distribución de los datos sobre ella:

- (H1): $\mathcal{M} \subset \mathbb{R}^D$ es una variedad de dimensión d, con $d < D$, que se puede escribir como $\mathcal{M} = \varphi(C)$, siendo $\varphi : C \mapsto \mathbb{R}^D$ una trasformación isométrica y $C \subset \mathbb{R}^d$ un conjunto convexo, compacto y tal que $\bar{C^o} = C$,

- (H2): $f : \mathcal{M} \mapsto \mathbb{R}_{\geq 0}$ es una función de densidad continua con $f_{\min} = \min_{\mathbf{x}\in\mathcal{M}} f(\mathbf{x}) > 0$.

El siguiente teorema muestra la convergencia del estimador cuando $n \to \infty$ a un objeto macroscópico no trivial. Su demostración constituye el objetivo central del próximo capítulo.

Teorema 1. *Sea $\mathbb{X}_n$ una muestra i.i.d de tamaño n distribuida a partir de una densidad $f : \mathcal{M} \mapsto \mathbb{R}_{\geq 0}$ de manera tal que valen (H1), (H2). Luego, para $\alpha > 1$ y dados $\mathbf{p}, \mathbf{q} \in \mathcal{M}$ se tiene*

$$\lim_{n\to\infty} n^{\beta} \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) = \mu_{\alpha,d} \inf_{\Gamma\subset\mathcal{M}} \int_{\Gamma} \frac{1}{f^{\beta}} \quad \textit{casi seguramente,} \tag{2.3}$$

donde $\beta = (\alpha - 1)/d$; $\mu_{\alpha,d}$ es una constante que depende del parámetro α y de la dimensión d de $\mathcal{M}$; y la minimización se realiza sobre todas las curvas continuas y rectificables Γ contenidas en la variedad $\mathcal{M}$ y que conectan $\mathbf{p}$ con $\mathbf{q}$. Más aún, si existe una única curva $\hat{\Gamma} \subset \mathcal{M}$ que conecta $\mathbf{p}$ con $\mathbf{q}$ y tal que

$$\int_{\hat{\Gamma}} \frac{1}{f^{\beta}} = \inf_{\Gamma\subset\mathcal{M}} \int_{\Gamma} \frac{1}{f^{\beta}}, \tag{2.4}$$

entonces la sucesión de curvas Γ_n que realizan el camino óptimo convergen uniformemente a $\hat{\Gamma}$.

Siguiendo el Principio de Fermat: *Definimos la distancia de Fermat entre todo par de puntos $\mathbf{p}, \mathbf{q} \in \mathcal{M}$ como*

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) = \inf_{\Gamma\subset\mathcal{M}} \int_{\Gamma} \frac{1}{f^{\beta}}. \tag{2.5}$$

Es fácil observar que $\mathcal{D}(\cdot,\cdot)$ define una distancia sobre $\mathcal{M}$. De está manera, el Teorema 1 establece que $\mu_{\alpha,d}^{-1} n^{\beta} \mathcal{D}_{\mathbb{X}_n}(\cdot,\cdot)$ es un estimador consistente de $\mathcal{D}(\cdot,\cdot)$ cuando $\mathbb{X}_n$ es una muestra i.i.d de f. Por lo tanto, para $\alpha > 1$ el estimador definido en (2.1) toma en cuenta tanto el soporte $\mathcal{M}$ donde se encuentran los puntos como la densidad f.

Observemos el parecido que existe con el Principio de Fermat en óptica. El mismo establece que la trayectoria Γ seguida por un haz de luz para llegar de un punto a otro es un extremo del funcional llamado camino óptico, el cual está dado por

$$\Gamma \mapsto \int_{\Gamma} \mathrm{n}(\mathbf{x}) dl, \tag{2.6}$$

donde n *es el índice de refracción del medio, definido como el cociente entre la velocidad de la luz en el vacío y la velocidad de la luz en el medio. Por ejemplo, para el vacío* $\mathrm{n} = 1$ *y para el agua* $\mathrm{n} \approx 1{,}33$*. El camino óptico representa el tiempo que tarda la luz en recorrer una determinada trayectoria. De esta manera, existe una analogía entre la distancia de Fermat y el Principio de Fermat donde* $f^{-\beta}$ *juega el rol del índice de refracción. El camino* Γ *que minimice* (2.5) *va a tratar de ir por regiones de densidad alta haciendo que la contribución de* $f^{-\beta}$ *sea lo más chica posible.*

La idea detrás de la definición de la distancia de Fermat es la de medir la cercanía entre puntos, típicamente en algún problema de análisis de datos o Machine Learning, mediante una magnitud que contemple tanto el soporte $\mathcal{M}$ de los datos (medir lo longitud de las geodésicas en vez de las lineas rectas) como su magnitud (identificar zonas de alta densidad que pueden ser interpretadas como clusters de puntos). En la Figura 2.1 se ilustra está situación.

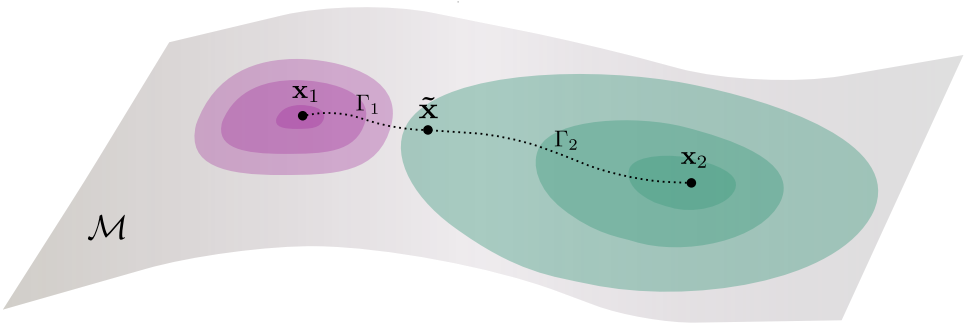


Figura 2.1: ¿Cómo funciona la distancia de Fermat?. *Supongamos que un conjunto de puntos es sampleado a partir de una densidad* f *con soporte en una superficie* $\mathcal{M}$ *de dimensión* d*, típicamente menor que la dimensión total del espacio, donde* f *es una densidad con dos modas* $\mathbf{x}_1$ *y* $\mathbf{x}_2$ *bien diferenciadas pero con dispersiones distintas. Consideremos también otro punto* $\tilde{\mathbf{x}} \in \mathcal{M}$ *como se muestra en la figura. Si medimos la distancia de* $\tilde{\mathbf{x}}$ *a* $\mathbf{x}_1$ *y* $\mathbf{x}_2$ *a partir de la distancia euclídea o de las longitud de la geodésicas* Γ_1 *y* Γ_2 *observaríamos que el punto* $\tilde{\mathbf{x}}$ *se encuentra más cerca de* $\mathbf{x}_1$ *que de* $\mathbf{x}_2$*. Sin embargo, dado que la distribución alrededor de* $\mathbf{x}_2$ *es más dispersa que la de* $\mathbf{x}_1$*, es deseable encontrar una distancia donde está situación esté contemplada y el punto* $\tilde{\mathbf{x}}$ *se encuentre más cerca de* $\mathbf{x}_2$ *que de* $\mathbf{x}_1$*. Es exactamente este el efecto que tiene la distancia de Fermat. La misma pesa la longitud de las geodésicas con una potencia inversa de la densidad, de manera que el peso total acumulado de* $\tilde{\mathbf{x}}$ *a* $\mathbf{x}_2$ *es menor que el peso acumulado de* $\tilde{\mathbf{x}}$ *a* $\mathbf{x}_1$*.*

2.2. Implementación

Todos los códigos y scripts fueron desarrollados en Python 3.6.5[1]. La implementación del estimador de la distancia de Fermat se efectúa por medio del algoritmo de Dijkstra de búsqueda de camino mínimo en grafos.

Si bien la minimización involucrada en la definición del estimador de la distancia de Fermat se realiza sobre todos los posibles caminos de puntos contenidos en $\mathbb{X}_n$, es posible demostrar que es posible restringir la búsqueda a pares consecutivos de puntos que sean k-vecinos más cercanos sin modificar significativamente el estimador. Dado un punto $\mathbf{x}$, definimos como $\mathcal{N}_k(\mathbf{x})$ al conjunto de los k vecinos más cercanos a $\mathbf{x}$ dentro del conjunto de puntos $\mathbb{X}_n$, es decir, los k puntos que se encuentran más cerca de $\mathbf{x}$ a partir de la distancia euclídea. Luego, dados $\alpha \geq 1$ y $k \in \mathbb{N}$ definimos el *estimador de la distancia de Fermat restringido* $\hat{\mathcal{D}}^k_{\mathbb{X}_N}(\cdot,\cdot)$ como

$$\hat{\mathcal{D}}^k_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \min_{\substack{(\mathbf{y}_1,\ldots,\mathbf{y}_K) \in \mathbb{X}_n^K, \\ \mathbf{y}_1 = \mathbf{p}, \mathbf{y}_K = \mathbf{q} \\ \mathbf{y}_{i+1} \in \mathcal{N}_k(\mathbf{x}_i)}} \sum_{i=1}^{K-1} |\mathbf{y}_{i+1} - \mathbf{y}_i|^\alpha. \tag{2.7}$$

[1] Todos los códigos son abiertos y se encuentran disponibles en *github.com/facusapienza21/d-distance*

La siguiente proposición muestra como se compara el estimador de la distancia de Fermat $\mathcal{D}_{\mathbb{X}_N}(\cdot,\cdot)$ con el estimador restringido. Su demostración será objeto de estudio del siguiente capítulo.

Proposición 1. *Dado $\varepsilon > 0$, existe $k_0 = \mathcal{O}(\log(n/\varepsilon))$ tal que*

$$\hat{\mathcal{D}}^{k_0}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \quad \textit{con probabilidad al menos } 1-\varepsilon. \tag{2.8}$$

Más precisamente, el $\mathbb{X}_n$-camino minimizante $\mathbf{y}^*_1,\ldots,\mathbf{y}^*_{K_n}$ satisface $\mathbf{y}^*_{i+1} \in \mathcal{N}_{k_0}(\mathbf{y}^*_i)$ para todo $i = 1,\ldots,K_n-1$ con probabilidad al menos $1-\varepsilon$.

Dado un grafo $G = (V,E)$, donde V es el conjunto de vértices y E el conjunto de aristas, el tiempo de ejecución del algoritmo de Dijkstra es $\mathcal{O}(|V|^2)$, por lo tanto el cálculo del estimador de la distancia de Fermat entre todos los pares de puntos es $\mathcal{O}(N^3)$, Cormen (2009). Sin embargo, utilizando colas de prioridad la complejidad se reduce a $\mathcal{O}(|E| + |V|\log|V|)$. Si se considera el grafo de k-vecinos más cercanos se tiene $|V| = \mathcal{O}(kN)$ y por lo tanto el estimador de la distancia de Fermat entre todo par de puntos puede calcularse en un tiempo de ejecución $\mathcal{O}(N^2(\log N)^2)$. De esta manera se logra reducir el tiempo de ejecución de $\mathcal{O}(N^3)$ a $\mathcal{O}(N^2(\log N)^2)$.

Restringir la búsqueda a k vecinos más cercanos simplemente representa una mejora en el tiempo de corrida del algoritmo y sin modificar las propiedades macroscópicas del estimador. En la Tabla 2.2 se muestran las distintas distancias presentadas.

	Distribución homogénea	Distribución no homogénea
	$\alpha = 1$	$\alpha > 1$
sin k-NN	distancia euclídea	distancia de Fermat $\mathcal{D}_{\mathbb{X}_n}(\cdot,\cdot)$
con k-NN	distancia geodésica (Isomap)	distancia de Fermat restringida $\mathcal{D}^k_{\mathbb{X}_n}(\cdot,\cdot)$

Figura 2.2: Esquema de distancias. *Cuando consideramos el estimador de la distancia de Fermat para $\alpha = 1$ recuperamos la distancia usual del espacio. Para datos no homogéneos, nosotros introducimos un nuevo estimador que contempla la distribución con la cual los datos están distribuidos. A su vez, la Proposición 1 nos permite restringir la búsqueda a caminos formados por pares de puntos consecutivos que sean k-vecinos más cercanos entre si sin modificar las propiedades del estimador original. Cuando $\alpha = 1$ y restringimos a los primeros vecinos recuperamos el algoritmo de Isomap.*

A continuación, se resumen los argumentos de entrada y la salida del algoritmo que calcula el estimador de la distancia de Fermat entre todo par de puntos en $\mathbb{X}_n$.

- **Parámetros**

  distance_matrix: Matriz cuadrada de $n \times n$ con entradas no negativas y diagonal igual a cero.

  Matriz de distancias original entre los datos (por ejemplo, la distancia euclidea).

  alpha: Número real mayor o igual que 1.

  Parámetro α de la distancia de Fermat.

  dimension: Número entero mayor o igual a 1.

*Dimensión d de la manifold donde viven los datos. Sólo es necesatio cuando **normalization**=True. Simplemente se introduce este parámetro cuando se conoce la dimensión de la superficie donde viven los puntos y se desea incluir la constante normalizadora n^β en el estimador de la distancia de Fermat. Para cualquier aplicación, este parámetro no tiene importancia.*

k_nn: Número natural.

Número de vecinos más cercanos a partir del cual se construye el grafo sobre el cual se va a calcular el estimador de la distancia de Fermat.

indices_to_do: *all* en caso de calcular todos o una lista de los puntos a calcular en caso contrario.

Puntos para los cuales se va a calcular el estimador de la distancia de Fermat. Por default es all y se calcula para todo par de puntos.

normalization: bool.

En caso de ser True, incorpora la constante normalizadora n^β.

- **Return**

out_dist: Matriz cuadrada de $n \times n$ con entradas no negativas y diagonal igual a cero.

Matriz con el estimador de la distancia de Fermat calculado entre todo par de puntos del dataset.

path: Matriz cuadrada de $n \times n$.

Matriz a partir del cual se puede reconstruir el camino mínimo entre cualquier par de puntos.

2.3. Experimentos

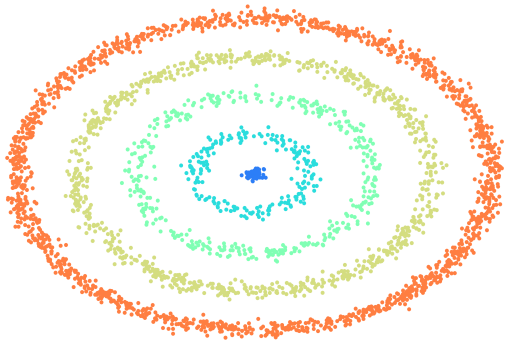
2.3.1. Anillos

Una manera ilustrativa de entender cómo funciona la distancia de Fermat es a partir de estudiar lo que sucede cuando un conjunto de puntos se encuentra localizado en anillos concéntricos de radios distintos (Figura 2.3(a)). Los mismos se obtienen luego de samplear radialmente 100, 200, 400, 900 y 1600 puntos a partir de distribuciones normales con medias 0, 1, 2, 3 y 4 y desvío estándar 0,1, respectivamente, y con distribución angular uniforme.

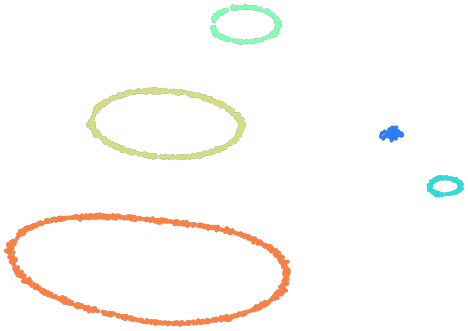
Notemos que la distancia de Fermat (2.5) entre puntos de distintas componentes conexas estrictamente separadas es infinito, mientras que la de puntos localizados en la misma componente es finita. De la misma manera, puntos separados por regiones con densidad f muy chica van a estar a distancia mucho mayor que puntos localizados en la misma región de densidad alta.

Por lo tanto, luego de calcular el estimador de la distancia de Fermat entre los puntos para $\alpha > 1$ y hacer una representación en dos dimensiones de los mismos puntos pero con la nueva distancia por medio del algoritmo de t-SNE, observamos como cada una de las componentes conexas queda completamente separada de las otras, si bien cada una de ellas respeta su estructura interna (Figura 2.3(b)). El hecho de que las componentes queden tan bien separadas es de gran utilidad si se desea efectuar un algoritmo de clustering.

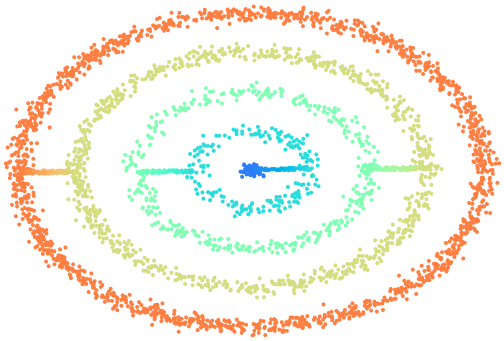
En el mismo contexto, una pregunta muy interesante que nos podemos hacer es qué sucede si realizamos el mismo procedimiento pero haciendo que los distintos anillos estén conectados entre sí por pequeños puentes (Figura 2.3(c)). La Figura 2.3(d) muestra como se respeta la estructura de los anillos a la vez que los puentes funcionan de conectores entre los mismos.



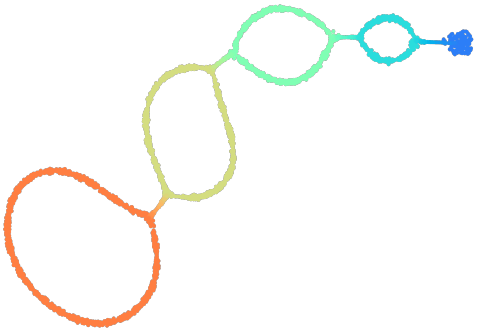
(a) Anillos dataset



(b) Representación de los Anillos utilizando t-SNE con el estimador de la distancia de Fermat para $\alpha = 2$.



(c) Anillos con puentes dataset



(d) Representación de los Anillos con Puentes utilizando t-SNE con el estimador de la distancia de Fermat para $\alpha = 1{,}2$.

Figura 2.3: Anillos.

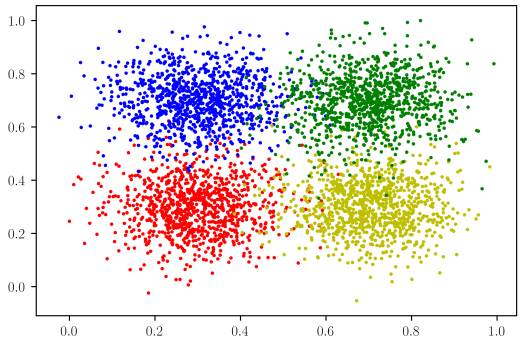
2.3.2. Normales en Swiss Roll

Uno de los ejemplos más utilizados para testear algoritmos de *manifold learning* es el famoso rollo suizo (*Swiss Roll* a partir de ahora), el cual fue introducido cuando discutimos acerca del algoritmo de *Isomap*. El objetivo es definir una distancia entre puntos dentro del Swiss Roll pero midiendo las geodésicas sobre la superficie. El algoritmo de *Isomap* está diseñado para medir geodésicas dentro de la superficie donde están soportados los datos, pero no establece nada acerca de la distribución de los datos dentro de ella. Para ver las ventajas que el estimador de la distancia de Fermat presenta respecto de este otro método consideremos una variación del clásico *Swiss Roll*.

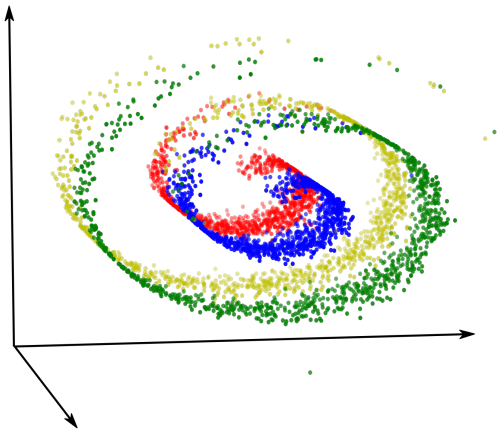
En dos dimensiones, consideremos cuatro distribuciones normales con misma matriz de covarianza pero distintas medias. Las medias de las normales están dadas por $\mu_1 = (0{,}3, 0{,}3)$, $\mu_2 = (0{,}3, 0{,}7)$, $\mu_3 = (0{,}7, 0{,}3)$ y $\mu_4 = (0{,}7, 0{,}7)$ y matriz de covarianza proporcional a la identidad y con desvió estándar igual a 0,2. Para cada normal se samplean un total de 1000 puntos, de manera tal que el conjunto de puntos totales tenga un tamaño igual a $n = 4000$ (Figura 2.4(a)). Luego, consideremos una transformación $h : \mathbb{R}^2 \mapsto \mathbb{R}^3$ cuya imagen sea el *Swiss Roll*, tal como se muestra en la Figura 2.4(b). En nuestro caso, consideramos la transformación dada por:

$$h(t, s) = (t\cos(\omega t), t\sin(\omega t), As)$$

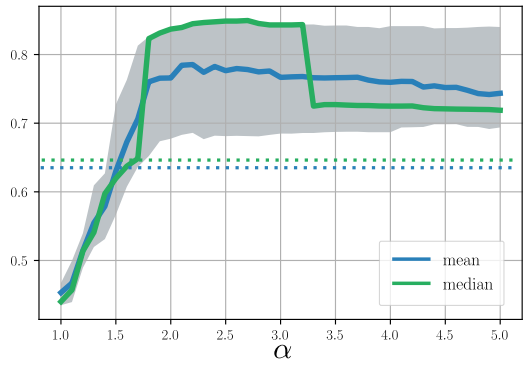
donde $A = 3$ y $\omega = 15$. La elección de los parámetros A y ω es tal que la superficie donde se mapean los datos tenga largo y ancho comparables. El hecho de que la transformación h no sea isométrica implica que, medidas sobre el *Swiss Roll*, las normales se mapean a distribuciones que no necesariamente son normales y que a su vez tienen matrices de covarianza distintas.



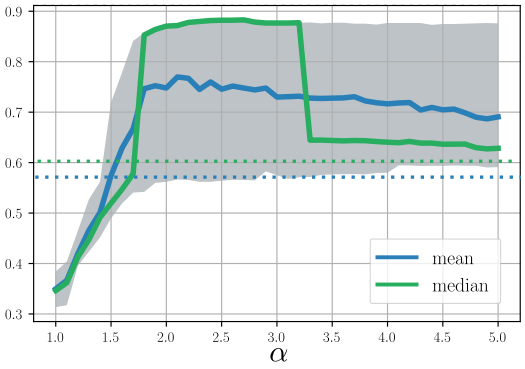
(a) Distribución de los datos en 2D



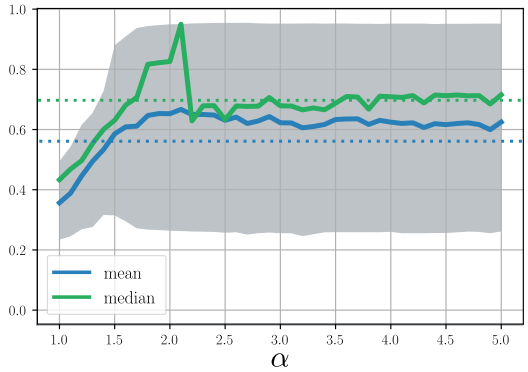
(b) Distribución de los datos en 3D



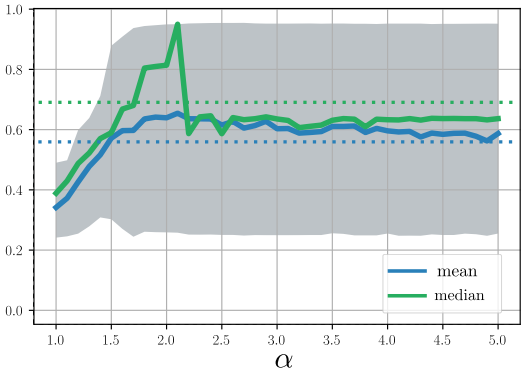
(c) Adjusted mutual information



(d) Adjusted Rand index



(e) Accuracy



(f) F1 score

Figura 2.4: Clustering en el Swiss Roll. *Consideremos un conjunto de puntos en dos dimensiones con clusters bien definidos (2.4(a)) mapeados en tres dimensiones tal como se muestra en 2.4(b). Luego, se calcula el estimador de la distancia de Fermat entre los puntos y a partir de dicha distancia efectuamos el algoritmo de K-medoids para encontrar clusters de puntos, eligiendo $K = 4$. Para* 1000 *corridas K-medoids, se calcula la performance media (azul), la mediana (verde) y la franja intercuantil (gris) para el adjusted mutual information (2.4(c)), adjusted Rand index (2.4(d)), accuracy (2.4(e)) y F_1 score (2.4(f)). Observamos que la performance del algoritmo mejora dentro de un rango de valores de α respecto de la distancia euclídea del espacio ($\alpha = 1$) y de la performance media (linea punteada azul) y mediana (linea punteada verde) que se obtiene a partir de Isomap y C-Isomap, barriendo sobre todos los valores posibles de k_{Isomap}. De esta manera, observamos como el estimador de la distancia de Fermat refleja mucho mejor la estructura intrínseca de los datos, en particular cuando se desea realizar una tarea de clustering.*

Para evaluar la performance del estimador de la distancia de Fermat, se calcula la matriz de distancias para distintos valores de α y se efectúan un total de $n_{iterations} = 1000$ corridas del algoritmo de K-medoids con distintas configuraciones iniciales elegidas al azar. Para cada una de las corridas, se calculan distintos índices entre la clasificación que resulta del algoritmo y la verdadera clasificación de los datos. Dichos índices incluyen: *adjusted mutual information* (2.4(c)), *adjusted Rand index* (2.4(d)), *accuracy* (2.4(e)), F_1 *score* (2.4(f)). Todos estos índices fueron definidos en el primer capítulo. Para cada indicador se muestra la media (azul), la mediana (verde) y la distancia intercuartil (sombra gris) observadas dentro de las $n_{iterations}$ iteraciones del algoritmo. A su vez, en linea punteada se muestra la performance observada cuando se utiliza la distancia devuelta por el algoritmo de *Isomap* o *C-Isomap* (en todos los caso se muestra la mejor de ambas y se selecciona el parámetro k_{Isomap} de manera de maximizar la performance). Para todos los indicadores se observa que existe un intervalo de valores de α para los cuales la performance del clustering mejora. Para el *adjusted mutual information* y el *adjusted Rand index* se observan performances superiores para $1{,}8 \leq \alpha \leq 3{,}2$ mientras que para el *accuracy* y F_1 *score* dicho intervalo se reduce a $1{,}7 \leq \alpha \leq 2{,}1$. En todos los casos no sólo se observar mejores resultados respecto de la distancia euclídea (caso $\alpha = 1$), sino que también se consiguen mejores resultados respecto de *Isomap* y *C-Isomap*.

Capítulo 3

Consistencia del estimador

El azar siempre ayuda.

— Sabiduría china

El objetivo de esta sección es dar una demostración del Teorema 1 enunciado en el Capítulo 2, el cual establece la convergencia del estimador de la distancia de Fermat. A su vez, al final del capítulo daremos una demostración de la Proposición 1, también enunciada en el capítulo anterior, que establece bajo qué condiciones el estimador de la distancia de Fermat y el estimador de la distancia de Fermat restringido son equivalentes.

El Teorema 1 está enunciado para una muestra independiente e idénticamente distribuida (i.i.d) con densidad f sobre una variedad compacta y conexa. Vamos a comenzar probando la convergencia del estimador para una muestra proveniente de un proceso puntual de Poisson de intensidad $nf(\mathbf{x})$ sobre un conjunto compacto conexo $C \subset \mathbb{R}^d$ con $\bar{C^o} = C$. Dicho problema es interesante de por sí y se encuadra dentro de la teoría de percolación euclídea de primera pasada. Luego, extenderemos los resultados a los casos donde el tamaño de la muestra está fijo (ensamble canónico) y al caso de una variedad $\mathcal{M}$ contenida en un espacio de dimensión mayor.

3.1. Preliminares

Dado un conjunto boreliano $A \subset \mathbb{R}^d$, vamos a notar por $|A|$ a su medida de Lebesgue y por $\#A$ al número de puntos en A. A su vez, vamos a notar por $|\cdot|$ a la norma euclídea sobre $\mathbb{R}^d$.

Dado un conjunto medible Borel $C \subset \mathbb{R}^d$, una configuración aleatoria de puntos $\mathbb{X} \subset C$ se dice un proceso puntual de Poisson con intensidad $\lambda : C \mapsto \mathbb{R}_{\geq 0}$ si para todo par de conjuntos medibles Borel disjuntos $A, B \subset C$ se tiene (Moller & Waagepetersen (2003); Kallenberg (2002))

$$\mathbb{P}\Big(\#(\mathbb{X} \cap A) = k, \#(\mathbb{X} \cap B) = j\Big) = \frac{e^{-(S(A)+S(B))} S(A)^k S(B)^j}{k!\, j!}, \tag{3.1}$$

donde $S(\cdot)$ es la función definida sobre los conjuntos borelianos contenidos en C dada por

$$S(A) = \int_A \lambda(\mathbf{x})dx.$$

Un proceso puntual de Poisson se dice homogéneo si su intensidad λ es constante. Una propiedad importante de los procesos de Poisson homogéneos es que, condicionado al número de partículas sobre un conjunto compacto, la distribución coincide con una muestra i.i.d uniforme sobre C. Notemos que (3.1) se traduce en que la cantidad de puntos contenidos en cualquier conjunto A sigue una distribución de Poisson de parámetro $S(A)$ independiente de la cantidad de puntos que haya en cualquier otro conjunto medible B con $A \cap B = \emptyset$.

Sea $\mathbb{X}$ un conjunto localmente finito de puntos dados por un proceso puntual. Nos referiremos a los puntos en $\mathbb{X}$ como partículas, para diferenciarlos de los demás puntos en $\mathbb{R}^d$. Para cualquier punto $\mathbf{p} \in \mathbb{R}^d$ se define el centro de su celda de Voronoi como

$$\mathbf{y}(\mathbf{p}) = \operatorname*{argmín}_{\mathbf{y}\in\mathbb{X}} |\mathbf{p} - \mathbf{y}|.$$

Para cada par de puntos $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$ definimos a $(\mathbf{y}_1, \ldots, \mathbf{y}_K)$ con $\mathbf{y}_1 = \mathbf{y}(\mathbf{p}), \mathbf{y}_K = \mathbf{y}(\mathbf{q})$ como un camino de $\mathbf{p}$ a $\mathbf{q}$ (o $\mathbb{X}$-camino de ser necesario). Dado un parámetro $\alpha > 1$, definimos el *estimador distancia de Fermat respecto de* $\mathbb{X}$ como

$$\mathcal{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q}) = \text{ínf} \left\{ \sum_{j=1}^{K-1} |\mathbf{y}_{i+1} - \mathbf{y}_i|^\alpha \colon K \geq 2, \text{ y } (\mathbf{y}_1, \ldots, \mathbf{y}_K) \text{ es un } \mathbb{X}\text{-camino de } \mathbf{p} \text{ a } \mathbf{q} \right\}. \tag{3.2}$$

Observemos que en tal caso $\{\mathcal{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q})\}$ es una familia de variables aleatorias indexadas por $(\mathbf{p}, \mathbf{q}) \in \mathbb{R}^{2d}$. Por otro lado, notemos que si tomáramos $\alpha \leq 1$ el estimador de la distancia de Fermat sería trivialmente $\mathcal{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q}) = |\,\mathbf{p} - \mathbf{q}\,|^\alpha$.

Si el conjunto $\mathbb{X}$ es finito, entonces la cantidad de $\mathbb{X}$-caminos sin partículas repetidas es finito y por lo tanto existe un camino que realiza el ínfimo. Por otro lado, de la continuidad de $\mathcal{D}_{\mathbb{X}}(\cdot, \cdot)$ respecto de $\mathbb{X}$ se sigue que el camino que realiza el mínimo es único con probabilidad uno.

3.2. Caso Poisson homogéneo

El caso donde $\mathbb{X}$ proviene de un proceso puntual de Poisson homogeneo con $\lambda = 1$ sobre $\mathbb{R}^d$ es introducido en Howard & Newman (1997) dentro del contexto de *Euclidean First Passage Percolation Theory*. Recomendamos al lector interesado consultar Howard & Newman (2001), donde se hace una revisión más en profundidad del problema incluyendo resultados de fluctuaciones.

Proposición 2 (Howard & Newman (1997), Lema 3 y Lema 4; Howard & Newman (2001), Teorema 2.2)**.** *Sea* $\mathbb{X}$ *un proceso puntual de Poisson en* $\mathbb{R}^d$ *con intensidad* $\lambda = 1$*. Entonces existe* $0 < \mu < \infty$ *tal que*

$$\lim_{|\mathbf{q}|\to\infty} \frac{\mathcal{D}_{\mathbb{X}}(\mathbf{0}, \mathbf{q})}{|\mathbf{q}|} = \mu \quad , \quad \textit{casi seguramente.} \tag{3.3}$$

Más aun, dado $\kappa_1 = \text{mín}\{1, d/\alpha\}$ *y* $\kappa_2 = 1/(4\alpha + 3)$*, para todo* $\varepsilon \in (0, \kappa_2)$ *existen constantes* c_0 *y* c_1 *que dependen de* ε *tales que*

$$\mathbb{P}\Big(|\mathcal{D}_{\mathbb{X}}(\mathbf{0}, l\mathbf{e}_1) - \mu l| \geq \eta \Big) \leq c_1 \exp\left(-c_0 \left(\eta/\sqrt{l}\right)^{\kappa_1}\right) \tag{3.4}$$

vale para todo $l > 0$ *y* η *que satisfaga* $l^{\frac{1}{2}+\varepsilon} \leq \eta \leq l^{\frac{1}{2}+\kappa_2-\varepsilon}$*.*

Sea $C \subset \mathbb{R}^d$ un conjunto convexo y compacto y sea $\mathbb{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ un proceso puntual de Poisson homogéneo en C con intensidad $\lambda_n = \mathcal{O}(n^\gamma)$ con $\gamma > 0$, de manera tal que $\lambda_n \to \infty$ cuando $n \to \infty$. Es decir, la cantidad de puntos N del conjunto $\mathbb{X}$ sigue una distribución de Poisson de parámetro $\lambda_n|C|$ y, condicionado a $N = k$, resulta que los puntos en $\mathbb{X}$ son i.i.d con distribución uniforme sobre C de tamaño k. Sea

$$\beta = \frac{\alpha - 1}{d}. \tag{3.5}$$

Mediante un reescalamiento adecuado, podemos probar la convergencia del estimador de la distancia de Fermat cuando $\mathbf{p}, \mathbf{q}$ están fijos en el espacio pero la cantidad de partículas en el conjunto compacto C tiende a infinito.

Proposición 3. *Dados $\mathbf{p}, \mathbf{q}$ en el interior de C se tiene*

$$\lim_{n\to\infty} \lambda_n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) = \mu|\mathbf{p} - \mathbf{q}|, \qquad \textit{casi seguramente.} \tag{3.6}$$

Más aun, dado $\delta > 0$ existen constantes positivas c_1, c_2, c_3, c_4, donde c_2 depende de δ, tales que si se cumple $|\,\mathbf{p} - \mathbf{q}\,| > \delta$ entonces

$$\mathbb{P}\Big(\big|\lambda_n^\beta \mathcal{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q}) - \mu|\,\mathbf{p} - \mathbf{q}\,|\big| \geq c_4 \lambda_n^{-1/3d} \Big) \leq c_1 \exp\left(-c_2 \lambda_n^{c_3}\right). \tag{3.7}$$

para todo n con $\lambda_n \geq 1$.

Demostración. Mediante traslaciones y rotaciones del conjunto C, podemos asumir sin perdida de generalidad que $\mathbf{p} = \mathbf{0}$ y $\mathbf{q} = \mathbf{e}_1 = (1, 0, \ldots, 0)$. En tal caso, es fácil ver que la intensidad λ_n y $\mathcal{D}_{\mathbb{X}_n}(\cdot, \cdot)$ transforman como

$$\lambda_n \mapsto \lambda_n |\,\mathbf{p} - \mathbf{q}\,|^d \quad , \quad \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \mapsto |\,\mathbf{p} - \mathbf{q}\,|^\alpha \mathcal{D}_{\hat{\mathbb{X}}_n}(\mathbf{0}, \mathbf{e}_1)$$

donde $\hat{\mathbb{X}}_n$ es un proceso puntual de Poisson de intensidad $\hat{\lambda}_n = \lambda_n |\,\mathbf{p} - \mathbf{q}\,|^d$ sobre el conjunto $\hat{C}$ que resulta de la transformación lineal que realiza el mapeo $\mathbf{p} \mapsto \mathbf{0}$, $\mathbf{q} \mapsto \mathbf{e}_1$. Por simplicidad de notación vamos a poner simplemente $\hat{\mathbb{X}}_n = \mathbb{X}_n$, $\hat{\lambda}_n = \lambda_n$ y $\hat{C} = C$. Luego, es fácil ver que los términos $|\,\mathbf{p} - \mathbf{q}\,|$ se cancelan y queda que (3.6) es equivalente a

$$\lim_{n\to\infty} \lambda_n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{0}, \mathbf{e}_1) = \mu. \tag{3.8}$$

En tal caso, la distribución de $\mathbb{X}_n$ coincide con la distribución de $\lambda_n^{-1/d}\mathbb{X} \cap C$, pues ambos son procesos puntuales de Poisson con misma intensidad. Reescalando nuevamente por un factor $\lambda_n^{1/d}$ tenemos que (3.8) es igual a

$$\lim_{n\to\infty} \frac{1}{\lambda_n^{1/d}} \mathcal{D}_{\mathbb{X}\cap\lambda_n^{1/d}C}\left(\mathbf{0}, \lambda_n^{1/d}\mathbf{e}_1\right) = \mu. \tag{3.9}$$

La única diferencia entre (3.9) y (3.3) es que en (3.3) la distancia es minimizada entre los $\mathbb{X}-$caminos mientras que en (3.9) la distancia es minimizada entre los $\left(\mathbb{X} \cap \lambda_n^{1/d} C\right)$-caminos. Para dos puntos cualquiera $\tilde{\mathbf{p}}$ y $\tilde{\mathbf{q}}$ y $a > 0$, consideramos la $a-$dilatación del segmento que une $\tilde{\mathbf{p}}$ con $\tilde{\mathbf{q}}$ definida como

$$[\![\tilde{\mathbf{p}}, \tilde{\mathbf{q}}]\!]_a := \left\{ \mathbf{x} \colon |\mathbf{x} - \mathbf{y}| < a \text{ para algún } \mathbf{y} \text{ en el segmento que une } \tilde{\mathbf{p}} \text{ con } \tilde{\mathbf{q}} \right\}. \tag{3.10}$$

Dado que $\mathbf{0}$ y $\mathbf{e}_1$ están en el interior del conjunto conexo C, se tiene que existe $a > 0$ tal que $[\![\mathbf{0}, \lambda_n^{1/d}\mathbf{e}_1]\!]_{a\lambda_n} \subset \lambda_n^{1/d} C$. Luego, vamos a probar que para cualquier valor de $a > 0$ se tiene

$$\lim_{n\to\infty} \frac{\mathcal{D}_{\mathbb{X}}\left(\mathbf{0}, \lambda_n^{1/d}\mathbf{e}_1\right)}{\lambda_n^{1/d}} = \lim_{n\to\infty} \frac{\mathcal{D}_{\mathbb{X}\cap[\![\mathbf{0}, \lambda_n^{1/d}\mathbf{e}_1]\!]_{a\lambda_n^{1/d}}}\left(\mathbf{0}, \lambda_n^{1/d}\mathbf{e}_1\right)}{\lambda_n^{1/d}}. \tag{3.11}$$

Sea Γ_n el $\mathbb{X}$-camino de partículas que realiza $\mathcal{D}_{\mathbb{X}}(\mathbf{0}, \lambda_n^{1/d}\mathbf{e}_1)$ y notemos por d_n^{max} a la distancia entre la geodésica Γ_n y el segmento que une $\mathbf{0}$ con $\lambda_n^{1/d}\mathbf{e}_1$. El Corolario 2.5 de Howard & Newman (2001) establece que para todo $\varepsilon > 0$ existe N_ε aleatorio tal que existen a lo sumo N_ε partículas para las cuales $d_n^{max} \geq (\lambda_n^{1/d})^{3/4+\varepsilon}$ y de manera tal que $N_\varepsilon < \infty$ en casi todo punto. En particular, va a existir $n_1 < \infty$ aleatorio para el cual se tiene $d_n^{max} < (\lambda_n^{1/d})^{3/4+\varepsilon}$ para todo $n > n_1$. Por otro lado, tomemos $\varepsilon < 1/4$ y n_2 de manera tal que $(\lambda_n^{1/d})^{\varepsilon - 1/4} < a$ para todo $n > n_2$. Eligiendo $n_0 = \max\{n_1, n_2\}$ tenemos que $d_n^{max} < a\lambda_n^{1/d}$ para todo $n > n_0$, lo cual inmediatamente implica (3.11).

Sabiendo que podemos restringir la búsqueda a los $\mathbb{X}_n \cap \lambda_n^{1/d} C$-caminos, eligiendo $l = |\mathbf{p} - \mathbf{q}|\lambda_n^{1/d}$ en (3.4) y usando la isotropía del proceso puntual de Poisson tenemos que

$$\mathbb{P}\bigg(\left|\mathcal{D}_{\mathbb{X}}(\mathbf{0}, l\mathbf{e}_1) - \mu l\right| \geq \eta \bigg) = \mathbb{P}\bigg(\left|\lambda_n^{\beta}\mathcal{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q}) - \mu |\,\mathbf{p} - \mathbf{q}\,|\right| \geq \eta \lambda_n^{-1/d} \bigg).$$

Luego, elegimos $\varepsilon = \kappa_2/2$ y $\eta = l^{\frac{1+\kappa_2}{2}} = |\,\mathbf{p} - \mathbf{q}\,|^{\frac{1+\kappa_2}{2}} \lambda_n^{\frac{1+\kappa_2}{2D}} \leq |\,\mathbf{p} - \mathbf{q}\,|^{\frac{1+\kappa_2}{2}} \lambda_n^{\frac{2}{3d}}$ para $\lambda_n \geq 1$, de manera tal que se desprende

$$\mathbb{P}\bigg(\left|\lambda_n^{\beta}\mathcal{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q}) - \mu |\,\mathbf{p} - \mathbf{q}\,|\right| \geq |\,\mathbf{p} - \mathbf{q}\,|^{\frac{1+\kappa_2}{2}} \lambda_n^{-1/3d} \bigg) \leq c_1 \exp\left(-c_0 |\,\mathbf{p} - \mathbf{q}\,|^{\frac{\kappa_1\kappa_2}{2}} \lambda_n^{\frac{\kappa_1\kappa_2}{2d}} \right). \tag{3.12}$$

Dado que la serie definida por el último término en (3.12) es sumable[1], a partir del lema de Borel-Cantelli concluimos la convergencia en (3.6) se da casi seguramente. Por otro lado, dado $\delta > 0$ tal que $|\,\mathbf{p} - \mathbf{q}\,| > \delta$, eligiendo $c_2 = c_2(\delta)$ como

$$c_2 = c_0 \delta^{\frac{\kappa_1\kappa_2}{2}} < c_0 |\,\mathbf{p} - \mathbf{q}\,|^{\frac{\kappa_1\kappa_2}{2}}$$

tenemos que se cumple (3.7), donde $c_3 = \kappa_1\kappa_2/(2d)$ y $c_4 = \operatorname{diam}(C)^{\frac{1+\kappa_2}{2}}$. □

3.3. Caso Poisson no homogéneo

Ahora sea $\mathbb{X}_n$ un proceso puntual de Poisson sobre C con intensidad $\lambda_n(\mathbf{x}) = nf(\mathbf{x})$, donde $f : C \mapsto \mathbb{R}_{\geq 0}$ es una función continua con

$$f_{\min} = \min_{\mathbf{x} \in C} f(\mathbf{x}) > 0 \quad , \quad f_{\max} = \max_{\mathbf{x} \in C} f(\mathbf{x}) < \infty. \tag{3.13}$$

Para demostrar la convergencia del estimador de la distancia de Fermat en el caso no homogéneo son necesarios algunos lemas previos. El siguiente resultado sobre acoplamientos de procesos de Poisson es de suma importancia para las demostraciones de las próximas secciones.

Lema 1 (Superposition and thinning, Moller & Waagepetersen (2003))**.** *Sea $\mathbb{X}$ un proceso puntual de Poisson con intensidad λ sobre C. Dados λ_- y λ_+ tales que $\lambda_- \leq \lambda \leq \lambda_+$ para todo $\mathbf{x} \in C$, es posible contruir dos procesos puntuales de Poisson $\mathbb{X}_-$ y $\mathbb{X}_+$ sobre C con intensidades λ_-, λ_+, respectivamente, tales que con probabilidad 1 se tiene $\mathbb{X}_- \subseteq \mathbb{X} \subseteq \mathbb{X}_+$.*

[1] En particular, la serie $(r^{n^x})_{n \in \mathbb{N}}$ es sumable para todo $x > 0$ y $0 \leq r < 1$. Consideremos $m \in \mathbb{N}$ tal que $1/m \leq x$. Luego

$$\sum_{n=1}^{\infty} r^{n^x} \leq \sum_{n=1}^{\infty} r^{n^{1/m}} = \sum_{n_1=1}^{\infty} \sum_{n_2=n_1^m}^{(n_1+1)^m - 1} r^{n_2^{1/m}} < \sum_{n_1=1}^{\infty} (n_1+1)^m r^{n_1} < \infty,$$

donde la convergencia de la última serie se desprende fácilmente del criterio de Cauchy.

3.3.1. Cotas para el proceso no homogéneo

Comenzamos encontrando cotas para los límites superiores e inferiores del estimador de la distancia de Fermat. Para ello vamos a basarnos en los resultados anteriormente demostrados para el caso donde las partículas provienen de un proceso de Poisson homogéneo sobre C.

Lema 2. *Sean $\mathbf{p}, \mathbf{q}$ puntos interiores de C y $\delta > 0$ con $|\,\mathbf{p}-\mathbf{q}\,| > \delta$. Sea $\mathbb{X}_n$ un proceso puntual de Poisson con intensidad $nf(\mathbf{x})$ sobre C. Luego, para todo $\varepsilon > 0$ se tiene que existen $n_0 = n_0(\varepsilon)$ determinísitico y constantes positivas c_1, c_2, c_3, con $c_2 = c_2(\delta)$, tales que*

$$\mathbb{P}\left(n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \leq \mu f_{max}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| - \varepsilon\right) \leq c_1 \exp\left(-c_2(f_{min}n)^{c_3}\right) \tag{3.14}$$

$$\mathbb{P}\left(n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \geq \mu f_{min}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| + \varepsilon\right) \leq c_1 \exp\left(-c_2(f_{min}n)^{c_3}\right) \tag{3.15}$$

para todo $n > n_0$.

Demostración. Observemos que dadas dos configuraciones localmente finitas $\mathbb{X}$ y $\hat{\mathbb{X}}$, si $\mathbb{X} \subseteq \hat{\mathbb{X}}$, entonces

$$\mathcal{D}_{\hat{\mathbb{X}}}(\mathbf{p},\mathbf{q}) \leq \mathcal{D}_{\mathbb{X}}(\mathbf{p},\mathbf{q}).$$

Consideremos dos procesos de Poisson homogéneos $\mathbb{X}_n^-$ y $\mathbb{X}_n^+$ con intensidades nf_{min} y nf_{max}, respectivamente, de manera tal que $\mathbb{X}_n^- \subset \mathbb{X}_n \subset \mathbb{X}_n^+$. Dicha construcción es posible debido al Lema 1 y al hecho que $nf_{min} \leq \lambda_n(\mathbf{x}) \leq nf_{max} \forall \mathbf{x} \in C$. Luego

$$\mathbb{P}\left(n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \leq \mu f_{max}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| - \varepsilon\right) \leq \mathbb{P}\left(n^{\beta}\mathcal{D}_{\mathbb{X}_n^+}(\mathbf{p},\mathbf{q}) \leq \mu f_{max}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| - \varepsilon\right)$$

$$\mathbb{P}\left(n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \geq \mu f_{min}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| + \varepsilon\right) \leq \mathbb{P}\left(n^{\beta}\mathcal{D}_{\mathbb{X}_n^-}(\mathbf{p},\mathbf{q}) \geq \mu f_{min}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| + \varepsilon\right).$$

Eligiendo n_0 de manera tal que $\varepsilon > c_4(f_{min}n)^{-1/3d}$ y que $f_{min}n \geq 1$ para todo $n > n_0$, utilizando la Proposición 3 se desprenden (3.14) y (3.15). □

3.3.2. Geodésicas de longitud acotada

Dado el $\mathbb{X}_n$-camino de partículas $(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_{K_n}^*)$ que conecta $\mathbf{p}$ con $\mathbf{q}$ y realiza $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})$, definimos la curva rectificable Γ_n como la poligonal que va uniendo los puntos $\mathbf{y}_i^*$ con $\mathbf{y}_{i+1}^*$, $i = 1, 2, \ldots, K_n - 1$. Con probabilidad uno este camino es único y está bien definido. Es claro que $\{\Gamma_n\}_{n\in\mathbb{N}}$ es una familia de curvas parametrizables continuas, es decir, que existe una función continua de la cual la curva es la imagen. Llamemos L_n a la longitud de Γ_n dada por

$$L_n = |\,\mathbf{p}-\mathbf{y}_1^*| + \sum_{i=1}^{K_n-1} |\mathbf{y}_{i+1}^* - \mathbf{y}_i^*| + |\mathbf{y}_{K_n}^* - \mathbf{q}\,|. \tag{3.16}$$

Proposición 4 (Caminos óptimos de longitud acotada)**.** *Sea $C \in \mathbb{R}^d$ un conjunto compacto, convexo y con $\bar{C^0} = C$. Sobre C consideremos un proceso puntual de Poisson $\mathbb{X}_n$ con intensidad $\lambda_n = nf(\mathbf{x})$. Dados $\mathbf{p}, \mathbf{q} \in C$ existen constantes positivas $\ell_{max}, c_5, c_6.c_7, n_0$, con c_6 función de $|\,\mathbf{p}-\mathbf{q}\,|$, tales que*

$$\mathbb{P}\left(L_n > \ell_{max}\right) \leq c_5 \exp\left(-c_6 n^{c_7}\right). \tag{3.17}$$

para todo $n > n_0$. En particular, tenemos casi seguramente

$$\limsup_{n\to\infty} L_n \leq \ell_{max}. \tag{3.18}$$

Demostración. A partir de la desigualdad de Hölder se tiene

$$L_n \leq \left(\sum_{i=1}^{K_n - 1} 1^{\frac{\alpha}{\alpha-1}} \right)^{\frac{\alpha-1}{\alpha}} \left(|\,\mathbf{p} - \mathbf{y}_1^*|^\alpha + \sum_{i=1}^{K_N - 1} |\mathbf{y}_i^* - \mathbf{y}_{i+1}^*|^\alpha + |\mathbf{y}_{K_n}^* - \mathbf{q}\,|^\alpha \right)^{\frac{1}{\alpha}},$$

es decir que $L_n^\alpha \leq \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) K_n^{\alpha-1}$. Luego

$$\begin{aligned}
\mathbb{P}\left(L_n > \ell_{max}\right) &\leq \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \left(K_n n^{-1/d}\right)^{\alpha-1} > \ell_{max} L_n^{\alpha-1} \right) \\
&\leq \mathbb{P}\left(f_{max}^\beta n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \left(\frac{K_n}{\lambda_n^{1/d} L_n}\right)^{\alpha-1} > \ell_{max} \right) \\
&\leq \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) > 2\mu f_{min}^{-\beta} |\,\mathbf{p} - \mathbf{q}\,| \right) \qquad (3.19) \\
&+ \mathbb{P}\left(\frac{K_n}{\lambda_n^{1/d} L_n} > \left(\frac{1}{2\mu |\,\mathbf{p}-\mathbf{q}\,|} \left(\frac{f_{min}}{f_{max}}\right)^\beta \ell_{max} \right)^{\frac{1}{\alpha-1}} \right). \qquad (3.20)
\end{aligned}$$

A partir del Lema 2 sabemos que (3.19) está acotado superiormente por una función que decae exponencialmente en n. Por otro lado, vamos a probar que existen constantes positivas c_8, c_9, c_{10}, donde c_{10} depende únicamente de $|\,\mathbf{p}-\mathbf{q}\,|$, tales que

$$\mathbb{P}\left(\frac{K_n}{\lambda_n^{1/d} L_n} > c_8 \right) \leq c_9 \exp\left(-c_{10} n^{1/d}\right), \qquad (3.21)$$

de manera tal que eligiendo ℓ_{max} en (3.20) con

$$\ell_{max} \geq 2\mu c_8^{\alpha-1} \left(\frac{f_{max}}{f_{min}}\right)^\beta |\,\mathbf{p}-\mathbf{q}\,| \qquad (3.22)$$

concluimos (3.17). Vemos como probar (3.21). Consideremos el cubrimiento de $\mathbb{R}^d$ dado por la familia de cubos $(C_i)_{i\in\mathbb{N}}$ de lado $\varepsilon = \varepsilon_0 n^{-1/d}$ con vértices contenidos en $\varepsilon_0 n^{-1/d}\mathbb{Z}^d$. Sea $m_n = \#(\{i \in \mathbb{N} : C_i \cap \Gamma_n \neq \emptyset\})$. Luego

$$K_n \leq \sum_{i: C_i \cap \Gamma_n \neq \emptyset} X_i,$$

donde $X_i = \#(\mathbb{X}_n \cap C_i) \sim \text{Poiss}(n\int_{C_i} f)$. Consideremos el evento

$$\begin{aligned}
E_n^m = \{&\exists \text{ un camino } C_{i_1}, \ldots, C_{i_m} \text{ formado por } m \text{ celdas consecutivas} \\
&\text{y que contiene al menos } m/2d \text{ partículas}\}.
\end{aligned}$$

Sea una familia de m celdas distintas $C_{i_1}, C_{i_2}, \ldots, C_{i_m}$. Luego, $U_m = \sum_{j=1}^m X_{i_j} \sim \text{Poiss}(n\int_{\cup C_{i_j}} f)$. Dada $V_m \sim \text{Poiss}(m\varepsilon_0^d f_{max})$, como $n\int_{\cup C_{i_j}} f \leq m\varepsilon_0^d f_{max}$, se tiene $U_m \prec_{\text{st}} V_m$. Recurriendo a cotas de Chernoff obtenemos

$$\begin{aligned}
\mathbb{P}\left(U_m \geq \frac{m}{2d}\right) &\leq \mathbb{P}\left(V_m \geq \frac{m}{2d}\right) \\
&= \mathbb{P}\left(e^{\theta V_m} \geq e^{\frac{\theta m}{2d}}\right) \\
&\leq \exp\left(-\frac{\theta m}{2d}\right) \mathbb{E}\left[e^{\theta V_m}\right] \\
&= \exp\left(-\frac{\theta m}{2d} + m\varepsilon_0^d f_{max}(e^\theta - 1)\right) \quad \forall \theta \in \mathbb{R}. \qquad (3.23)
\end{aligned}$$

La cantidad de posibles caminos formados por m celdas adyacentes que unen $\mathbf{p}$ con $\mathbf{q}$ está acotada superiormente por $(2d)^m$, pues partiendo desde cualquier celda es posible moverse a lo sumo a alguna de

las $2d$ celdas vecinas. Por lo tanto

$$\mathbb{P}(E_n^m) \leq \left[(2d)\exp\left(-\frac{\theta}{2d}\right)\exp\left(\varepsilon_0^d f_{max}(e^\theta - 1)\right)\right]^m.$$

Sea $\theta > 0$ tal que $(2d)e^{-\theta/2d} < e^{-1}/2$ y $\varepsilon_0 > 0$ tal que $e^{\varepsilon_0^d f_{max}(e^\theta - 1)} < 2$, de manera tal que $\mathbb{P}(E_n^m) \leq e^{-m}$. Notemos que cualquier camino conexo que conecte $\mathbf{p}$ con $\mathbf{q}$ debe atravesar por lo menos $\eta_1 \varepsilon_0^{-1} |\,\mathbf{p} - \mathbf{q}\,| n^{1/d}$ celdas, con $\eta_1 > 0$ alguna constante geométrica que depende de d. Sea el evento

$$F_n = \left\{\frac{m_n}{2d} \leq K_n\right\} \subset \bigcup_{m \geq \eta_1 \varepsilon_0^{-1} |\,\mathbf{p} - \mathbf{q}\,| n^{1/d}} E_n^m,$$

de manera tal que vale

$$\mathbb{P}(F_n) \leq \sum_{m=\lfloor \eta_1 \varepsilon_0^{-1} |\,\mathbf{p} - \mathbf{q}\,| n^{1/d} \rfloor}^{\infty} \mathbb{P}\left(E_n^m\right) \leq e(1 - e^{-1})^{-1} e^{-\eta_1 \varepsilon_0^{-1} |\,\mathbf{p} - \mathbf{q}\,| n^{1/d}}.$$

Sea el camino óptimo $(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_{K_n}^*)$ y $(\nu_1, \nu_2, \ldots, \nu_{m_n})$ el camino conexo de celdas atravesadas por el camino óptimo. En F_n^c hay al menos $m_n/3d$ índices i para los cuales se cumple que d es divisor de i, $i + d - 1 < m_n$ y ν_j no contiene ninguna partícula para todo j con $i \leq j < i + d$. Luego, a partir del Principio del Palomar es fácil ver que cada uno de estos trozos de camino óptimo que atraviesa d celdas desocupadas aporta al menos ε a la longitud de la curva L_n, es decir $(m_n/3d)\varepsilon \leq L_n$. Luego

$$K_n \leq \frac{m_n}{2d} \leq \frac{3}{2\varepsilon_0} n^{1/d} L_n \leq \frac{3}{2\varepsilon_0 f_{min}^{1/d}} \lambda_n^{1/d} L_n \quad \text{en } F_n^c,$$

es decir, eligiendo

$$c_8 = \frac{3}{2\varepsilon_0 f_{min}^{1/d}} \quad , \quad c_9 = e(1 - e^{-1})^{-1} \quad , \quad c_{10} = \eta_1 \varepsilon_0^{-1} |\,\mathbf{p} - \mathbf{q}\,|$$

se desprende (3.21). Finalmente, tomando $c_6 = \min\{c_2(|\,\mathbf{p} - \mathbf{q}\,|), c_{10}(|\,\mathbf{p} - \mathbf{q})|\}$ concluimos la proposición. □

3.3.3. Existencia de la curva que realiza la distancia de Fermat

Otro resultado que va a ser de importante para probar la convergencia de las curvas Γ_n es el siguiente lema, el cual establece condiciones suficientes para que una familia de curvas sea compacta.

Lema 3 (Myers (1945))**.** *Sea en un espacio métrico compacto E una familia $\mathcal{S}$ de curvas continuas con las siguientes propiedades:*

1. *Cada una de las curvas en $\mathcal{S}$ puede ser parametrizada de manera rectificable.*
2. *El límite inferior l de la longitud de las curvas en $\mathcal{S}$ es finito.*

Entonces existe una subsucesión Γ_k de curvas en $\mathcal{S}$ paramétrizadas por una función $h_k : [0,1] \mapsto E$ tales que h_k converge uniformemente a alguna función $h : [0,1] \mapsto E$ asociada a una curva continua con longitud no mayor que l.

En particular, el lema establece que una familia de curvas continuas y rectificables contenidas en un conjunto compacto C y con longitud acotada por alguna constante es un espacio compacto con la siguiente

métrica

$$\ell_{\text{Myers}}(\gamma,\sigma) = \min_{\substack{h:[0,1]\mapsto C \text{ parametrización de } \gamma \\ g:[0,1]\mapsto C \text{ parametrización de } \sigma}} \max_{t\in[0,1]} |h(t)-g(t)|. \tag{3.24}$$

Dado $\delta > 0$ y una curva γ sea la δ-dilatación de γ dada por

$$\gamma_\delta = \{\mathbf{r} \in C : \exists \mathbf{s} \in \gamma \text{ con } |\mathbf{r}-\mathbf{s}| < \delta\}. \tag{3.25}$$

Notemos que el hecho de que las dos curvas $\gamma, \sigma \subset C$ cumplan $\ell_{\text{Myers}}(\gamma,\sigma) < \delta$ implica $\gamma \subset \sigma_\delta$ y $\sigma \subset \gamma_\delta$.

Proposición 5. *Sea $C \subset \mathbb{R}^d$ un dominio compacto arcoconexo y una función $h : C \to \mathbb{R}_{\geq 0}$ continua. Luego, dados $\mathbf{p}, \mathbf{q} \in C$, existe una curva $\Gamma^* \subset C$ continua y rectificable que conecta $\mathbf{p}$ con $\mathbf{q}$ tal que*

$$\int_{\Gamma^*} h = \inf_{\Gamma \subset C} \int_\Gamma h, \tag{3.26}$$

donde el ínfimo se realiza sobre todas las curvas continuas y rectificables Γ contenidas en C y que conectan $\mathbf{p}$ con $\mathbf{q}$.

Demostración. Sea $(\Gamma_n)_{n\in\mathbb{N}}$ una sucesión de curvas continuas rectificables contenidas en C y que conectan $\mathbf{p}$ con $\mathbf{q}$, de manera tal que

$$\lim_{n\to\infty} \int_{\Gamma_n} h = \inf_{\Gamma\subset C} h.$$

Luego, a partir del Lema 3 se tiene que existe una curva rectificable y continua Γ^* parametrizada por alguna función $P : [0,1] \mapsto C$ y una subsucesión $(\Gamma_{n_k})_{k\in\mathbb{N}}$ parametrizadas por $P_k : [0,1] \mapsto C$ de manera tal que P_k converge uniformemente a P. Dado que h es continua sobre un compacto, es uniformemente continua y por lo tanto se tiene que $h \circ P_k$ converge uniformemente a $h \circ P$. Luego

$$\lim_{n\to\infty} \int_{\Gamma_n} h = \lim_{k\to\infty} \int_0^1 (h\circ P_k)(s)ds = \int_0^1 (h\circ P)(s)ds = \int_{\Gamma^*} h.$$

Concluimos que el ínfimo se realiza sobre la curva Γ^*. □

3.3.4. Restricción a un entorno

Vamos a demostrar que dados dos puntos $\mathbf{p}, \mathbf{q} \in C$ podemos restringir la búsqueda del $\mathbb{X}_n$-camino que realiza $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})$ a un entorno del segmento que conecta $\mathbf{p}$ con $\mathbf{q}$ y cuyo diámetro sea proporcional a $|\,\mathbf{p}-\mathbf{q}\,|$.

Lema 4. *Sean $\mathbf{p}, \mathbf{q} \in C$. Luego, existen constantes positivas a, c_{11} independientes de $\mathbf{p}$ y $\mathbf{q}$ y c_{12}, n_0 que dependen de $|\,\mathbf{p}-\mathbf{q}\,|$ tales que*

$$\mathbb{P}\left(\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \neq \mathcal{D}_{\mathbb{X}_n \cap [\![\mathbf{p},\mathbf{q}]\!]_{a|\,\mathbf{p}-\mathbf{q}\,|}}(\mathbf{p},\mathbf{q})\right) \leq c_{11}\exp\left(-c_{12}n^{c_3}\right) \tag{3.27}$$

para todo $n > n_0$.

Demostración. Sea $\mathbf{r} \notin [\![\mathbf{p},\mathbf{q}]\!]_{a|\,\mathbf{p}-\mathbf{q}\,|}$. Dado $\delta_1 < \mu f_{min}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,|/3$, consideremos los eventos

$$A_n(\mathbf{r}) = \left\{ n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{r}) \leq n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) + \delta_1 \right\}$$
$$B_n(\mathbf{r}) = \left\{ n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{r}) \geq \mu f_{max}^{-\beta}|\,\mathbf{p}-\mathbf{r}| - \delta_1 \right\}$$

$$C_n = \left\{ n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \leq \mu f_{min}^{-\beta} |\,\mathbf{p}-\mathbf{q}\,| + \delta_1 \right\}.$$

Sobre $A_n(\mathbf{r}) \cap B_n(\mathbf{r}) \cap C_n$ se cumple que

$$\mu f_{max}^{-\beta} |\,\mathbf{p}-\mathbf{r}| \leq n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{r}) + \delta_1 \leq n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) + 2\delta_1 \leq \mu f_{min}^{-\beta} |\,\mathbf{p}-\mathbf{q}\,| + 3\delta_1 < 2\mu f_{min}^{-\beta} |\,\mathbf{p}-\mathbf{q}\,|.$$

Luego, para toda elección posible de $\mathbf{r} \notin [\![\mathbf{p},\mathbf{q}]\!]_{a|\,\mathbf{p}-\mathbf{q}\,|}$ se tiene $|\,\mathbf{p}-\mathbf{r}| > a|\,\mathbf{p}-\mathbf{q}\,|$. Eligiendo

$$a = 3\left(\frac{f_{max}}{f_{min}}\right)^\beta \tag{3.28}$$

tenemos que $\mathbb{P}(A_n(\mathbf{r}) \cap B_n(\mathbf{r}) \cap C_n) = 0$ y a partir del Lema 2 existen $c_2 = c_2(|\,\mathbf{p}-\mathbf{q}\,|)$, $n_0 = n_0(|\,\mathbf{p}-\mathbf{q}\,|)$ independientes de $\mathbf{r}$ y constantes positivas c_1, c_3 tales que

$$\mathbb{P}(A_n(\mathbf{r})) \leq \mathbb{P}(B_n^c(\mathbf{r})) + \mathbb{P}(C_n^c) \leq 2c_1 \exp\left(-c_2(f_{min}n)^{c_3}\right) \quad \forall n > n_0.$$

Supongamos que $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})$ es estrictamente menor que el estimador de la distancia de Fermat restringida al conjunto $C \cap [\![\mathbf{p},\mathbf{q}]\!]_{a|\,\mathbf{p}-\mathbf{q}\,|}$. En tal caso existe una partícula $\mathbf{z} \in \mathbb{X}_n \cap [\![\mathbf{p},\mathbf{q}]\!]^c_{a|\,\mathbf{p}-\mathbf{q}\,|}$ tal que

$$\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{z}) + \mathcal{D}_{\mathbb{X}_n}(\mathbf{z},\mathbf{q}) \geq \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{z}).$$

Consideremos el siguiente cubrimiento por bolas

$$C \setminus [\![\mathbf{p},\mathbf{q}]\!]_{a|\,\mathbf{p}-\mathbf{q}\,|} \subset \bigcup_{\mathbf{v}\in\mathcal{V}} B\left(\mathbf{v}, \delta_0 n^{-1/d}\right)$$

donde $\mathcal{V} \subset C \setminus [\![\mathbf{p},\mathbf{q}]\!]_{a|\,\mathbf{p}-\mathbf{q}\,|}$ es un conjunto de puntos fijos en el espacio elegido de manera tal que existe una constante $\eta_2 > 0$ con $\#(\mathcal{V}) < \eta_2 n$ y $\delta_0^\alpha < \delta_1$. Sea $\mathbf{v}_\mathbf{z} \in \mathcal{V}$ tal que $\mathbf{z} \in B\left(\mathbf{v}_\mathbf{z}, \delta_0 n^{-1/d}\right)$. A partir de la desigualdad triangular obtenemos

$$n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{z}) \geq n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{v}_z) - n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{z},\mathbf{v}_z) \geq n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{v}_z) - \delta_0^\alpha n^{-1/d} \geq n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{v}_z) - \delta_1.$$

Finalmente tenemos

$$\begin{aligned}
\mathbb{P}\left(\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \neq \mathcal{D}_{\mathbb{X}_n \cap [\![\mathbf{p},\mathbf{q}]\!]_{a|\,\mathbf{p}-\mathbf{q}\,|}}(\mathbf{p},\mathbf{q})\right) &\leq \mathbb{P}\left(\exists \mathbf{v} \in \mathcal{V} : n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \geq n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{v}) - \delta_1\right) \\
&\leq \sum_{\mathbf{v}\in\mathcal{V}} \mathbb{P}\left(A_n(\mathbf{v})^c\right) \\
&\leq 2c_1\eta_2 n \exp\left(-c_2(f_{min}n)^{c_3}\right) \quad \forall n > n_0.
\end{aligned}$$

Tomando $c_{11} = 2c_1\eta_2$ y c_{12} tal que $-c_{12}n < -c_2 f_{min}^{c_3} n^{c_3} + \log n$ obtenemos (3.27), donde al igual que c_2 tenemos que c_{12} es función de $|\,\mathbf{p}-\mathbf{q}\,|$. □

3.3.5. Espaciado entre puntos consecutivos del camino óptimo

Para finalizar con los lemas preliminares antes de la demostración del teorema, vamos a estudiar cómo se comporta el espaciado entre puntos consecutivos del $\mathbb{X}_n$-camino óptimo $(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_{K_n}^*)$ que realiza $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})$.

Lema 5. *Dados $\delta > 0$ y $0 \leq \gamma < 1/d$, existen constantes positivas c_{13}, c_{14}. tales que*

$$\mathbb{P}\left(\max_{i<K_n} |\mathbf{y}_i^* - \mathbf{y}_{i+1}^*| > \delta n^{-\gamma}\right) \leq c_{13} n^{\gamma/d} \exp\left(-c_{14} n^{1-\gamma d}\right). \tag{3.29}$$

En particular

$$\mathbb{P}\left(\max_{i<K_n} |\mathbf{y}_i^* - \mathbf{y}_{i+1}^*| > \delta \right) \leq c_{13} \exp\left(-c_{14} n\right). \tag{3.30}$$

Demostración. Dado cualquier par de puntos consecutivos $\mathbf{y}_i^*, \mathbf{y}_{i+1}^*$ del camino óptimo se tiene que

$$\mathbb{X}_n \cap \left\{ \mathbf{x} \in C : |\mathbf{x} - \mathbf{y}_{i+1}^*|^\alpha + |\mathbf{x} - \mathbf{y}_i^*|^\alpha < |\mathbf{y}_{i+1}^* - \mathbf{y}_i^*|^\alpha \right\} = \emptyset.$$

En particular, $\max_{i<K_n} |\mathbf{y}_i^* - \mathbf{y}_{i+1}^*| > \delta n^{-\gamma}$ implica que existe una región sobre C con volumen $\delta^d n^{-\gamma/d}$ sobre la cual no hay partículas. Dicha región va a contener estrictamente a un cubo de lado $\eta_3 \delta n^{-\gamma}$, siendo η_3 alguna constante que depende de α y d. Luego, consideremos la familia cubos con vértices en nodos adyacentes de la red $\eta_3 \delta n^{-\gamma} / 2\mathbb{Z}^d$. Notemos que existen $\mathcal{O}(n^{\gamma d})$ de estos cubos. Por otro lado, la probabilidad de que no hayan partículas en uno de los cubos es $\mathcal{O}\left(\exp(c_{14} n^{1-\gamma d})\right)$. □

3.3.6. Prueba del caso Poisson no homogéneo

El siguiente resultado prueba la convergencia del *estimador de la distancia de Fermat* $\mathcal{D}_{\mathbb{X}_n}(\cdot,\cdot)$ a un objeto macroscópico no trivial para el caso donde $\mathbb{X}_n$ es un proceso puntual de Poisson no homogéneo sobre un conjunto compacto C. El mismo tiene interés en si mismo porque describe el comportamiento de las geodésicas para un proceso puntual de Poisson no homogéneo.

Teorema 2. *Sea $C \subset \mathbb{R}^d$ un conjunto convexo, compacto y tal que $\bar{C^o} = C$. Consideremos $f : C \mapsto \mathbb{R}_{\geq 0}$ una función de densidad continua con $f_{min} = \min_{\mathbf{x}\in C} f(\mathbf{x}) > 0$. Sea $\mathbb{X}_n$ un proceso puntual de Poisson sobre C con intensidad $nf(\mathbf{x})$. Luego, dados $\mathbf{p}, \mathbf{q}$ en el interior de C se tiene*

$$\lim_{n\to\infty} n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \mu \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta} \quad \textit{casi seguramente,} \tag{3.31}$$

donde la minimización se realiza sobre todas las curvas continuas y rectificables Γ contenidas en C y que conectan $\mathbf{p}$ con $\mathbf{q}$. Más aun, si existe una única curva $\tilde{\Gamma} \subset C$ rectificable y continua tal que

$$\int_{\tilde{\Gamma}} \frac{1}{f^\beta} = \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta}, \tag{3.32}$$

entonces las curvas Γ_n que realizan $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})$ convergen casi seguramente a $\tilde{\Gamma}$ con la topología inducida por ℓ_{Myers}.

Demostración. Dado $\varepsilon > 0$ tenemos

$$\begin{aligned}\mathbb{P}\left(\left| n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) - \mu \inf_{\Gamma\subset C} \int_\Gamma \frac{1}{f^\beta} \right| > \varepsilon \right) &= \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) > \mu \inf_{\Gamma\subset C} \int_\Gamma \frac{1}{f^\beta} + \varepsilon \right) \\ &\quad + \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) < \mu \inf_{\Gamma\subset C} \int_\Gamma \frac{1}{f^\beta} - \varepsilon \right).\end{aligned} \tag{3.33}$$

Para ver la convergencia casi segura, probemos que ambas probabilidades involucradas en (3.33) son sumables en n.

Consideremos una curva $\Gamma^* \subset C$ continua y rectificable tal que $\int_{\Gamma^*} \frac{1}{f^\beta} < \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} + \varepsilon/(4\mu)$. Tomando $\varepsilon < 1$, tenemos que la longitud $|\Gamma^*|$ está superiormente acotada por:

$$|\Gamma^*| < \ell_{max}^* = f_{max}^\beta \left(\inf_\Gamma \int_\Gamma \frac{1}{f^\beta} + \frac{1}{4\mu} \right). \tag{3.34}$$

Consideremos un conjunto finito de puntos $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M$ sobre Γ^* de manera tal que $\mathbf{r}_1 = \mathbf{p}$, $\mathbf{r}_M = \mathbf{q}$ y $\delta/2 < |\mathbf{r}_{i+1} - \mathbf{r}_i| < \delta$. Notemos que en tal caso $M = M(\delta) < \lfloor 2\ell^*_{max}/\delta \rfloor$. Luego

$$\int_{\Gamma^*} \frac{1}{f^\beta} = \sum_{i=1}^{M-1} \int_{\Gamma_i^*} \frac{1}{f^\beta},$$

donde Γ_i^* es el tramo de la curva Γ^* que conecta los puntos $\mathbf{r}_i$ y $\mathbf{r}_{i+1}$. Dado que la función $f^{-\beta}$ es integrable Riemann sobre Γ^*, existe δ_1 tal que si $|\mathbf{r}_{i+1} - \mathbf{r}_i| < \delta_1$ entonces

$$\sum_{i=1}^{M-1} \frac{1}{\left[\text{mín}_{\Gamma_i^*} f\right]^\beta} |\mathbf{r}_i - \mathbf{r}_{i+1}| < \int_{\Gamma^*} \frac{1}{f^\beta} + \frac{\varepsilon}{4}.$$

Elegimos $\delta = \text{mín}\{\delta_1, 1\}$. Por otro lado, como f es continua sobre un conjunto compacto y está acotada inferiormente por $f_{min} > 0$, se tiene que $f^{-\beta}$ es uniformemente continua. Dado $\varepsilon_2 < \varepsilon f_{min}^\beta/(4\mu M)$ existe $\delta_2 > 0$ tal que $|\mathbf{r} - \mathbf{s}| < \delta_2$ implica $|f(\mathbf{r})^{-\beta} - f(\mathbf{s})^{-\beta}| < \varepsilon_2$. Si para cada $i = 1, 2, \ldots, M-1$ consideramos el conjunto $C_i = \{\mathbf{r} \in C : \exists \mathbf{u} \in \Gamma_i^* \text{ con } |\mathbf{u} - \mathbf{r}| \leq \delta_2/2\}$, es claro

$$\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \leq \mathcal{D}_{\mathbb{X}_n \cap \left(\cup_{i=1}^{M-1} C_i\right)}(\mathbf{p}, \mathbf{q}) \leq \sum_{i=1}^{M-1} \mathcal{D}_{\mathbb{X}_n \cap C_i}(\mathbf{r}_i, \mathbf{r}_{i+1}). \tag{3.35}$$

Por otro lado tenemos que

$$\begin{aligned}
\mu \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta} + \varepsilon &> \mu \int_{\Gamma^*} \frac{1}{f^\beta} + \frac{3\varepsilon}{4} \\
&> \mu \sum_{i=1}^{M-1} \frac{1}{\left[\text{mín}_{\Gamma_i^*} f\right]^\beta} |\mathbf{r}_{i+1} - \mathbf{r}_i| + \frac{\varepsilon}{2} \\
&> \mu \sum_{i=1}^{M-1} \frac{1 - \varepsilon_2}{\left[\text{mín}_{C_i} f\right]^\beta} |\mathbf{r}_{i+1} - \mathbf{r}_i| + \frac{\varepsilon}{2} \qquad (3.36) \\
&\geq \mu \sum_{i=1}^{M-1} \frac{1}{\left[\text{mín}_{C_i} f\right]^\beta} |\mathbf{r}_{i+1} - \mathbf{r}_i| + \frac{\varepsilon}{2} - \frac{\mu M \delta}{f_{min}^\beta} \varepsilon_2 \\
&> \mu \sum_{i=1}^{M-1} \frac{1}{\left[\text{mín}_{C_i} f\right]^\beta} |\mathbf{r}_{i+1} - \mathbf{r}_i| + \frac{\varepsilon}{4},
\end{aligned}$$

donde en (3.36) usamos la continuidad uniforme de $f^{-\beta}$ para remplazar el mínimo sobre Γ_i^* por el mínimo sobre C_i. Luego

$$\begin{aligned}
\mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \geq \mu \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta} + \varepsilon\right) &\leq \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n \cap \left(\cup_{i=1}^{M-1} C_i\right)}(\mathbf{p}, \mathbf{q}) \geq \mu \int_{\Gamma^*} \frac{1}{f^\beta} + \frac{3\varepsilon}{4}\right) \qquad (3.37) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{M-1} n^\beta \mathcal{D}_{\mathbb{X}_n \cap C_i}(\mathbf{r}_i, \mathbf{r}_{i+1}) \geq \mu \sum_{i=1}^{M-1} \frac{1}{\left[\text{mín}_{C_i} f\right]^\beta} |\mathbf{r}_{i+1} - \mathbf{r}_i| + \frac{\varepsilon}{4}\right) \\
&\leq \sum_{i=1}^{M-1} \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n \cap C_i}(\mathbf{r}_i, \mathbf{r}_{i+1}) \geq \mu \frac{1}{\left[\text{mín}_{C_i} f\right]^\beta} |\mathbf{r}_{i+1} - \mathbf{r}_i| + \frac{\varepsilon}{4M}\right) \\
&\leq M c_1 \exp\left(-c_2 (f_{min} n)^{c_3}\right) \quad \forall n > n_0(\varepsilon)
\end{aligned}$$

donde utilizamos el Lema 2 y la constante c_2 depende únicamente de δ.

Analicemos ahora la otra probabilidad involucrada en (3.33). Dado $\varepsilon > 0$, llamemos $(p_n)_{n \in \mathbb{N}}$ a la sucesión definida por

$$p_n = \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \leq \mu \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta} - \varepsilon\right). \tag{3.38}$$

Dado $\delta > 0$, sea el evento $E_n = \{\text{máx}_{j<K_n} |\mathbf{y}_j^* - \mathbf{y}_{j+1}^*| < \delta/2\}$. Sobre $\{L_n \leq \ell_{max}\} \cap E_n$ se tiene que existen partículas aleatorias $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k \in \Gamma_n \cap \mathbb{X}_n$ con $\delta/2 < |\mathbf{z}_{i+1} - \mathbf{z}_i| < \delta$ para todo $i = 1, 2, \ldots, k-1$, donde $k \leq k_{max} = 2\ell_{max}\delta^{-1}$. Luego

$$\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) = \sum_{i=0}^{k} \mathcal{D}_{\mathbb{X}_n}(\mathbf{z}_i, \mathbf{z}_{i+1}), \tag{3.39}$$

donde notamos por $\mathbf{z}_0 = \mathbf{p}$, $\mathbf{z}_{k+1} = \mathbf{q}$. Consideremos un cubrimiento del conjunto C de la forma

$$C \subset \bigcup_{\mathbf{v} \in \mathcal{V}} B\left(\mathbf{v}, \delta_0 n^{-1/d}\right), \tag{3.40}$$

siendo $\delta_0 > 0$ y $\mathcal{V} \subset C$ elegido de manera tal que exista una constante η_4 con $\#(\mathcal{V}) \leq \eta_4 n$. Sean $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k \in \mathcal{V}$ tales que $\mathbf{z}_i \in B(\mathbf{w}_i, \delta_0 n^{-1/d})$ para toda elección de $i \leq k$. Luego, eligiendo δ_0 de manera tal que $2\delta_0^{\alpha} k_{max} < \varepsilon/2$, utilizando la desigualdad triangular en (3.39) obtenemos

$$\begin{aligned} n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) &\geq \sum_{i=0}^{k} n^{\beta} \left[\mathcal{D}_{\mathbb{X}_n}(\mathbf{w}_i, \mathbf{w}_{i+1}) - \mathcal{D}_{\mathbb{X}_n}(\mathbf{w}_i, \mathbf{z}_i) - \mathcal{D}_{\mathbb{X}_n}(\mathbf{w}_{i+1}, \mathbf{z}_{i+1})\right] \\ &\geq \sum_{i=0}^{k} n^{\beta} \left[\mathcal{D}_{\mathbb{X}_n}(\mathbf{w}_i, \mathbf{w}_{i+1}) - 2\delta_0^{\alpha} n^{-\alpha/d}\right] \\ &\geq \sum_{i=0}^{k} n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{w}_i, \mathbf{w}_{i+1}) - \frac{\varepsilon}{2}. \end{aligned}$$

Sea también $\delta_0 < \delta/8$ de manera tal que $|\mathbf{w}_i - \mathbf{w}_{i+1}| > |\mathbf{z}_i - \mathbf{z}_{i+1}| - |\mathbf{w}_i - \mathbf{z}_i| - |\mathbf{w}_{i+1} - \mathbf{z}_{i+1}| > \delta/2 - \delta/8 - \delta/8 = \delta/4$. Luego

$$\begin{aligned} p_n \leq \mathbb{P}\Bigg(&\exists \mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathcal{V} \text{ con } k \leq k_{max} \text{ y } \frac{\delta}{8} < |\mathbf{v}_i - \mathbf{v}_{i+1}| < \delta \text{ tales que} \\ &\sum_{i=0}^{k} n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{v}_i, \mathbf{v}_{i+1}) \leq \mu \inf_{\Gamma \subset C} \int_{\Gamma} \frac{1}{f^{\beta}} - \frac{\varepsilon}{2} \ , \ L_n \leq \ell_{max} \ , \ E_n\Bigg) + \mathbb{P}\left(L_n > \ell_{max}\right) + \mathbb{P}\left(E_n^c\right) \end{aligned} \tag{3.41}$$

La Proposición 4 y el Lema 5 establecen los últimos dos sumandos en (3.41) son sumables en n. Por otro lado, notemos que la cantidad de caminos posibles $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in \mathcal{V}$ donde $k \leq k_{max}$ está acotada superiormente por $(\eta_4 n)^{k_{max}}$. Consideremos cualquiera de estos caminos. Sean los eventos

$$\begin{aligned} A_i &= \left\{\mathcal{D}_{\mathbb{X}_n}(\mathbf{v}_i, \mathbf{v}_{i+1}) = \mathcal{D}_{\mathbb{X}_n \cap [\![\mathbf{v}_i, \mathbf{v}_{i+1}]\!]_{a|\mathbf{v}_i - \mathbf{v}_{i+1}|}}(\mathbf{v}_i, \mathbf{v}_{i+1})\right\} \\ B_i &= \left\{n^{\beta}\mathcal{D}_{\mathbb{X}_n \cap [\![\mathbf{v}_i, \mathbf{v}_{i+1}]\!]_{a|\mathbf{v}_i - \mathbf{v}_{i+1}|}}(\mathbf{v}_i, \mathbf{v}_{i+1}) \geq \mu \frac{1}{\text{máx}_{\mathbf{x} \in [\![\mathbf{v}_i, \mathbf{v}_{i+1}]\!]_{a|\mathbf{v}_i - \mathbf{v}_{i+1}|}} f(\mathbf{x})^{\beta}} |\mathbf{v}_i - \mathbf{v}_{i+1}| - \frac{\varepsilon}{8k_{max}}\right\} \cap A_i. \end{aligned}$$

De los Lemas 2 y 4 se desprende que

$$\mathbb{P}(B_i^c) \leq c_1 \exp\left(-c_2 f_{min}^{c_3} n^{c_3}\right) + c_{11} \exp\left(-c_{12} n^{c_3}\right) \quad \forall n > n_0, \quad i = 1, 2, \ldots, k-1, \tag{3.42}$$

donde c_2, c_{12} y n_0 dependen de δ. Sea $\delta > 0$ elegido de tal manera que $|\mathbf{r} - \mathbf{s}| < (a+1)\delta$ asegure que $|f^{-\beta}(\mathbf{r}) - f^{-\beta}(\mathbf{s})| < \varepsilon_3 = \varepsilon\mu^{-1} f_{min}^{\beta} \ell_{max}^{-1}/4$. Como $f^{-\beta}$ es continua en un compacto, es uniformemente continua y por lo tanto siempre va a existir dicho $\delta > 0$. En tal caso tenemos

$$\sum_{i=0}^{k} \frac{1}{\text{máx}_{[\![\mathbf{v}_i, \mathbf{v}_{i+1}]\!]_{a|\mathbf{v}_i - \mathbf{v}_{i+1}|}} f^{\beta}} |\mathbf{v}_i - \mathbf{v}_{i+1}| > \sum_{i=0}^{k} \frac{1 - \varepsilon_3}{\text{mín}_{[\![\mathbf{v}_i, \mathbf{v}_{i+1}]\!]} f^{\beta}} |\mathbf{v}_i - \mathbf{v}_{i+1}| > \int_{\gamma(\mathbf{v}_0, \ldots, \mathbf{v}_k)} \frac{1}{f^{\beta}} - \varepsilon_3 f_{min}^{-\beta} \ell_{max}$$

donde $[\![\mathbf{v}_i, \mathbf{v}_{i+1}]\!]$ es el segmento que une $\mathbf{v}_i$ con $\mathbf{v}_{i+1}$ y $\gamma(\mathbf{v}_0, \dots, \mathbf{v}_k)$ es la poligonal que conecta los puntos $\mathbf{v}_0 = \mathbf{p}, \mathbf{v}_1, \dots, \mathbf{v}_k = \mathbf{q}$. Luego

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=0}^{k} n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{v}_i, \mathbf{v}_{i+1}) \leq \mu \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta} - \frac{\varepsilon}{2} \ , \ L_n \leq \ell_{max} \ , \ E_n\right) \\
\leq\ &\mathbb{P}\Bigg(\sum_{i=0}^{k} \left[\mu \frac{1}{\max_{[\![\mathbf{v}_i, \mathbf{v}_{i+1}]\!]_{a|\mathbf{v}_i - \mathbf{v}_{i+1}|}} f^\beta} |\mathbf{v}_i - \mathbf{v}_{i+1}| - \frac{\varepsilon}{8k_{max}}\right] \leq \mu \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta} - \frac{\varepsilon}{2} \ , \\
&\qquad L_n \leq \ell_{max} \ , \ E_n \ , \ \bigcap_{i=0}^{k} B_i\Bigg) \quad + \quad \sum_{i=0}^{k} \mathbb{P}(B_i^c) \\
\leq\ &\mathbb{P}\left(\mu \int_{\gamma(\mathbf{v}_0, \dots, \mathbf{v}_k)} \frac{1}{f^\beta} \leq \mu \inf_{\Gamma \subset C} \int_\Gamma \frac{1}{f^\beta} - \frac{\varepsilon}{8} \ , \ L_n \leq \ell_{max} \ , \ E_n \ , \ \bigcap_{i=0}^{k} B_i\right) + \sum_{i=0}^{k} \mathbb{P}(B_i^c). \qquad (3.43)
\end{aligned}$$

Notemos que la primer probabilidad involucrada en (3.43) es igual a cero, dado que implica que la curva $\gamma(\mathbf{v}_0, \dots, \mathbf{v}_k)$ integra menos que el ínfimo sobre todas las curvas $\Gamma \subset C$. Juntando todo, concluimos que

$$p_n \leq \mathbb{P}\left(L_n > \ell_{max}\right) + \mathbb{P}\left(E_n^c\right) + \sum_{\substack{\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathcal{V} \\ |\mathbf{v}_i - \mathbf{v}_{i+1}| > \delta/8}} \sum_{i=0}^{k} \mathbb{P}(B_i^c)$$

lo cual también es sumable. Esto concluye la prueba de la parte principal del teorema. Veamos ahora la convergencia de las curvas Γ_n al único óptimo $\hat{\Gamma}$. Dado $\varepsilon_4 > 0$, veamos que el evento $\ell_{\text{Myers}}(\Gamma_n, \hat{\Gamma}) \geq \varepsilon_4$ sólo puede ocurrir un número finito de veces. A partir de los mismos argumentos que usamos en el Lema 5 es fácil ver que existe $\varepsilon_5 > 0$ para el cual

$$\int_{\hat{\Gamma}} \frac{1}{f^\beta} + \varepsilon_5 < \inf_{\Gamma : \ell_{\text{Myers}}(\Gamma, \hat{\Gamma}) \geq \varepsilon_4} \int_\Gamma \frac{1}{f^\beta},$$

pues en caso contrario existiría una sucesión de curvas que no convergen a $\hat{\Gamma}$ pero cuyo estimador de la distancia de Fermat si lo hace, lo cual es un absurdo dado que el ínfimo se realiza únicamente sobre $\hat{\Gamma}$. Notemos que la misma construcción con la cual demostramos la convergencia del estimador de la distancia de Fermat puede ser utilizada para probar que dado $\varepsilon > 0$ existe $\delta > 0$ tal que

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left| n^\beta \mathcal{D}_{\mathbb{X}_n \cap \Gamma_\delta}(\mathbf{p}, \mathbf{q}) - \mu \int_\Gamma \frac{1}{f^\beta} \right| > \varepsilon\right) < \infty \quad \forall \Gamma \text{ con } |\Gamma| < \ell^*_{max}, \qquad (3.44)$$

donde $\Gamma_\delta = \{\mathbf{r} \in C : \exists \mathbf{s} \in \Gamma \text{ con } |\mathbf{r} - \mathbf{s}| < \delta\}$. Tomemos $\varepsilon = \varepsilon_5/2$ y δ_5 tal que vale (3.44). Por otro lado, del Lema 3 se desprende que existe un conjunto finito de curvas $\gamma^1, \gamma^2, \dots, \gamma^m \in C \setminus \{\gamma : \ell_{\text{Myers}}(\gamma, \hat{\Gamma}) < \varepsilon_4\}$ tales que para toda curva $\Gamma \subset C$ continua rectificable, con longitud menor a ℓ^*_{max} y tal que $\ell_{\text{Myers}}(\Gamma_n, \hat{\Gamma}) \geq \varepsilon_4$, existe γ^j con $\ell_{\text{Myers}}(\Gamma, \gamma^j) < \min\{\varepsilon_4, \delta_5\}$. Luego, a partir de la misma estrategia que antes tenemos

$$\begin{aligned}
\mathbb{P}\left(\ell_{\text{Myers}}(\Gamma_n, \gamma^i) < \delta\right) &\leq \mathbb{P}\left(n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) = n^\beta \mathcal{D}_{\mathbb{X}_n \cap \gamma^i_{\delta_5}}(\mathbf{p}, \mathbf{q})\right) \\
&\leq \mathbb{P}\left(\mu \int_{\hat{\Gamma}} \frac{1}{f^\beta} + \frac{\varepsilon_5}{2} > n^\beta \mathcal{D}_{X_n}(\mathbf{p}, \mathbf{q}) = n^\beta \mathcal{D}_{\mathbb{X}_n \cap \gamma^i_{\delta_5}}(\mathbf{p}, \mathbf{q}) > \mu \int_{\gamma^i} \frac{1}{f^\beta} - \frac{\varepsilon_5}{2}\right) \\
&+ \mathbb{P}\left(\left| n^\beta \mathcal{D}_{\mathbb{X}_n \cap \gamma^i_{\delta_5}}(\mathbf{p}, \mathbf{q}) - \mu \int_{\gamma^i} \frac{1}{f^\beta} \right| > \frac{\varepsilon_5}{2}\right) \\
&+ \mathbb{P}\left(\left| n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) - \mu \int_{\hat{\Gamma}} \frac{1}{f^\beta} \right| > \frac{\varepsilon_5}{2}\right)
\end{aligned}$$

de donde se deduce que

$$\sum_{i=1}^{m}\sum_{n=1}^{\infty}\mathbb{P}\Big(\ell_{\text{Myers}}(\Gamma_n,\gamma^i)<\delta\Big)<\infty.$$

Por lo tanto, concluimos que existen finitas Γ_n contenidas en $C\setminus\{\gamma:\ell_{\text{Myers}}(\gamma,\hat{\Gamma})<\varepsilon_4\}$, tal como queríamos probar. □

3.4. Ensamble canónico

Ahora extenderemos el resultado al caso donde el estimador de la distancia de Fermat se calcula sobre un conjunto $\mathbb{X}_n=\{\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n\}$, proveniente de una muestra i.i.d de tamaño n con densidad f. El siguiente corolario del Teorema 2 establece la convergencia del estimador de la distancia de Fermat al mismo objeto macroscópico cuando la cantidad de partículas está fija.

Corolario 1. *Sea $C\subset\mathbb{R}^d$ un conjunto convexo, compacto y tal que $\bar{C^o}=C$. Sea una función de densidad $f:C\mapsto\mathbb{R}_{\geq 0}$ continua con $f_{min}=\min_{\mathbf{x}\in C}f(\mathbf{x})>0$. Sea $(\mathbb{X}_n)_{n\in\mathbb{N}}$ una sucesión de muestras i.i.d de tamaño n con densidad f. Luego, dados $\mathbf{p},\mathbf{q}$ en el interior de C se tiene*

$$\lim_{n\to\infty}n^\beta\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})=\mu\inf_{\Gamma\subset C}\int_\Gamma\frac{1}{f^\beta}\quad\textit{casi seguramente,}\tag{3.45}$$

donde la minimización se realiza sobre todas las curvas continuas y rectificables Γ contenidas en C y que conectan $\mathbf{p}$ con $\mathbf{q}$.

Demostración. Sea $\varepsilon>0$ y consideremos $\mathbb{X}_n^+$, $\mathbb{X}_n^-$ dos procesos puntuales de Poisson sobre C con intensidades $n(1+\varepsilon)f(\mathbf{x})$, $n(1-\varepsilon)f(\mathbf{x})$, respectivamente. Llamemos $M_n^+=\#(\mathbb{X}_n^+)$, $M_n^-=\#(\mathbb{X}_n^-)$. Luego,

$$\lim_{N\to\infty}\frac{M_n^+}{n}=1+\varepsilon,\qquad\lim_{N\to\infty}\frac{M_n^-}{n}=1-\varepsilon,\qquad\text{casi seguramente.}$$

Sea el evento $\Omega_n=\{M_n^-\leq n\leq M_n^+\}$. Usando cotas de Chernoff tal como hicimos en (3.23) podemos ver que existe $c_{15}=c_{15}(\varepsilon)>0$ tal que $\mathbb{P}(\Omega_n^c)\leq e^{-c_{15}n}$ y por lo tanto $\mathbb{P}(\Omega_n^c)$ es sumable para toda elección de $\varepsilon>0$. A partir del Lema de Borel-Cantelli obtenemos que existe N_0 aleatorio tal que para todo $n>N_0$ vale Ω_n, o equivalentemente

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty}\bigcap_{n=k}^{\infty}\Omega_n\right)=1.$$

A partir del Lema 1, podemos construir $\mathbb{X}_n^+,\mathbb{X}_n^-$ de manera tal que sobre el evento Ω_n se tenga $\mathbb{X}_n^-\subseteq\mathbb{X}_n\subseteq\mathbb{X}_n^+$. Luego

$$n^\beta\mathcal{D}_{\mathbb{X}_n^+}(\mathbf{p},\mathbf{q})\leq n^\beta\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})\leq n^\beta\mathcal{D}_{\mathbb{X}_n^-}(\mathbf{p},\mathbf{q})\quad\text{en }\Omega_n.$$

Usando el Teorema 2 y tomando límite inferior y superior obtenemos

$$\liminf_{n\to\infty}n^\beta\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})\geq\frac{1}{(1+\varepsilon)^\beta}\mu\inf_{\Gamma\subset C}\int_\Gamma\frac{1}{f^\beta},\tag{3.46}$$

$$\limsup_{n\to\infty}n^\beta\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})\leq\frac{1}{(1-\varepsilon)^\beta}\mu\inf_{\Gamma\subset C}\int_\Gamma\frac{1}{f^\beta}\quad\text{sobre }\bigcup_{k=1}^{\infty}\bigcap_{n=k}^{\infty}\Omega_n.\tag{3.47}$$

Si llamamos por A_ε al evento definido por las desigualdades (3.46) y (3.47), tenemos que $\mathbb{P}(A_\varepsilon)=1$. Finalmente

$$\mathbb{P}\left(\lim_{n\to\infty}n^\beta\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})=\mu\inf_{\Gamma\subset C}\int_\Gamma\frac{1}{f^\beta}\right)=\mathbb{P}\left(\bigcap_{k=1}^{\infty}A_{\frac{1}{k}}\right)=1,$$

de donde concluimos la prueba. □

3.5. Variedades

En esta sección generalizaremos los resultados obtenidos a variedades con dimensión posiblemente menor al espacio ambiente. Por simplicidad, nos limitaremos al caso donde la variedad se puede escribir como la imagen de una transformación conforme. Es nuestra creencia que los mismos resultados valen para variedades más generales pero la demostración se vuelve más engorrosa.

A partir de este momento notaremos por d a la dimensión intrínseca de la variedad y D la dimensión del espacio ambiente donde está contenida la variedad.

3.5.1. Preliminares

Sea $C \subset \mathbb{R}^d$ un conjunto convexo, compacto y tal que $\bar{C^o} = C$. Consideremos una transformación $\varphi : C \mapsto \mathbb{R}^D$, con $d \leq D$, tal que φ es un difeomorfísmo (es decir, es una transformación diferenciable, biyectiva y con inversa φ^{-1} diferenciable). Definimos la variedad compacta $\mathcal{M}$ como la imagen de φ, es decir, $\mathcal{M} = \varphi(C)$.

Sea $J_\varphi(\mathbf{x}) \in \mathbb{R}^{D\times d}$ la matriz Jacobiana de φ definida como

$$\left(J_\varphi(\mathbf{x})\right)_{ij} = \frac{\partial \varphi_i}{\partial x_j}(\mathbf{x}).$$

La transformación φ se dice conforme si localmente preserva los ángulos, es decir, si para todo $\mathbf{x} \in C$ y todo par de vectores $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$ tangentes a $\mathcal{M}$ en el punto $\varphi(\mathbf{x})$ se tiene

$$\left(J_\varphi(\mathbf{x})\mathbf{v}\right)^t \left(J_\varphi(\mathbf{x})\mathbf{w}\right) = c(\mathbf{x})\mathbf{v}^t\mathbf{w}, \tag{3.48}$$

donde $c(\mathbf{x}) > 0$ es un factor local de escala continuo. Dado que C tiene interior no vacío, la condición (3.48) es equivalente a $J_\varphi(\mathbf{x})^T J_\varphi(\mathbf{x}) = c(\mathbf{x})\mathbb{I}_d$, donde $\mathbb{I}_d$ es la identidad en $\mathbb{R}^d$.

Si $c(\mathbf{x}) = 1$ para todo $\mathbf{x} \in C$ entonces φ se denomina isometría. Las isometrias tienen la propiedad de preservar la longitud de curvas. Dada una curva $\sigma : [0,1] \mapsto C$ diferenciable y $\gamma = \varphi \circ \sigma$, donde φ es una transformación isométrica, si llamamos $L(\cdot)$ a la longitud de una curva, tenemos

$$L(\gamma) = \int_0^1 \left|\frac{d}{ds}\varphi(\sigma(s))\right| ds = \int_0^1 \left|J_\varphi(\sigma(s))\frac{d\sigma}{ds}(s)\right| ds = \int_0^1 \sqrt{c(\sigma(s))}\left|\frac{d\sigma}{ds}(s)\right| ds = L(\sigma). \tag{3.49}$$

Dada la variedad $\mathcal{M}$ y dos puntos $\mathbf{p}, \mathbf{q} \in \mathcal{M}$, se define la distancia geodésica entre $\mathbf{p}$ y $\mathbf{q}$ como la menor longitud de todas las curvas contenidas en $\mathcal{M}$ que conectan $\mathbf{p}$ con $\mathbf{q}$. Más precisamente, la distancia geodésica $\ell_{\mathcal{M}}(\cdot,\cdot)$ está dada por

$$\ell_{\mathcal{M}}(\mathbf{p},\mathbf{q}) = \text{ínf}\left\{L(\gamma) : \gamma(0) = \mathbf{p}, \gamma(1) = \mathbf{q}, \gamma \subset \mathcal{M}\right\}.$$

En particular, dada una isometría $\varphi : C \mapsto \mathcal{M}$ de (3.49) se deduce

$$\ell_{\mathcal{M}}\left(\varphi(\mathbf{x}), \varphi(\mathbf{y})\right) = |\mathbf{x} - \mathbf{y}| \quad \forall \mathbf{x}, \mathbf{y} \in C.$$

El siguiente lema establece que localmente la distancia geodésica y la distancia euclídea entre pares de puntos que se encuentran cerca son similares.

Lema 6. *Sea $\mathcal{M}$ una variedad compacta regular y tomemos $\mathbf{p}, \mathbf{q} \in \mathcal{M}$. Luego, dado $\varepsilon > 0$ existe $\delta > 0$ tal que si $|\,\mathbf{p} - \mathbf{q}\,| < \delta$ entonces*

$$(1-\varepsilon)\ell_{\mathcal{M}}(\mathbf{p},\mathbf{q}) \leq |\,\mathbf{p} - \mathbf{q}\,| \leq \ell_{\mathcal{M}}(\mathbf{p},\mathbf{q}). \tag{3.50}$$

Demostración. La segunda desigualdad es inmediata de la definición de geodésica. Veamos como probar la primer desigualdad. Por simplicidad, notemos $\ell = \ell_{\mathcal{M}}(\mathbf{p}, \mathbf{q})$. Dada una variedad $\mathcal{M}$, se define su mínimo radio de curvatura $r_0 = r_0(\mathcal{M})$ como

$$\frac{1}{r_0} = \max_{\gamma, s}\{|\gamma''(s)|\}$$

donde el máximo se busca entre todas las curvas $\gamma : [0, t] \mapsto \mathcal{M}$ parametrizadas por longitud de arco (es decir, con $|\gamma''| = 1$) y todos los posibles valores $s \in [0, t]$. Para una variedad $\mathcal{M}$ compacta y diferenciable tenemos que r_0 es una función continua sobre $\mathcal{M}$ y estrictamente positiva. Sea $\gamma : [0, \ell] \mapsto \mathcal{M}$ la geodésica parametrizada por longitud de arco que realiza la distancia ℓ entre $\mathbf{p}$ y $\mathbf{q}$. Luego

$$\mathbf{p} - \mathbf{q} = \int_0^\ell \gamma'(s)ds = \int_0^\ell \left[\gamma'(0) + \int_0^s \gamma''(t)dt\right] ds = \gamma'(0)\ell + \int_0^\ell \int_0^s \gamma''(t)dtds.$$

Tomando norma y acotando por r_0 obtenemos

$$|\,\mathbf{p} - \mathbf{q} - \ell\gamma'(0)| \leq \frac{\ell^2}{2}\frac{1}{r_0},$$

lo cual implica

$$|\,\mathbf{p} - \mathbf{q}\,| \geq |\ell\gamma'(0)| - \frac{\ell^2}{2r_0} = \ell\left(1 - \frac{\ell}{2r_0}\right).$$

Notemos que eligiendo $\ell_{\mathcal{M}}(\mathbf{p}, \mathbf{q}) < 2r_0\varepsilon$ obtenemos (3.50). Sin embargo, nosotros necesitamos encontrar cotas que sean función de $|\,\mathbf{p} - \mathbf{q}\,|$. Para ello, consideremos la función

$$\phi(\mathbf{p}, \mathbf{q}) = \frac{\ell_{\mathcal{M}}(\mathbf{p}, \mathbf{q})}{|\,\mathbf{p} - \mathbf{q}\,|} \quad , \mathbf{p} \neq \mathbf{q},$$

definida sobre $\mathcal{M}\times\mathcal{M}$, donde ponemos $\phi(\mathbf{p}, \mathbf{p}) = 1$. Luego, para todo pares de puntos $\mathbf{p}, \mathbf{q}$ con $\ell_{\mathcal{M}}(\mathbf{p}, \mathbf{q}) < 2r_0\varepsilon$ vale que $\phi(\mathbf{p}, \mathbf{q}) < (1-\varepsilon)^{-1}$. Por otro lado, sobre el conjunto compacto $\mathcal{M}\times\mathcal{M}\setminus\{(\mathbf{p}, \mathbf{q}) : \ell_{\mathcal{M}}(\mathbf{p}, \mathbf{q}) > r_0\varepsilon\}$ la función $\phi(\cdot, \cdot)$ es continua y por lo tanto realiza su máximo m_ϕ. Sea $M_\phi = \max\{(1 - \varepsilon)^{-1}, m_\phi\}$. Eligiendo $\delta = 2r_0M_\phi^{-1}\varepsilon$ tenemos que $|\,\mathbf{p} - \mathbf{q}\,| < \delta$ implica $\ell_{\mathcal{M}}(\mathbf{p}, \mathbf{q}) < 2r_0\varepsilon$ y por lo tanto obtenemos (3.50). □

3.5.2. Teorema principal sobre variedades

Teorema 3. *Sea $\mathbb{X}_n$ una muestra i.i.d de tamaño n distribuida a partir de una densidad $f : \mathcal{M} \mapsto \mathbb{R}_{\geq 0}$, donde*

- *$\mathcal{M} \subset \mathbb{R}^D$ es una variedad de dimensión d, con $d < D$, que se puede escribir como $\mathcal{M} = \varphi(C)$, siendo $\varphi : C \mapsto \mathbb{R}^D$ una transformación isométrica y $C \subset \mathbb{R}^d$ un conjunto convexo, compacto y tal que $\bar{C^o} = C$,*
- *$f : \mathcal{M} \mapsto \mathbb{R}_{\geq 0}$ es una función continua con $f_{min} = \min_{\mathbf{x}\in\mathcal{M}} f(\mathbf{x}) > 0$*

Luego, dados $\mathbf{p}, \mathbf{q} \in \mathcal{M}$ se tiene

$$\lim_{n\to\infty} n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) = \mu_{\alpha,d} \inf_{\Gamma\subset\mathcal{M}} \int_\Gamma \frac{1}{f^\beta} \quad \textit{casi seguramente,} \tag{3.51}$$

donde $\mu_{\alpha,d}$ es una constante que depende del parámetro α y de la dimensión d de $\mathcal{M}$; y la minimización se realiza sobre todas las curvas continuas y rectificables Γ contenidas en la variedad $\mathcal{M}$ y que conectan

$\mathbf{p}$ *con* $\mathbf{q}$. *Más aún, si existe una única curva* $\hat{\Gamma} \subset \mathcal{M}$ *que conecta* $\mathbf{p}$ *con* $\mathbf{q}$ *y tal que*

$$\int_{\hat{\Gamma}} \frac{1}{f^\beta} = \inf_{\Gamma \subset \mathcal{M}} \int_\Gamma \frac{1}{f^\beta}, \tag{3.52}$$

entonces la sucesión de curvas Γ_n *que realizan el camino óptimo convergen uniformemente a* $\hat{\Gamma}$.

Demostración. Dados $\mathbb{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, consideremos $\mathbb{Z}_n = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\}$ tal que $\mathbf{x}_i = \varphi(\mathbf{z}_i)$ para todo $i = 1, 2, \dots, n$. A partir de un cambio de variables es fácil observar que $\mathbb{Z}_n$ es una muestra i.i.d con densidad

$$g(\mathbf{x}) = f(\varphi(\mathbf{x}))\sqrt{\det\left(J_\varphi(\mathbf{x})^t J_\varphi(\mathbf{x})\right)} = f(\varphi(\mathbf{x})).$$

Dado que φ es una isometría, se tiene que dado cualquier camino de puntos $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K \in \mathcal{M}$ vale

$$\sum_{i=1}^{K-1} |\mathbf{y}_{i+1} - \mathbf{y}_i|^\alpha \leq \sum_{i=1}^{K-1} \ell_{\mathcal{M}}\left(\mathbf{y}_{i+1}, \mathbf{y}_i\right)^\alpha = \sum_{i=1}^{K-1} |\varphi^{-1}(\mathbf{y}_{i+1}) - \varphi^{-1}(\mathbf{y}_i)|^\alpha,$$

de donde se desprende que $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \leq \mathcal{D}_{\mathbb{Z}_n}\left(\varphi^{-1}(\mathbf{p}), \varphi^{-1}(\mathbf{q})\right)$. Por otro lado, notemos que dado que φ es una isometría vale

$$J = \mu \inf_{\Gamma \subset \mathcal{M}} \int_\Gamma \frac{1}{f^\beta} = \mu \inf_{\sigma \subset C} \int_\sigma \frac{1}{g^\beta},$$

donde el segundo ínfimo se efectúa sobre todas las curvas $\sigma \subset C$ que conectan los puntos $\varphi^{-1}(\mathbf{p})$ y $\varphi^{-1}(\mathbf{q})$. Luego se tiene que

$$\lim_{n\to\infty} n^\beta \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \leq \lim_{n\to\infty} n^\beta \mathcal{D}_{\mathbb{Z}_n}(\varphi^{-1}(\mathbf{p}), \varphi^{-1}(\mathbf{q})) = J, \quad \text{casi seguramente.}$$

Por otro lado, dado $\varepsilon > 0$, sea $\delta > 0$ como en el Lema 6. Para cada $n \in \mathbb{N}$ sea $(\mathbf{y}_1^*, \mathbf{y}_2^*, \dots, \mathbf{y}_{K_n}^*)$ el camino que realiza la distancia de Fermat entre los puntos $\mathbf{p}$ y $\mathbf{q}$. Sea el evento $E_n = \{\max_{1\leq j\leq K_n - 1} |\mathbf{y}_{j+1}^* - \mathbf{y}_j^*| < \delta\}$. A partir del Lema 5 tenemos que E_n sucede con probabilidad exponencialmente chica en n. A partir del Lema 6 deducimos que

$$\begin{aligned}
\mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) &= \sum_{i=1}^{K_n - 1} |\mathbf{y}_{i+1}^* - \mathbf{y}_i^*|^\alpha \\
&\geq (1-\varepsilon)^\alpha \sum_{i=1}^{K_n - 1} \ell_{\mathcal{M}}\left(\mathbf{y}_i^*, \mathbf{y}_{i+1}^*\right)^\alpha \\
&= (1-\varepsilon)^\alpha \sum_{i=1}^{K_n - 1} |\varphi^{-1}\left(\mathbf{y}_{i+1}^*\right) - \varphi^{-1}\left(\mathbf{y}_i^*\right)|^\alpha \\
&\geq (1-\varepsilon)^\alpha \mathcal{D}_{\mathbb{Z}_n}\left(\varphi^{-1}(\mathbf{p}), \varphi^{-1}(\mathbf{q})\right) \quad \text{en } E_n.
\end{aligned}$$

Dado que $\mathbb{P}(E_n)$ es sumable (Lema 5), tenemos que

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} E_n\right) = 1,$$

y por lo tanto

$$\lim_{n\to\infty} \mathcal{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q}) \geq (1-\varepsilon)^\alpha \lim_{n\to\infty} \mathcal{D}_{\mathbb{Z}_n}\left(\varphi^{-1}(\mathbf{p}), \varphi^{-1}(\mathbf{q})\right) = (1-\varepsilon)^\alpha J.$$

Como esto vale para todo $\varepsilon > 0$, concluimos (3.51). La convergencia de las curvas se deduce de los mismos argumentos de compacidad esbozados anteriormente. □

3.6. Restricción a k vecinos más cercanos

Veamos ahora que podemos restringir la búsqueda del $\mathbb{X}_n$-camino que realiza $\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q})$ a caminos formados por puntos que sean k-vecinos más cercanos. Sea nuevamente $\mathbb{X}_n$ una muestra i.i.d con densidad f soportada en una variedad compacta $\mathcal{M}$. Dado $k \geq 1$ y un punto $\mathbf{x} \in \mathbb{X}_n$, el k-vecino más cercano a $\mathbf{x}$, el cual notamos por $\mathbf{x}^{(k)}$, queda definido como

$$\mathbf{x}^{(1)} = \operatorname*{argmín}_{\mathbf{y}\in\mathbb{X}_n\setminus\{\mathbf{x}\}} |\mathbf{y}-\mathbf{x}| \quad \text{si } k=1, \qquad \mathbf{x}^{(k)} = \operatorname*{argmín}_{\mathbf{y}\in\mathbb{X}_n\setminus\{\mathbf{x},\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(k-1)}\}} |\mathbf{y}-\mathbf{x}| \quad \text{si } k>1.$$

Sea $\mathcal{N}_k(\mathbf{x}) = \{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(k)}\}$ el conjunto de k-vecinos más cercanos del punto $\mathbf{x}$. Dados $\mathbf{p},\mathbf{q} \in \mathbb{X}_n$, un parámetro $\alpha \geq 1$ y $k \in \mathbb{N}$, definimos el *estimador de la distancia de Fermat restringido* como

$$\hat{\mathcal{D}}^k_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \min_{\substack{(\mathbf{y}_1,\ldots,\mathbf{y}_K)\in\mathbb{X}_n^K,\\ \mathbf{y}_1=\mathbf{p},\mathbf{y}_K=\mathbf{q}\\ \mathbf{y}_{i+1}\in\mathcal{N}_k(\mathbf{x}_i)}} \sum_{i=1}^{K-1} |\mathbf{y}_{i+1}-\mathbf{y}_i|^\alpha. \tag{3.53}$$

La siguiente proposición establece que con probabilidad arbitrariamente grande el camino óptimo (y por lo tanto, el estimador mismo) no se ve modificado cuando restringimos la búsqueda a los vecinos más cercanos.

Proposición 6. *Dado $\varepsilon > 0$, existen constantes positivas c_{16}, c_{17} tales que dado $k_0 > c_{16}\log(n/\varepsilon) + c_{17}$ se tiene*

$$\hat{\mathcal{D}}^{k_0}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \quad \textit{con probabilidad al menos } 1-\varepsilon. \tag{3.54}$$

Más precisamente, el $\mathbb{X}_n$-camino minimizante $\mathbf{y}^*_1,\ldots,\mathbf{y}^*_{K_n}$ satisface $\mathbf{y}^*_{i+1} \in \mathcal{N}_{k_0}(\mathbf{y}^*_i)$ para todo $i = 1,\ldots,K_n-1$ con probabilidad al menos $1-\varepsilon$.

Demostración. Al igual que antes, basta con probar el resultado cuando $\mathbb{X}_n$ es un proceso puntual de Poisson sobre C con intensidad $nf(\mathbf{x})$. Fijado $n \in \mathbb{N}$, sea $\mathbf{y}^*_1,\mathbf{y}^*_2,\ldots,\mathbf{y}^*_{K_n}$ el camino óptimo. Definimos

$$knn_{max} = \min\left\{k \in \mathbb{N} : \mathbf{y}^*_{i+1} \in \mathcal{N}_k(\mathbf{y}^*_i) \text{ para todo } i < K_n\right\}. \tag{3.55}$$

Notemos que el camino óptimo no se va a ver modificado cuando restringimos la búsqueda a los k_0 vecinos más cercanos si y sólo si se tiene $knn_{max} \leq k_0$. Dado de que no hay ninguna partícula en el $\mathbb{X}_n$-camino minimizante entre $\mathbf{y}^*_i$ y $\mathbf{y}^*_{i+1}$, si definimos

$$B_i = \left\{\mathbf{x} : |\mathbf{y}^*_i - \mathbf{x}|^\alpha + |\mathbf{y}^*_{i+1}-\mathbf{x}|^\alpha < |\mathbf{y}^*_{i+1}-\mathbf{y}^*_i|^\alpha\right\}, \quad 1 \leq i < K_n,$$

entonces tenemos que $B_i \cap \mathbb{X}_n = \emptyset$. Sea $r_i = |\mathbf{y}^*_{i+1}-\mathbf{y}^*_i|$ aleatorio. Es fácil ver que $B_i \cap B(\mathbf{y}^*_i, r_i)$ tiene interior no vacío y que existe una constante determinística $\delta > 0$, tal que $B_{\delta r_i}(\mathbf{x}) \subset B_i$ para $\mathbf{x} = (\mathbf{y}^*_{i+1}-\mathbf{y}^*_i)/2$. Dado $k \in \mathbb{N}$, definimos el evento

$$\begin{aligned} A_k = \Big\{&\exists \text{ bola } B_r \subset C \text{ de radio } r \text{ con al menos } k \text{ partículas en su interior}\\ &\text{tal que existe otra bola } B_{\delta r} \text{ de radio } \delta r \text{ con } B_{\delta r} \subset B_r \text{ y}\\ &\text{que no contiene ninguna particula en su interior}\Big\}. \end{aligned}$$

Por lo tanto, el hecho de que algún punto $\mathbf{y}^*_{i+1}$ sea exactamente el k vecino más cercano al punto $\mathbf{y}^*_i$ implica A_k, de manera tal que

$$\left\{knn_{max} \geq k_0\right\} \subset \bigcup_{k=k_0}^{\infty} A_k. \tag{3.56}$$

Definimos

$$r_{min} = \frac{1}{3}\left(\frac{k}{2f_{max}n}\right)^{1/d} \quad , \quad r_{max} = 2\sqrt{d}\left(\frac{2k}{f_{min}n}\right)^{1/d}, \tag{3.57}$$

donde claramente se tiene que $r_{min} < r_{max}$. Luego tenemos que

$$\mathbb{P}(A_k) = \mathbb{P}(A_k, r < r_{min}) + \mathbb{P}(A_k, r > r_{max}) + \mathbb{P}(A_k, r \in [r_{min}, r_{max}]). \tag{3.58}$$

Veamos como podemos acotar cada una de las probabilidades involucradas en (3.58).

$$\begin{aligned}\mathbb{P}(A_k, r < r_{min}) &\leq \mathbb{P}\left(\exists \text{ bola de radio } r_{min} \text{ con al menos } k \text{ partículas}\right) \\ &\leq \mathbb{P}\left(\exists \text{ cubo de lado } 2r_{min} \text{ con al menos } k \text{ partículas}\right).\end{aligned}$$

Consideremos sobre $\mathbb{R}^d$ todos los cubos de lado $3r_{min}$ con vértices en $r_{min}\mathbb{Z}^d$. Notemos que la cantidad de estos cubos que tienen intersección no vacía con C está acotado por $\eta_C^1 n/k$, donde η_C^1 es alguna constante positiva que depende de la geometría de C. Por otro lado, cualquier cubo de lado $2r_{min}$ está estrictamente contenido en uno de los cubos de la red. Dado que la cantidad de partículas en cualquiera de los cubos de la red sigue una distribución Poisson con parámetro menor o igual a $3^d r_{min}^d f_{max} n = k/2$, usando cotas de Chernoff tal como hicimos en (3.23) tenemos que

$$\mathbb{P}(A_k, r < r_{min}) \leq \eta_C^1 \frac{n}{k}\mathbb{P}\left(V_1 \geq k\right) \leq \eta_C^1 \frac{n}{k} e^{-\theta_1 k} \tag{3.59}$$

donde $V_1 \sim \text{Poiss}(k/2)$ y θ_1 es alguna constante numérica positiva. Por otro lado

$$\begin{aligned}\mathbb{P}\left(A_k, r > r_{max}\right) &\leq \mathbb{P}\left(\exists \text{ bola de radio } r_{max} \text{ con } k \text{ partículas}\right) \\ &\leq \mathbb{P}\left(\exists \text{ cubo de lado } r_{max}/\sqrt{d} \text{ con a lo sumo } k \text{ partículas}\right).\end{aligned}$$

Ahora consideremos la familia de cubos de lado $r_{max}/(2\sqrt{d})$ cuyos vértices están contenidos en $(r_{min}/(2\sqrt{d}))\mathbb{Z}^d$. Nuevamente, existen a lo sumo $\eta_C^2 n/k$ de estos cubos que tienen intersección no vacía con C, donde η_C^2 es alguna constante positiva. A su vez, es claro que la existencia de un cubo de lado $r_{max}/\sqrt{d}$ con a lo sumo k partículas asegura que al existe uno de los cubos de la familia con a lo sumo k partículas. La cantidad de partículas en cada cubo fijo de la red sigue una distribución de Poisson con intensidad mayor o igual a $r_{max}^d f_{min} n/(2^d d^{d/2}) = 2k$. Luego

$$\mathbb{P}(A_k, r > r_{max}) \leq \eta_C^2 \frac{n}{k}\mathbb{P}\left(V \leq k\right) \leq \eta_C^2 \frac{n}{k} e^{-\theta_2 k}, \tag{3.60}$$

donde $V_2 \sim \text{Poiss}(2k)$ y θ_2 es alguna constante numérica positiva. Por último, notemos que

$$\begin{aligned}\mathbb{P}\left(A_k, r \in [r_{min}, r_{max}]\right) &\leq \mathbb{P}\left(\exists \text{ bola de radio } \delta r_{min} \text{ sin particulas en su interior}\right) \\ &\leq \mathbb{P}\left(\exists \text{ cubo de lado } \delta r_{min}/\sqrt{d} \text{ sin particulas en su interior}\right).\end{aligned}$$

Procedemos igual que antes, donde ahora consideramos la grilla $(\delta r_{min}/2\sqrt{d})\mathbb{Z}^d$. Existen a lo sumo $\eta_C^3 n/k$ de estos cubos con con intersección no vacía con C, con η_C^3 alguna constante positiva, y la cantidad de partículas dentro de cada uno de estos cubos sigue una distribución de Poisson con intensidad menor o igual que $r_{min}^d f_{max} n/(2^d d^{d/2}) = k/(2^{d+1} 3^d d^{d/2})$. Luego

$$\mathbb{P}\left(A_k, r \in [r_{min}, r_{max}]\right) \leq \eta_C^3 \frac{n}{k}\mathbb{P}\left(V_3 = 0\right) \leq \eta_C^3 \frac{n}{k} e^{-\theta_3 k}, \tag{3.61}$$

donde $V_3 \sim \text{Poiss}(k/(2^{d+1} 3^d d^{d/2}))$ y $\theta_3 > 0$ es alguna constante numérica. Finalmente obtenemos que

$$\mathbb{P}(A_k) \leq \eta_C \frac{n}{k} e^{-\theta k}, \tag{3.62}$$

donde $\theta = \text{mín}\{\theta_1, \theta_2, \theta_3\}$ y $\eta_C = \eta_C^1 + \eta_C^2 + \eta_C^3$. Volviendo a (3.56) tenemos

$$\mathbb{P}\left(knn_{max} \geq k_0\right) \leq \sum_{k=k_0}^{\infty} \eta_C \frac{n}{k} e^{-\theta k} \leq \eta_C \frac{n}{k_0} (1 - e^{-\theta})^{-1} e^{-\theta k_0} < \eta_C n (1 - e^{-\theta})^{-1} e^{-\theta k_0}. \tag{3.63}$$

Para asegurar que $\mathbb{P}\left(knn_{max} \geq k_0\right) < \varepsilon$ basta con pedir

$$k < \frac{1}{\theta} \log\left(\frac{\eta_C}{1 - e^{-\theta}} \frac{n}{\varepsilon}\right),$$

de donde concluimos la demostración de la proposición. □

Conclusiones

En la presente tesis fue introducida la *distancia de Fermat* junto con un estimador consistente de la misma. La motivación de su definición y estudio proviene de sus aplicaciones en análisis de datos, por ejemplo, en un problema de Machine Learning o estadística donde entender la estructura intrínseca de los datos es la clave para poder extraer información valiosa de los mismos.

A lo largo del primer capítulo vimos cuales son algunas de las técnicas más conocidas cuando se desea reducir la dimensión de los datos. Dado que los datos típicamente viven en espacios de dimensión muy grande, estas representaciones en espacios de dimensión menor permiten definir métricas que contemplan la estructura de los datos. Entre ellas destacamos *Isomap*, un algoritmo que estima la longitud de las geodésicas sobre la variedad en la cual están soportados los puntos.

En el segundo capítulo estudiamos cómo el estimador de la distancia de Fermat logra captar la densidad de los datos y define una métrica mucho más útil cuando se desea efectuar una tarea de clustering. De esta manera, dos puntos van a estar cerca si existe un camino que los conecte que pase por regiones donde la densidad es alta. Vimos como la performance del algoritmo K-medoids mejora cuando se usa el estimador de la distancia de Fermat como input en comparación con la distancia euclídea y con los estimadores provistos por *Isomap* y *C-Isomap*. En cuanto a este aspecto, resta realizar más experimentos evaluando la distancia de Fermat con datos reales.

Por otro lado, la siguiente convergencia a un objeto macroscópico no trivial,

$$n^{\beta}\mathcal{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \xrightarrow{n\to\infty} \mu \inf_{\Gamma} \int_{\Gamma} \frac{1}{f^{\beta}} \quad \text{casi seguramente,}$$

nos invita a pensar un muchas otras aplicaciones. Por ejemplo, la constante normalizadora n^{β} depende únicamente del parámetro α y de la dimensión intrínseca de los datos d. Por lo tanto, es posible definir un estimador para d. A su vez, la convergencia uniforme de los los caminos óptimos a la curva que realiza el ínfimo macroscópico nos invita a pensar que puntos consecutivos del camino óptimo permiten realizar una transición suave entre los puntos $\mathbf{p}$ y $\mathbf{q}$.

Más allá de su aplicación práctica, el estudio de $\mathcal{D}_{\mathbb{X}_n}(\cdot,\cdot)$ dentro de la teoría de percolación euclídea de primera pasada permitió abordar interrogantes que no habían sido planteados hasta el momento, de los cuales en muchos casos pudimos dar una respuesta. En el tercer capítulo tratamos el comportamiento de las geodésicas cuando se trabaja con una muestra proveniente de un proceso puntual de Poisson no homogéneo. Vimos que macroscópicamente las geodésicas convergen a una curva que queda caracterizada por la intensidad del proceso. También vimos que la longitud de la curva óptima es acotada.

Bibliografía

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory, ICDT 2001*. Springer Berlin Heidelberg, 2001.

Barnard, J. M. and Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *Journal of Chemical Information and Computer Sciences*, 32(6):644–649, 1992.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Bernstein, M., Silva, V. De, Langford, J. C., and Tenenbaum, J. B. Graph approximations to geodesics on embedded manifolds. Technical report, 2000.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Borg, I. and Groenen, P. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 40(3):277–280, 2003.

Cormen, T. H. *Introduction to algorithms*. MIT press, 2009.

de Silva, V. and Tenenbaum, J. B. Global versus local methods in nonlinear dimensionality reduction. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, pp. 721–728. MIT Press, 2002.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Howard, C. D. and Newman, C. M. Euclidean models of first-passage percolation. *Probability Theory and Related Fields*, 108(2):153–170, 1997.

Howard, C. Douglas and Newman, Charles M. Special invited paper: Geodesics and spanning trees for euclidean first passage percolaton. *Ann. Probab.*, 29(2):577–623, 04 2001. doi: 10.1214/aop/1008956686. URL `https://doi.org/10.1214/aop/1008956686`.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec 1985.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

Kallenberg, O. *Foundations of Modern Probability*. Springer, second edition, 2002.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Lawson, D. J. and Falush, D. Similarity matrices and clustering algorithms for population identifcation using genetic data. In *Annual Review of Human Genomics*, number 13, pp. 337–361, 2012.

LeCun, Y., Huang, F. J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pp. II–104. IEEE, 2004.

Matoušek, Jiří. *Lectures on discrete geometry*, volume 212. Springer New York, 2002.

Meilă, M. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98 (5):873 – 895, 2007.

Moller, J. and Waagepetersen, R. P. *Statistical inference and simulation for spatial point processes*. CRC Press, 2003.

Morse, M. D. and Patel, J. M. An efficient and accurate method for evaluating time series similarity. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pp. 569–580. ACM, 2007.

Myers, S. B. Arcs and geodesics in metric spaces. *Transactions of the American Mathematical Society*, 57(2):217–227, 1945.

Powers, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

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

van der Maaten, L. J. P. and Hinton, G. E. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, pp. 2579–2605, 2008.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct): 2837–2854, 2010.

Wang, F., Tan, C., Li, P., and König, A. C. Efficient document clustering via online nonnegative matrix factorizations. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 908–919. SIAM, 2011.