Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Matemáticas

Aprendizaje justo basado en modelos de mezcla

Autora: Inés Velasco Barrero
Tutor: Eustasio del Barrio Tellado
Curso: 2024-2025

Resumen

Este TFG está motivado por un algoritmo propuesto por H. Inouzhe, cuyo objetivo es determinar cotas para la distancia en variación total y estimar la parte común entre dos distribuciones, con vistas a construir algoritmos de clasificación más justos. Se estudian las herramientas teóricas necesarias para su desarrollo, incluyendo formulaciones equivalentes del modelo de mezcla de probabilidades, propiedades topológicas de los conjuntos de recortes y el análisis de métricas como las de Wasserstein y las basadas en la discrepancia media promedio.

Abstract

This work is motivated by an algorithm proposed by H. Inouzhe, aimed at determining bounds for the total variation distance and estimating the common part between two distributions, with a view toward constructing fairer classification algorithms. The necessary theoretical tools for its development are studied, including equivalent formulations of the probabilistic mixture model, topological properties of trimming sets, and the analysis of metrics such as the Wasserstein distance and those based on the maximum mean discrepancy.

Índice general

Introducción

El objetivo principal en Aprendizaje Automático es la obtención de reglas para predecir una variable de interés, Y, denominada de forma genérica como "etiqueta", a partir de otras variables más fácilmente medibles, X, a las que se suele llamar "atributos". Desde un punto de vista formal, sea $\mathcal{X}$ el espacio de los atributos e $\mathcal{Y}$ el espacio de las etiquetas. Se quiere determinar una función $h : \mathcal{X} \to \mathcal{Y}$ tal que $\hat{Y} = h(X)$ sea una predicción de la variable respuesta Y, para cada $X \in \mathcal{X}$.

Un caso particular es el de la clasificación binaria, donde el conjunto de posibles etiquetas es $\mathcal{Y} = \{-1, 1\}$. Se trata de asignar a cada individuo de la población una de estas dos clases, a partir de ciertas variables aleatorias conocidas (los atributos). Para ello, se busca la regla de decisión $\hat{Y} := h(x)$ que acierte la clase del individuo las máximas veces posibles, es decir, que minimice la probabilidad de fallo $P(\hat{Y} \neq Y)$.

Muchos de los algoritmos de predicción aprenden de un conjunto de datos de los que se conocen tanto los atributos como la etiqueta:

$$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}, \quad i \in \{1, ..., n\}.$$

A partir de esta información (datos de entrenamiento), el modelo ajusta una regla que le permite predecir la clase $\{-1, 1\}$ de nuevos individuos, para los cuales solo se observan los atributos x, pero no se conoce la etiqueta y. Esto es lo que se denomina Aprendizaje Supervisado.

El desarrollo de algoritmos de clasificación es uno de los problemas más actuales. Se usan en medicina para diagnosticar enfermedades a partir de datos de otros pacientes anteriores, en ciberseguridad para detectar correos electrónicos maliciosos, en el sector automovilístico para hacer coches más seguros a partir del reconocimiento de imágenes, en finanzas para evaluar el riesgo de que una persona no devuelva un préstamo bancario, etc. En un futuro, prácticamente se tomarán todas las decisiones a partir de algoritmos. Por esta razón, surgen nuevos conflictos, principalmente, relacionados con la ética.

Por ejemplo, se considera un algoritmo para predecir si un alumno de bachillerato estudiará una carrera STEM o no. Si ahora hay menos mujeres que hombres con puestos de trabajo STEM, y el algoritmo tiene como conjunto de entrenamiento datos actuales, podría relacionar a las mujeres con carreras que no sean de ciencias. Sería un algoritmo injusto, en el que la variable sexo tendría bastante peso para determinar la carrera de futuros estudiantes. Para que el algoritmo no carezca de ética, el sexo de un estudiante no debe influir en la regla de decisión. Esto es solo un ejemplo ilustrativo de un problema real en inteligencia artificial: detectar, corregir y prevenir sesgos causados por el uso de algoritmos.

El aprendizaje "justo" se plantea en relación con situaciones en las que, además de los atributos admisibles, hay otras variables involucradas, S, cuyo efecto sobre la predicción se desea limitar o eliminar. A estas variables se las conoce como "atributos protegidos".

Para limitar la influencia de los atributos protegidos en la clasificación de los individuos no es suficiente con determinar una regla de clasificación que no tenga en cuenta esas variables, ya que probablemente los atributos protegidos S no sean independientes de otros atributos X.

Formalizando el problema, se denota por $\mathcal{S}$ al espacio de los atributos protegidos. Si se tiene un atributo protegido binario, entonces $\mathcal{S} = \{0, 1\}$. La ley de probabilidad a partir de la cual están generados los datos del grupo $S = 0$ es posiblemente distinta a la ley de los datos del grupo $S = 1$. Se denota $P_1 := \mathcal{L}(X|S = 0)$ y $P_2 := \mathcal{L}(X|S = 1)$. En el caso general, un atributo protegido va a dividir a la población en varios subgrupos, y la ley de probabilidad que genera los datos de cada subgrupo será, posiblemente, distinta para cada uno de ellos.

El cambio de marco de trabajo debido a la inclusión de atributos protegidos en el problema produce cierto conflicto de intereses en el problema de Aprendizaje Automático. Por un lado, se busca garantizar que las reglas obtenidas produzcan resultados similares para individuos con características similares. Por otro lado, las decisiones no deberían, globalmente, estar influenciadas por los atributos que se tratan de proteger. Frecuentemente, el aprendizaje involucra casos con diferentes atributos y también con diferentes atributos protegidos. En esta situación, el entrenamiento de reglas resulta problemático.

Un primer paso necesario para entrenar reglas adecuadas sería la búsqueda de subgrupos comparables entre las poblaciones con diferente atributo protegido. Así, se podrán entrenar reglas de clasificación sobre datos de estos subgrupos, que serán éticas. Con esta idea en mente, se busca escribir la ley de probabilidad de cada grupo con diferente atributo protegido como la mezcla de una probabilidad común a todos ellos y de otra diferente para cada grupo:

$$P_i = (1 - \alpha)P_0 + \alpha R_i, \quad i = \{1, 2, ..., n\},\ \alpha \in [0, 1]. \tag{1}$$

Además, el nivel al que dos poblaciones comparten una parte común se puede describir en términos de la distancia en variación total. Cuanto menor sea la distancia en variación total entre dos probabilidades P_1 y P_2, más peso tendrá la probabilidad común P_0: las distribuciones de los datos de cada grupo serán similares.

Este TFG está motivado por la propuesta de H. Inouzhe de un algoritmo que persigue obtener cotas probables para la distancia en variación total y una estimación consistente de la parte común entre dos probabilidades. Las cotas se buscan dentro de una partición fija $\mathcal{P} = \{\alpha_0 = 0 < \alpha_1 < \alpha_2 < ... < \alpha_k = 1\}$ del intervalo $[0, 1]$. Este algoritmo consta de tres pasos, descritos en la sección 1.3.2. Se resumen brevemente:

1. Fijar $\alpha_i \in \mathcal{P}$ (se empieza fijando $\alpha_i = \alpha_1$).

2. Calcular los recortes óptimos para una métrica d entre probabilidades:

$$(P_{\alpha_i}, Q_{\alpha_i}) := \underset{\substack{R \in \mathcal{R}_{\alpha_i}(P_1) \\ S \in \mathcal{R}_{\alpha_i}(P_2)}}{\operatorname{argmin}} d(R, S),$$

donde $\mathcal{R}_\alpha(P_j) := \{R \text{ probabillidad } : R(A) \leq \frac{1}{1-\alpha} P_j(A), \quad \forall A \text{ medible}\}$ es el conjunto de recortes de nivel α de P_j, con $j = 1, 2$.

3. Realizar el test de hipótesis nula

$$H_0 : P_{\alpha_i} = Q_{\alpha_i}$$

Si se rechaza la hipótesis nula, se cambia α_i por α_{i+1} y se vuelve a repetir el algoritmo. Si no se rechaza la hipótesis nula, entonces $d_{TV}(P_1, P_2) \leq \alpha_i$ y es la mejor cota que se ha podido encontrar para la distancia en variación total entre P_1 y P_2, dentro de los valores de la partición $\mathcal{P}$. Además, $P_{\alpha_i} = Q_{\alpha_i}$ es la estimación de la parte común de ambas probabilidades.

Este TFG está relacionado con el trabajo que estoy desarrollando a través de la beca de colaboración con el Departamento de Estadística e Investigación Operativa. Se ha dedicado al estudio de las herramientas matemáticas relacionadas con los distintos aspectos involucrados en el algoritmo. No se aborda el desarrollo del algoritmo (ni teórico, ni computacional); este tema queda como trabajo futuro, dentro del proyecto, más grande, de la beca de colaboración.

Por eso, se dedica el capítulo 1 a estudiar formulaciones equivalentes al modelo de mezcla descrito en la ecuación 1. Ese capítulo permitirá reescribir el problema de mezcla en términos de otro problema de distancia entre conjuntos de recorte. La métrica elegida debería hacer que los conjuntos de recorte tuvieran buenas propiedades topológicas, garantizando, por ejemplo, la existencia de minimizadores. Las métricas de Wasserstein, asociadas al transporte óptimo, resultan adecuadas y por ello se les dedica el capítulo 3. Se estudia la existencia de minimizadores en las versiones de Kantorovich y de Monge y las propiedades de la métrica de Wasserstein asociada. Se estudian también las propiedades topológicas de los conjuntos de recortes respecto de esta métrica.

Los problemas computacionales asociados al problema de transporte óptimo clásico se resuelven con el problema entrópico. Por eso, se ha dedicado un capítulo final al estudio de este problema entrópico.

Una vez estimados los recortes más próximos (respecto a la métrica de Wasserstein) se trata de comparar esos dos recortes. Para ello, resulta conveniente emplear otra métrica. Las métricas basadas en la discrepancia media promedio parecen una buena opción y por ello se dedica el capítulo 2 del TFG a ellas. La estimación de la distancia de Wasserstein se ve afectada por la maldición de la dimensionalidad mientras que las métricas MMD parecen ser más resistentes a este problema, esta es una razón por la que estas métricas parecen más adecuadas que la de Wasserstein. En todo caso, estudiar este aspecto de forma más detallada se plantea como trabajo futuro.

Capítulo 1

Modelos de mezcla: recortes y métricas

1.1. Recortes de una probabilidad

Un método estadístico para predecir resultados en función de ciertas variables no puede ser sensible a pequeñas modificaciones de los datos. Ese es el principio motivador de la "Estadística Robusta", desarrollada por P. Huber. En este contexto son interesantes los "recortes" de una probabilidad.

En un espacio finito $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, un recorte de una probabilidad P de nivel $\alpha \in [0, 1)$ es otra probabilidad Q cuya masa está concentrada en los mismos puntos que la primera, pero con pesos ligeramente modificados. Es decir, si $P = \sum_{i=1}^{n} \beta_i \delta_{x_i}$, entonces $Q = \sum_{i=1}^{n} w_i \delta_{x_i}$, con $0 \leq (1-\alpha)w_i \leq \beta_i$, para cada $i \in \{1, ...n\}$.

Un ejemplo práctico en el que se ilustra la utilidad de los recortes es el siguiente: Si $\{X_1, X_2, ..., X_n\}$ son n variables aleatorias independientes de una determinada distribución P, una aproximación a esa ley de probabilidad es la empírica asociada a la muestra:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Para dar menos peso a las observaciones atípicas, se puede recortar P_n de la siguiente forma. Sea $\alpha \in [0, 1)$ tal que $\alpha n \in \mathbb{N}$,

$$\tilde{P}_n = \frac{1}{n(1-\alpha)} \sum_{i=\lfloor \frac{n\alpha}{2} \rfloor + 1}^{\lfloor n(1-\frac{\alpha}{2}) \rfloor} \delta_{x_{(i)}}.$$

Con este recorte, se ordenan las observaciones $\{X_{(1)}, X_{(2)}, ..., X_{(n)}\}$. El parámetro α indica el porcentaje de observaciones que se quieren eliminar (las más alejadas de los valores centrales). El peso total que tenían esas observaciones atípicas se divide entre las $n(1-\alpha)$ observaciones restantes.

En esta sección, se definen los entornos de contaminación de una probabilidad, que fueron propuestos por P. Huber. También se generaliza la definición de probabilidad recortada (que se ha definido únicamente para el caso discreto). Los recortes

de una probabilidad son una propuesta de Cuesta, Matrán y coautores. Más detalles sobre estos aspectos se pueden encontrar en [3]. Ambos conceptos están relacionados con los modelos de mezcla de probabilidades y la parte común entre diferentes leyes de probabilidad. Por esta razón, su estudio es clave si se quieren encontrar subgrupos comparables en las poblaciones con diferente atributo protegido.

Los resultados que se presentan a continuación están en su mayor parte en [3], se han adaptado o completado las demostraciones.

Definición 1.1.1. *Sea $\mathcal{X}$ un espacio con una σ-álgebra $\mathcal{M}$ y sea $\varepsilon \in [0,1)$. Sea P_0 una probabilidad en este espacio. Se define el entorno de contaminación de nivel ε de P_0 como el conjunto de probabilidades:*

$$\nu_\varepsilon(P_0) := \{(1-\varepsilon)P_0 + \varepsilon R : \ R \text{ probabilidad en } (\mathcal{X}, \mathcal{M})\}.$$

La definición es correcta: Si P_1, P_2 son probabilidades en $(\mathcal{X}, \mathcal{M})$, entonces

$$(1-\varepsilon)P_1 + \varepsilon P_2$$

es probabilidad en $(\mathcal{X}, \mathcal{M})$ (es medida positiva por ser suma de medidas positivas y el peso total es 1).

La idea de entorno de contaminación es clara: es el conjunto de probabilidades que se definen a partir de pequeñas modificaciones de P_0. Por eso, tiene sentido pensar que "cuanto más se permita modificar P_0, más grande será el conjunto de probabilidades que forman el entorno de contaminación". Esta idea se formaliza en la siguiente proposición.

Proposición 1.1.2. *Sea P_0 una probabilidad en un espacio medible $(\mathcal{X}, \mathcal{M})$. Los entornos de contaminación, para cada $\varepsilon \in [0,1)$, son crecientes:*

- $\nu_0(P_0) = \{P_0\}$.
- *Sean $\varepsilon_1, \varepsilon_2 \in (0,1)$ tales que $\varepsilon_1 < \varepsilon_2$, entonces $\nu_{\varepsilon_1}(P_0) \subset \nu_{\varepsilon_2}(P_0)$.*

Demostración. Se supone $Q \in \nu_{\varepsilon_1}(P_0)$, es decir, $Q = (1-\varepsilon_1)P_0 + \varepsilon_1 R$ con R probabilidad. Entonces,

$$\begin{aligned} Q &= (1-\varepsilon_2)P_0 + [(\varepsilon_2 - \varepsilon_1)P_0 + \varepsilon_1 R] = (1-\varepsilon_2)P_0 + \varepsilon_2 \left[(1 - \frac{\varepsilon_1}{\varepsilon_2})P_0 + \frac{\varepsilon_1}{\varepsilon_2} R\right] \\ &= (1-\varepsilon_2)P_0 + \varepsilon_2 R', \quad R' = (1 - \frac{\varepsilon_1}{\varepsilon_2})P_0 + \frac{\varepsilon_1}{\varepsilon_2} R \quad \text{es probabilidad en } (\mathcal{X}, \mathcal{M}). \end{aligned}$$

Por lo tanto, $Q \in \nu_{\varepsilon_2}(P_0)$. □

Para la siguiente definición, se utilizan los conceptos de probabilidad absolutamente continua y derivada de Radon-Nikodym, que se recuerdan en el apéndice A.

Definición 1.1.3. *Sean $\alpha \in [0,1)$ y P una probabilidad en un espacio medible $(\mathcal{X}, \mathcal{M})$. El conjunto de recortes de P de nivel α (denotado por $\mathcal{R}_\alpha(P)$) es el conjunto de probabilidades en $(\mathcal{X}, \mathcal{M})$ absolutamente continuas respecto de P cuya derivada de Radon-Nikodym es menor o igual que $\frac{1}{1-\alpha}$ P-casi seguro, es decir,*

$$\mathcal{R}_\alpha(P) := \{Q \ll P : \frac{dQ}{dP} \leq \frac{1}{1-\alpha} \, P\text{-c.s.}\}.$$

Es fácil comprobar que la definición de recorte de una probabilidad coincide con la dada previamente para el caso discreto. Al recortar una probabilidad se modifica ligeramente la probabilidad, sin dar masa a conjuntos que antes no la tenían y sin aumentar excesivamente la probabilidad de cualquier conjunto. Se tiene una caracterización de los recortes de una probabilidad, dada por la siguiente proposición.

Proposición 1.1.4. *Sean P y Q probabilidades en $(\mathcal{X}, \mathcal{M})$ y $\alpha \in [0, 1)$. Son equivalentes:*

1. $Q \in \mathcal{R}_\alpha(P)$.
2. *Para todo conjunto $A \in \mathcal{M}$,*

$$Q(A) \leq \frac{1}{1-\alpha} P(A).$$

Además, si $(\mathcal{X}, \mathcal{M}) = (\mathbb{R}^d, \beta^d)$, las condiciones 1 y 2 son equivalentes también a

3. $$\int f dQ \leq \frac{1}{1-\alpha} \int f dP, \quad \forall f \geq 0 \text{ continua y acotada.}$$

Demostración. Si $Q \in \mathcal{R}_\alpha(P)$, entonces Q es absolutamente continua respecto de P y se verifica:

$$Q(A) = \int_A \frac{dQ}{dP} dP \leq \int_A \frac{1}{1-\alpha} dP = \frac{1}{1-\alpha} P(A).$$

Recíprocamente, si $Q(A) \leq \frac{1}{1-\alpha} P(A)$, para todo $A \in \mathcal{M}$, Q es absolutamente continua respecto de P. Además, si el conjunto $A = \{\frac{dQ}{dP} > \frac{1}{1-\alpha}\} \in \mathcal{M}$ tuviese probabilidad P estrictamente positiva, entonces

$$\frac{1}{1-\alpha} P(A) \geq Q(A) = \int_A \frac{dQ}{dP} dP > \int_A \frac{1}{1-\alpha} dP = \frac{1}{1-\alpha} P(A).$$

Se llega a un absurdo. Por lo tanto, $P(A) = 0$ y $\frac{dQ}{dP} \leq \frac{1}{1-\alpha}$ P-c.s.

Por último, si $(\mathcal{X}, \mathcal{M}) = (\mathbb{R}^d, \beta^d)$, entonces las probabilidades son regulares (ver apéndice B). Se supone que se tiene 3. Sea U un abierto en $\mathcal{X}$, se considera la sucesión de funciones continuas y acotadas definidas de la siguiente forma:

$$f_n(x) = \text{mín}\{1, n d(x, U^c)\}, \quad \forall n \in \mathbb{N}.$$

Como U^c es cerrado, $d(x, U^c) = 0$ si y solo si $x \notin U$. Es fácil ver que $\{f_n\}_{n=1}^\infty$ es una sucesión creciente que converge puntualmente a la función indicadora de U. Por el teorema de la convergencia monótona,

$$\begin{aligned} Q(U) &= \int \mathcal{X}_U dQ = \lim_{n\to\infty} \int f_n dQ = \lim_{n\to\infty} \int f_n \frac{dQ}{dP} dP \\ &\leq \lim_{n\to\infty} \frac{1}{1-\alpha} \int f_n dP = \frac{1}{1-\alpha} \int \mathcal{X}_U dP = \frac{1}{1-\alpha} P(U). \end{aligned}$$

Por la regularidad de las probabilidades en $\mathbb{R}^d$, si $A \in \beta^d$, se tiene que

$$\begin{aligned} Q(A) &= \text{ínf}\{Q(U) : E \subseteq U,\, U \text{ abierto}\} \\ &\leq \frac{1}{1-\alpha}\,\text{ínf}\{P(U) : E \subseteq U,\, U \text{ abierto}\} = \frac{1}{1-\alpha}P(A). \end{aligned}$$

La implicación contraria es fácil usando la regla de la cadena (ver A.0.3). Si $Q \in \mathcal{R}_\alpha(P)$, para cualquier función f positiva, continua y acotada, se tiene que:

$$\int f dQ \leq \int f \frac{dQ}{dP} dP \leq \frac{1}{1-\alpha} \int f dP.$$

□

A continuación, se relaciona el conjunto de recortes con el entorno de contaminación.

Proposición 1.1.5. *Sea $\alpha \in [0,1)$. Sean P, Q dos probabilidades en $(\mathcal{X}, \mathcal{M})$. Entonces,*

$$Q \in \mathcal{R}_\alpha(P) \quad \Leftrightarrow \quad P \in \nu_\alpha(Q).$$

Demostración. Sea $Q \in \mathcal{R}_\alpha(P)$, es decir, Q es una probabilidad tal que

$$(1-\alpha)Q(A) \leq P(A), \quad \forall A \in \mathcal{M}.$$

Se define

$$R := \frac{1}{\alpha}\left(P - (1-\alpha)Q\right).$$

R es medida, por ser suma de medidas, y positiva, por la hipótesis. Además,

$$R(X) = \frac{1}{\alpha}\left(P(X) - (1-\alpha)Q(X)\right) = \frac{1}{\alpha}\left(1 - (1-\alpha)\right) = 1.$$

Entonces R es una probabilidad y se cumple $P = (1-\alpha)Q + \alpha R$. Queda probado que $P \in \nu_\alpha(Q)$.

Recíprocamente, si $P \in \nu_\alpha(Q)$, existe una probabilidad R en $(\mathcal{X}, \mathcal{M})$ tal que

$$P(A) = (1-\alpha)Q(A) + \alpha R(A), \quad \forall A \in \mathcal{M}.$$

Entonces $P(A) \geq (1-\alpha)Q(A)$. Por la caracterización de recortes vista en 1.1.4, $Q \in \mathcal{R}_\alpha(P)$. □

1.2. Distancia en variación total entre dos probabilidades

Primero, se recuerda la definición de distancia en variación total.

Definición 1.2.1. *Sean P_1 y P_2 probabilidades en $(\mathcal{X}, \mathcal{M})$. Se define la distancia en variación total entre P_1 y P_2 como:*

$$d_{TV}(P_1, P_2) := \sup_{A \in \mathcal{M}} |P_1(A) - P_2(A)|.$$

Si P_1 y P_2 son probabilidades en un espacio $(\mathcal{X}, \mathcal{M})$, siempre existe otra medida positiva μ en ese mismo espacio tal que P_1 y P_2 son absolutamente continuas respecto de μ. Por ejemplo, $\mu = P_1 + P_2$.

En el siguiente lema, se da una caracterización de la distancia en variación total en términos de la derivada de Radon-Nikodym de P_1 y P_2 respecto de μ.

Si $(\mathcal{X}, \mathcal{M}) = (\mathbb{R}^d, \beta^d)$ y las probabilidades P_1 y P_2 tienen función de densidad, se trata de un caso particular de lo que se acaba de explicar, tomando μ como la medida de Lebesgue.

Lema 1.2.2. *Sean P_1 y P_2 probabilidades en $(\mathcal{X}, \mathcal{M})$ y μ una medida positiva en ese espacio tal que P_1 y P_2 son absolutamente continuas respecto de μ. Para $i = 1, 2$ se denota por f_i a la derivada de Radon-Nikodym de P_i respecto de μ. Entonces,*

$$d_{TV}(P_1, P_2) = 1 - \int \text{mín}(f_1, f_2) d\mu = \frac{1}{2} \int |f_1 - f_2| d\mu.$$

Demostración. Sea $A_0 = \{x \in \mathcal{X} : f_1(x) > f_2(x)\}$. Para cualquier conjunto $A \in \mathcal{X}$,

$$P_1(A) - P_2(A) = \int_A (f_1 - f_2) d\mu = \int_{A \cap A_0} (f_1 - f_2) d\mu + \int_{A \cap A_0^C} (f_1 - f_2) d\mu.$$

Teniendo en cuenta que $\int_{A \cap A_0^c} (f_1 - f_2) d\mu \leq 0$, se puede acotar esa diferencia de probabilidades superiormente:

$$P_1(A) - P_2(A) \leq \int_{A \cap A_0} (f_1 - f_2) d\mu \leq \int_{A_0} (f_1 - f_2) d\mu = P_1(A_0) - P_2(A_0).$$

Análogamente, se tiene que

$$-(P_1(A_0) - P_2(A_0)) = P_1(A_0^C) - P_2(A_0^C) \leq P_1(A) - P_2(A).$$

Como ambas desigualdades se verifican para cualquier conjunto $A \in \mathcal{X}$, tomando superiores, se deduce que

$$d_{TV}(P_1, P_2) = P_1(A_0) - P_2(A_0) = \int_{A_0} (f_1 - f_2) d\mu = \int (f_1 - f_2)_+ d\mu.$$

Usando la igualdad $(f_1 - f_2)_+ = f_1 - \text{mín}(f_1, f_2)$, se concluye que

$$d_{TV}(P_1, P_2) = 1 - \int \text{mín}(f_1, f_2) d\mu.$$

Por otro lado,

$$1 = \int f_1 d\mu = \int f_2 d\mu \Rightarrow 0 = \int (f_1 - f_2) d\mu = \int_{A_0} (f_1 - f_2) d\mu + \int_{A_0^C} (f_1 - f_2) d\mu.$$

Entonces,

$$\int (f_1 - f_2)_- d\mu = -\int_{A_0^C} (f_1 - f_2) d\mu = \int_{A_0} (f_1 - f_2) d\mu = \int (f_1 - f_2)_+ d\mu.$$

$$\int |f_1 - f_2| d\mu = \int (f_1 - f_2)_+ d\mu + \int (f_1 - f_2)_- d\mu = 2 \int (f_1 - f_2)_+ d\mu.$$

Con esta igualdad, es claro que

$$d_{TV}(P_1, P_2) = \int (f_1 - f_2)_+ d\mu = \frac{1}{2} \int |f_1 - f_2| d\mu.$$

□

Volviendo al problema del aprendizaje justo, si un atributo protegido binario divide a la población en dos subgrupos, cada uno de ellos tendrá una distribución de probabilidad diferente (P_1 y P_2). La distancia en variación total mide cuánto se parecen esas dos probabilidades.

Como ya se ha explicado, para entrenar reglas de clasificación justas, el primer paso es buscar la parte común P_0 entre esas dos probabilidades. Para ello, se busca representar, para cierto $\alpha \in (0, 1)$, P_1 y P_2 de la siguiente forma:

$$P_i = (1 - \alpha) P_0 + \alpha R_i, \quad i = \{1, 2\}.$$

para ciertas probabilidades R_1 y R_2.

En otras palabras, se quiere determinar una probabilidad P_0 tal que P_1 y P_2 pertenezcan al entorno de contaminación de P_0, para cierto nivel $\alpha \in (0, 1)$. O, equivalentemente, encontrar una probabilidad P_0 que pertenezca al conjunto de recortes de P_1 y de P_2, del mismo nivel $\alpha \in (0, 1)$. Además, cuanto más pequeño sea α, más se parecerán las probabilidades P_1 y P_2. Por lo tanto, el objetivo es encontrar el menor $\alpha \in (0, 1)$ que verifique esa condición.

El último resultado de esta sección, que se puede encontrar en [2], relaciona los recortes de una probabilidad con la distancia en variación total. La idea es la siguiente: cuanto menor es la distancia en variación total entre las probabilidades P_1 y P_2, más peso tiene la probabilidad común P_0.

Proposición 1.2.3. *Sean P_1 y P_2 dos probabilidades en $(\mathcal{X}, \mathcal{M})$ y sea $\alpha \in [0, 1)$. Los siguientes enunciados son equivalentes:*

1. *Existe una probabilidad P_0 en $\mathcal{X}$ tal que P_1 y P_2 pertenecen al entorno de contaminación de P_0 de nivel α. Es decir,*

$$P_1 = (1 - \alpha) P_0 + \alpha R_1,$$

$$P_2 = (1 - \alpha) P_0 + \alpha R_2,$$

   con R_1 y R_2 probabilidades en $\mathcal{X}$.

2. *Los conjuntos de recortes de nivel α de P_1 y de P_2 tienen intersección no vacía.*

3. *La distancia en variación total entre P_1 y P_2 es menor o igual que α.*

Demostración. Se ha probado en 1.1.5 que los dos primeros puntos son equivalentes, hay que probar la equivalencia con el tercero.

Si $P_1, P_2 \in \nu_\alpha(P_0)$, entonces $|P_1(A) - P_2(A)| = |\alpha(R_1(A) - R_2(A))| \leq \alpha$, para todo $A \in \mathcal{M}$. Se concluye que $d_{TV}(P_1, P_2) \leq \alpha$.

Para probar el recíproco, hay que probar que existe una probabilidad P_0 tal que $P_0 \in R_\alpha(P_1) \cap R_\alpha(P_2)$. Sea μ una medida en $(\mathcal{X}, \mathcal{M})$ con $P_1 \ll \mu$, $P_2 \ll \mu$. Por ejemplo, $\mu = P_1 + P_2$. Y sea

$$f_i = \frac{dP_i}{d\mu}, \quad i = 1, 2.$$

Por hipótesis (y usando la caracterización de la distancia en variación total dada en el lema 1.2.2),

$$d_{TV}(P_1, P_2) = 1 - \int \text{mín}(f_1, f_2) d\mu \leq \alpha \Leftrightarrow \int \text{mín}(f_1, f_2) d\mu \geq 1 - \alpha.$$

Se considera la función $f_0 := \frac{\text{mín}(f_1, f_2)}{\int \text{mín}(f_1, f_2) d\mu}$ y sea $dP_0 := f_0 d\mu$. Es decir,

$$P_0(A) = \int_A f_0 d\mu, \quad \forall A \in \mathcal{M}.$$

Así definida, P_0 es una medida positiva que cumple que $P_0(X) = \frac{\int \text{mín}(f_1, f_2) d\mu}{\int \text{mín}(f_1, f_2) d\mu} = 1$. Es decir, es una probabilidad. Además, para $i \in \{1, 2\}$:

$$P_0(A) = \frac{1}{\int \text{mín}(f_1, f_2) d\mu} \int_A \text{mín}(f_1, f_2) d\mu \leq \frac{1}{1 - \alpha} \int_A f_i d\mu = \frac{1}{1 - \alpha} P_i(A).$$

Se ha probado la existencia de una probabilidad $P_0 \in \mathcal{R}_\alpha(P_1) \cap \mathcal{R}_\alpha(P_2)$. □

La proposición que se acaba de enunciar es de gran utilidad. Sirve para establecer cotas para la distancia en variación total entre dos probabilidades P_1 y P_2. Si para un nivel $\alpha \in [0, 1)$ se encuentra un elemento común a $\mathcal{R}_\alpha(P_1)$ y $\mathcal{R}_\alpha(P_2)$, entonces la distancia en variación total entre P_1 y P_2 está acotada por α.

1.3. Estimación de la parte común de dos conjuntos de datos

1.3.1. Maldición de la dimensionalidad

La demostración de la proposición 1.2.3 es constructiva, es decir, da un método para definir la parte común P_0 entre dos probabilidades P_1 y P_2. De esta forma, se tiene caracterizada la parte común de dos conjuntos de datos.

Si, por ejemplo, P_1 y P_2 son dos probabilidades en $(\mathbb{R}^d, \beta^d)$ que tienen densidades f_1, f_2, respectivamente, se ha visto en el lema 1.2.2 que

$$d_{TV}(P_1, P_2) = 1 - \int \text{mín}(f_1, f_2) dx.$$

También se ha visto, en 1.2.3, que la densidad de la parte común P_0 es el mínimo de ambas densidades normalizado. Es decir, si se define

$$P_0(A) = \frac{1}{\int \text{mín}(f_1, f_2) dx} \int_A \text{mín}(f_1, f_2) dx, \quad \forall A \in \mathcal{X},$$

entonces $P_0 \in R_\alpha(P_1) \cap R_\alpha(P_2)$, con $\alpha = d_{TV}(P_1, P_2) = 1 - \int \min(f_1, f_2)dx$.

Por lo tanto, estimando las densidades f_1 y f_2, se tiene una estimación de la parte común entre esas dos distribuciones y de su distancia en variación total. Pero esta estrategia no es la adecuada. El problema es que los estimadores de funciones de densidad no son resistentes a lo que se conoce como "maldición de la dimensionalidad".

Al aumentar la dimensión en la que se trabaja, los datos se dispersan tanto que se vuelven difíciles de analizar. Los métodos que funcionan bien en pocas dimensiones (como estimar densidades o distancias) se vuelven ineficaces. En esto consiste la maldición de la dimensionalidad. Si se estima la función de densidad a partir de una muestra de tamaño n, por ejemplo en los histogramas, estimadores kernel, etc., cuando la dimensión es muy alta, $d \gg n$, los datos están muy separados entre sí, no cubren prácticamente nada del espacio. Hay resultados teóricos que demuestran que el mejor estimador posible de una densidad en $\mathbb{R}^d$ comete un error de orden $n^{-a/(d+b)}$ para ciertas constantes $a, b > 0$, lo que demuestra que la precisión de los estimadores se deteriora con d.

Por eso, hay que buscar otras alternativas para encontrar la parte común a dos conjuntos de datos. Los procedimientos que se proponen en este trabajo están basados en distancias entre probabilidades. En concreto, se estudia el problema del transporte óptimo, con algunas variaciones, y la máxima discrepancia en media.

1.3.2. Estimación de cotas para la distancia en variación total

Sea $\mathcal{X} \subset \mathbb{R}^d$ el espacio en el que toman valores los atributos (variables aleatorias fácilmente medibles que se usan para determinar una regla de clasificación). Sea S un atributo protegido binario, es decir, $S \in \mathcal{S} = \{0, 1\}$. Se consideran dos probabilidades en $\mathcal{X}$:

$$P := \mathcal{L}(X|S = 0), \quad Q := \mathcal{L}(X|S = 1).$$

H. Inouzhe propuso un procedimiento, alternativo a la estimación de las funciones de densidad, para estimar una cota $\alpha \in [0, 1]$ de la distancia en variación total entre P y Q, y determinar la parte común P_0 a ambas probabilidades. Se explica a continuación:

Se considera una partición del intervalo $[0, 1]$:

$$\mathcal{P} = \{\alpha_0 = 0 < \alpha_1 < \alpha_2 < ... < \alpha_k = 1\}.$$

Al comienzo, se fija $\alpha = \alpha_1$. El algoritmo consta de tres etapas:

1. Para $\alpha_i \in \mathcal{P}$ fijo, la pregunta es si $d_{TV}(P, Q) \leq \alpha_i$. Se ha probado ya que α es una cota para la distancia en variación total entre P y Q si, y solo si, $\mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q) \neq \emptyset$. Por eso, el siguiente paso consiste en buscar el elemento común a los dos conjuntos de recortes. Esto es un problema de distancia óptima recortada.

2. Se considera una métrica d en una clase de probabilidades en $\mathbb{R}^d$. Se calculan los recortes óptimos, definidos como:

$$(P_{\alpha_i}, Q_{\alpha_i}) := \underset{\substack{R\in\mathcal{R}_{\alpha_i}(P)\\ S\in\mathcal{R}_{\alpha_i}(Q)}}{\operatorname{argmin}} d(R,S).$$

3. El tercer paso consiste en contrastar si $P_{\alpha_i} = Q_{\alpha_i}$. Se hace este contraste con un test de hipótesis nula.

$$H_0 : P_{\alpha_i} = Q_{\alpha_i}$$

Si se rechaza la hipótesis nula, entonces, $d_{TV}(P,Q) > \alpha_i$. Se aumenta el valor de la cota α, cambiando α_i por α_{i+1} y se vuelve a repetir el algoritmo. Si no se rechaza la hipótesis nula, entonces $d_{TV}(P,Q) \leq \alpha_i$ y es la mejor cota que se ha podido encontrar para la distancia en variación total entre P y Q, dentro de los valores de la partición $\mathcal{P}$. Además, $P_{\alpha_i} = Q_{\alpha_i}$ es la parte común de ambas probabilidades.

Para concluir este capítulo, se detallan los pasos 2 y 3 del algoritmo, en los que se centrarán el resto de capítulos del trabajo:

En el paso 2, hay que escoger una distancia adecuada entre probabilidades en $\mathbb{R}^d$. Una opción es trabajar con la distancia de Wasserstein $\mathcal{W}_2$, que se definirá en el capítulo 3, a partir del estudio del problema de transporte óptimo. Si P y Q tienen momentos de orden 2 finitos, su distancia de Wasserstein viene dada por la siguiente expresión:

$$\mathcal{W}_2(P,Q) = \min_{\pi\in\prod(P,Q)} \Big(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 d\pi(x,y) \Big)^{\frac{1}{2}},$$

donde se ha denotado por $\prod(P,Q)$ al conjunto de probabilidades en $\mathbb{R}^d \times \mathbb{R}^d$ con marginales P y Q. En el artículo [1], se garantiza que existe un minimizador de $\mathcal{W}_2(R,S)$, donde $R \in \mathcal{R}_\alpha(P)$ y $S \in \mathcal{R}_\alpha(Q)$. Es decir, existe (P_α, Q_α) probabilidades tales que

$$(P_\alpha, Q_\alpha) := \underset{\substack{R\in\mathcal{R}_{\alpha}(P)\\ S\in\mathcal{R}_{\alpha}(Q)}}{\operatorname{argmin}} \mathcal{W}_2(R,S).$$

Por lo tanto, si existe un recorte P_0 común a P y Q, esa probabilidad minimiza la distancia de Wasserstein entre ambos conjuntos de recortes (el mínimo es 0) y se tiene, entonces, que

$$P = (1-\alpha)P_0 + \alpha R_1,$$

$$Q = (1-\alpha)P_0 + \alpha R_2,$$

con R_1 y R_2 probabilidades.

En la práctica, se calcula la distancia de Wasserstein entre R y S a partir de aproximaciones empíricas. Para ello, se consideran aproximaciones discretas a esas dos probabilidades y se calcula su respectiva distancia de Wasserstein. Formalizando esta idea, sea $n \in \mathbb{N}$, y sea $\{x_1, ...x_n\}$ una muestra de la distribución R. Análogamente, sea $\{y_1, ...y_n\}$ una muestra de la distribución S. Se definen $R_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ y

$S_n = \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j}$. La distancia de Wasserstein entre R_n y S_n viene dada por la siguiente fórmula:

$$\mathcal{W}_2(R_n, S_n) = \text{mín}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - y_j\|^2\pi_{i,j}, \text{ sujeto a } \pi_{i,j} \geq 0;\ \sum_{j=1}^{n}\pi_{i,j} = \frac{1}{n};\ \sum_{i=1}^{n}\pi_{i,j} = \frac{1}{n}.$$

Este procedimiento (aproximar la distancia por su versión empírica) no es válido para la distancia en variación total entre R y S, ya que si, por ejemplo, R y S tienen densidad y no están concentradas en conjuntos disjuntos, entonces $d_{TV}(R, S) < 1$; pero si se generan muestras de cualquier tamaño de ambas probabilidades, no tendrán puntos en común, lo que implica que

$$d_{TV}(R_n, S_n) = 1, \quad \forall n \in \mathbb{N}.$$

Por lo tanto, $\lim_{n\to\infty} d_{TV}(R_n, S_n) = 1 \neq d_{TV}(R, S)$. En este aspecto, la distancia de Wasserstein presenta ventajas con respecto a la distancia en variación total ya que

$$\lim_{n\to\infty} \mathcal{W}_p(R_n, S_n) = \mathcal{W}_p(R, S).$$

Esto se demostrará en el capítulo 3, en concreto, en la sección 3.3.1.

Sin embargo, el cálculo de la distancia de Wasserstein, con la fórmula empírica, tiene un coste computacional muy alto, del orden de $O(n^3)$. Además, al calcularla a partir de una muestra, se ve afectada por la maldición de la dimensionalidad. Por estas razones, se propone modificar la distancia de Wasserstein, contaminándola con la divergencia de Kullback-Leibler (mide cuánto de diferentes son dos probabilidades). Este tema se trata en el capítulo 4, donde se estudia el coste de transporte entrópico $W_{2,\varepsilon}$.

Para cada $\varepsilon > 0$, si P y Q son probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos, se define

$$\mathcal{W}^2_{2,\varepsilon}(P, Q) = \inf_{\pi\in\prod(P,Q)}\left[\int \frac{\|x-y\|^2}{2}d\pi(x, y) + \varepsilon D(\pi|P\otimes Q)\right],$$

donde $D(\pi|P\otimes Q)$ es la divergencia de Kullback-Leibler de π respecto de la medida producto $P\otimes Q$, que también se definirá en el capítulo 4.

El coste de transporte entrópico $\mathcal{W}_{2,\varepsilon}$ se calcula, gracias a una formulación dual, a través de una iteración de punto fijo, que hace que sea menos costoso y que no se vea afectado por la maldición de la dimensionalidad.

En cuanto al paso 3 del método, el test que se usa para contrastar si $P_\alpha = Q_\alpha$ está basado en la máxima discrepancia en media MMD, que se estudia en el capítulo 2.

$$MMD[\mathcal{F}, p, q] := \sup_{f\in\mathcal{F}}(\int f dp - \int f dq) = \sup_{f\in\mathcal{F}}(E_p f - E_q f),$$

donde p y q son probabilidades en $\mathbb{R}^d$ y $\mathcal{F}$ es una clase fija de funciones $f : \mathbb{R}^d \to \mathbb{R}$. Bajo ciertas condiciones, la máxima discrepancia en media entre dos probabilidades es 0 si, y solo si, las probabilidades son iguales; esto se formaliza en la sección 2.3.

Se plantea un test en el que se estima la máxima discrepancia en media entre los recortes P_α y Q_α por una versión muestral, calculable en la práctica. En esa sección también se dan varias opciones de estimadores de esta métrica, que se calculan en la práctica únicamente a partir de productos escalares de los datos de una muestra, lo que resulta computacionalmente muy ventajoso.

Capítulo 2

Métricas de máxima discrepancia media

2.1. Contexto y Motivación

En estadística, es importante representar los datos en un espacio en el que sea fácil trabajar con ellos. Se ilustra esta idea con un ejemplo, en el contexto del Aprendizaje Automático (determinación de reglas para predecir respuestas en función de variables conocidas). En concreto, en la clasificación binaria, la variable respuesta Y, llamada etiqueta, solo puede tomar los valores 1 o -1, que determinan la clase del individuo. El objetivo es predecir Y a partir de otras variables aleatorias X, los atributos.

Para los datos dibujados en la gráfica, se puede encontrar un clasificador lineal (recta) que divide a los individuos de la clase verde de los rojos perfectamente (al menos para el conjunto de entrenamiento).

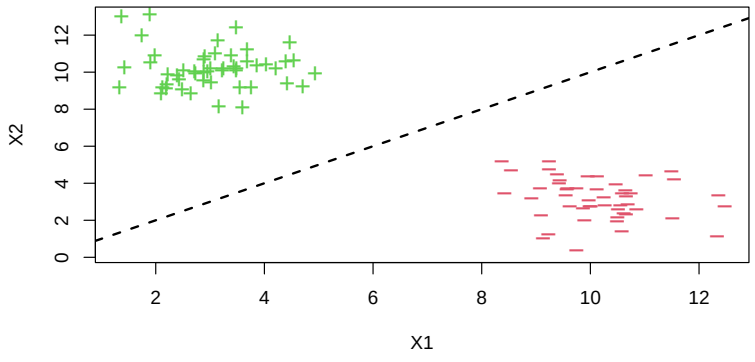

Figura 2.1: Clasificación lineal

En este ejemplo, se han generado 100 observaciones de la siguiente forma. Para $i = 1, ..., 100$:

$$X_i = (1 - Z_i)U_i + Z_i V_i,$$

$$Y_i = 2Z_i - 1,$$

donde

$$\begin{cases} Z_i \sim B(\frac{1}{2}) \quad i.i.d. \\ U_i \sim N((3,10), Id) \quad i.i.d. \\ V_i \sim N((10,3), Id) \quad i.i.d. \end{cases}$$

Es decir, los datos del grupo $Y = 1$ siguen una distribución normal bivariante centrada en el punto $(3, 10)$ con matriz de covarianzas la identidad, mientras que los individuos del grupo $Y = -1$ están generados a partir de una distribución normal con media en $(10, 3)$ y misma matriz de covarianzas.

No siempre es posible encontrar un hiperplano que separe los datos. Por ejemplo, para $i = 1, ..., 100$, se generan observaciones de la siguiente forma:

$$X_i = (1 - Z_i) \begin{bmatrix} U_i \\ 10U_i^2 + E_i \end{bmatrix} + Z_i \begin{bmatrix} V_i \\ 100V_i^2 + 3 + E_i \end{bmatrix},$$

$$Y_i = 2Z_i - 1,$$

donde

$$\begin{cases} Z_i \sim B(\frac{1}{2}) \quad i.i.d. \\ U_i \sim U(-1,1) \quad i.i.d. \\ V_i \sim U(-0{,}3, 0-3) \quad i.i.d. \\ E_i \sim N(0, 0{,}5) \quad i.i.d. \end{cases}$$

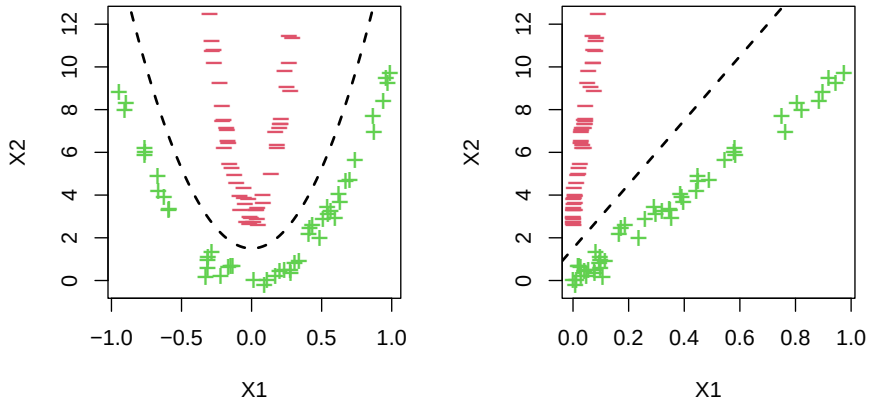

Figura 2.2: Clasificación mediante una regla cuadrática

Se puede ver en el gráfico de la izquierda que los datos no se pueden clasificar con una regla lineal. Quedarían bien clasificados mediante una regla cuadrática (parábola). Sin embargo, al cambiar la representación de los datos, es decir, al representarlos en otro espacio, sí quedan perfectamente divididos por una recta, como se ve en la figura de la derecha. Entonces, en el nuevo espacio, se pueden usar técnicas de clasificación lineal, como Support Vector Machine o el discriminante de Fisher.

Para transformar el primer conjunto de puntos (gráfico de la izquierda) en el segundo (gráfico de la derecha), se ha aplicado a los datos la transformación:

$$\phi : \mathbb{R}^2 \to \mathbb{R}^2, \qquad \phi(x_1, x_2) = (x_1^2, x_2).$$

No siempre es tan fácil encontrar la función que representa los datos en un espacio donde es más fácil operar con ellos. Habrá que probar diferentes transformaciones. Una forma cómoda de trabajar con aplicaciones ϕ que representan los datos en otro espacio es a partir de los núcleos.

En la siguiente sección, se definen los núcleos, y se explica una forma de representar los datos en un espacio de Hilbert, para trabajar únicamente con sus productos internos.

2.2. Espacio de Hilbert reproductor del núcleo

Sea $\mathcal{X}$ el espacio donde toman valores las variables aleatorias. Se ha visto, en el ejemplo anterior, que cambiar la representación de los datos puede simplificar su tratamiento, como en el problema de clasificación. En este capítulo, se estudiarán técnicas para transformar los datos para que tomen valores en un espacio de Hilbert, donde se puede trabajar con un producto interno. Por lo tanto, si $\mathcal{H}$ es un espacio de Hilbert, una aplicación $\phi : \mathcal{X} \to \mathcal{H}$ representa los datos de $\mathcal{X}$ en $\mathcal{H}$. Además, en muchos casos, no es necesario conocer explícitamente la función ϕ, únicamente hace falta conocer los productos internos entre los elementos del espacio $\mathcal{H}$ (que sean imagen por ϕ de algún elemento de $\mathcal{X}$).

Esta técnica se va a aplicar, en particular, a la obtención de una caracterización para la máxima discrepancia en media entre dos probabilidades. Se deducirá una expresión en la que solo intervengan los productos internos de los datos representados en el nuevo espacio (que es de Hilbert).

Con esta motivación, se define el núcleo, una aplicación de $\mathcal{X} \times \mathcal{X}$ en $\mathbb{R}^d$ que recoge únicamente la información de los productos internos. Los resultados expuestos en esta sección están basados en [9] y [18].

Definición 2.2.1. *Se considera $\phi : \mathcal{X} \to \mathcal{H}$ una aplicación con llegada a un espacio de Hilbert. En este contexto, se denomina núcleo del espacio de Hilbert $\mathcal{H}$ a la función*

$$\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$
$$\kappa(x, z) = \langle \phi(x), \phi(z) \rangle.$$

Proposición 2.2.2. *Si $\phi : \mathcal{X} \to \mathcal{H}$ es una aplicación con llegada a un espacio de Hilbert. El núcleo $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ verifica dos propiedades:*

1. *κ es una función simétrica.*
2. *κ es finitamente semidefinida positiva. Esto quiere decir que para cualquier subconjunto finito $\{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}$, la matriz*

$$\mathbf{K} = \begin{bmatrix} \kappa(x_1, x_1) & \kappa(x_1, x_2) & \cdots & \kappa(x_1, x_n) \\ \kappa(x_2, x_1) & \kappa(x_2, x_2) & \cdots & \kappa(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(x_n, x_1) & \kappa(x_n, x_2) & \cdots & \kappa(x_n, x_n) \end{bmatrix}$$

es semidefinida positiva.

Demostración. Por las propiedades del producto interno, es claro que κ es una función simétrica.

En cuanto a la segunda propiedad, sea $\{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}$. Para todo vector $\mathbf{v} \in \mathbb{R}^n$, se tiene:

$$\mathbf{v}^\top \mathbf{K}\mathbf{v} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \kappa(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \langle \phi(x_i), \phi(x_j)\rangle = \left\langle \sum_{i=1}^{n} v_i \phi(x_i), \sum_{j=1}^{n} v_j \phi(x_j) \right\rangle.$$

Entonces,

$$\mathbf{v}^\top \mathbf{K}\mathbf{v} = \left\| \sum_{i=1}^{n} v_i \phi(x_i) \right\|^2 \geq 0.$$

□

Hasta ahora, se ha definido la función núcleo a partir de un espacio de Hilbert $\mathcal{H}$ conocido y una aplicación $\phi : \mathcal{X} \to \mathcal{H}$ que representa los datos en dicho espacio. Pero, como ya se ha enfatizado antes, en muchos casos no es necesario conocer explícitamente el espacio $\mathcal{H}$, solo los valores que toma el núcleo.

Por esta razón, el próximo objetivo es determinar cuándo una función es un núcleo. Es decir, si, dada una función $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, existe un espacio de Hilbert $\mathcal{H}$ y una función $\phi : \mathcal{X} \to \mathcal{H}$ tal que $\kappa(x, z) = \langle \phi(x), \phi(z)\rangle$ para todo $(x, z) \in \mathcal{X} \times \mathcal{X}$. El siguiente teorema da una caracterización de los núcleos: cualquier función $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ simétrica y finitamente semidefinida positiva es un núcleo. Es decir, es una especie de recíproco de la proposición 2.2.2.

Así, se tiene una forma sencilla de trabajar con los datos en un espacio de Hilbert, que no hace falta conocer explícitamente. Se considera una función $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ con las propiedades que se han mencionado (simétrica y finitamente semidefinida positiva). Entonces, ya se conocerían los productos internos de los datos trasladados a cierto espacio de Hilbert.

Teorema 2.2.3 (Aronszajn)**.** *Sea $\mathcal{X}$ un espacio medible. Una función $\kappa : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ es de la forma*

$$\kappa(x, z) = \langle \phi(x), \phi(z)\rangle, \quad \forall x, z \in \mathcal{X},$$

donde $\phi : \mathcal{X} \to \mathcal{H}$ es una aplicación con llegada en un espacio de Hilbert $\mathcal{H}$ y $\langle \cdot, \cdot \rangle$ es el producto interno de $\mathcal{H}$, si y solo si, κ es simétrica y finitamente semidefinida positiva.

Demostración. Se va a construir un espacio de Hilbert $\mathcal{H}$ y una función $\phi : \mathcal{X} \to \mathcal{H}$ que tiene como núcleo la función κ. Primero, se considera el espacio de funciones de $\mathcal{X}$ en $\mathbb{R}$

$$\mathcal{G} = \left\{ \sum_{i=1}^{l} \alpha_i \kappa(x_i, \cdot) : \ l \in \mathbb{N},\ x_i \in \mathcal{X},\ \alpha_i \in \mathbb{R},\ i = 1, \ldots, l \right\}. \tag{2.1}$$

Con la suma y el producto de funciones por escalares habituales, $\mathcal{G}$ es un espacio vectorial. Ahora, se define

$$\phi : \mathcal{X} \to \mathcal{G}, \qquad \phi(x) = \kappa(x, \cdot).$$

Para cada $x \in \mathcal{X}$, $\phi(x)$ es una función de $\mathcal{X}$ en $\mathbb{R}$ que pertenece a $\mathcal{G}$.

Si $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$ y $g(x) = \sum_{j=1}^{m} \beta_j k(z_j, x)$ son elementos de $\mathcal{G}$, se define la aplicación $\langle \cdot, \cdot \rangle : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ como:

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, z_j) = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(z_j). \tag{2.2}$$

$\langle \cdot, \cdot \rangle$ es simétrico y lineal respecto de ambos argumentos. Si $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ y K es la matriz de κ en $x_1, \ldots, x_n$, debido a que κ es finitamente semidefinida positiva, se deduce que

$$\langle f, f \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \geq 0, \quad \forall f \in \mathcal{F}.$$

Por lo tanto, la aplicación $\langle \cdot, \cdot \rangle$ es un producto interno, a priori no estricto, es decir, $\langle f, f \rangle$ puede ser 0 para algún f no nulo. Entonces, $\langle \cdot, \cdot \rangle$ tiene asociada una seminorma $\|f\| = \langle f, f \rangle^{1/2}$. Para probar que $\langle \cdot, \cdot \rangle$ es un producto interno estricto, falta comprobar que si $\langle f, f \rangle = 0$ implica $f = 0$. Para ello, se utiliza la siguiente propiedad de los núcleos:

$$f(x) = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) = \langle f, \kappa(x, \cdot) \rangle = \langle f, \phi(x) \rangle, \quad \forall x \in \mathcal{X}. \tag{2.3}$$

Si $\langle f, f \rangle = 0$, entonces

$$|f(x)| = |\langle f, \phi(x) \rangle| \leq \|f\| \|\phi(x)\| = 0, \quad \forall x \in \mathcal{X}.$$

Se ha usado la desigualdad de Cauchy-Schwarz, que se verifica también para seminormas.

En conclusión, $(\mathcal{G}, \langle \cdot, \cdot \rangle)$ es un espacio vectorial con producto interno y $\phi : \mathcal{X} \to \mathcal{G}$ es una función que verifica $\langle \phi(x), \phi(y) \rangle = \langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle = \kappa(x, y)$. Por último, siempre que se tiene un espacio vectorial normado $\mathcal{G}$, se puede construir su compleción $\mathcal{H}$ (se incluye con una isometría en su doble dual $\mathcal{G}''$, que es una compleción del espacio original). Si, además, la norma del espacio vectorial original proviene de un producto interno, la norma de la compleción también. Esto se debe a que $\mathcal{G}$ es denso en su compleción $\mathcal{H}$. Un argumento para probar lo que se acaba de afirmar es el siguiente: Sean $f, g \in \mathcal{H}$, existen sucesiones en $\mathcal{G}$ tales que $f = \lim_{n\to\infty} f_n$ y $g = \lim_{n\to\infty} g_n$. Como $\mathcal{G}$ es un espacio de Hilbert, la norma verifica la identidad del paralelogramo:

$$\|f_n + g_n\|^2 + \|f_n - g_n\|^2 = 2\|f_n\|^2 + 2\|g_n\|^2.$$

Como la norma es continua, tomando límites, se obtiene que la norma en $\mathcal{H}$ también verifica la identidad del paralelogramo y, por lo tanto, viene de un producto interno.

$\mathcal{H}$ es el espacio de Hilbert que verifica las propiedades del teorema. Se denomina espacio de Hilbert reproductor del núcleo (RKHS, por sus iniciales en inglés).

□

La demostración del teorema de Aronszajn 2.2.3 es constructiva, es decir, caracteriza el espacio de Hilbert $\mathcal{H}$ a partir de la aplicación κ ($\mathcal{H}$ es un espacio de funciones de $\mathcal{X}$ en $\mathbb{R}$) y también se prueba que $\phi(x) = \kappa(x, \cdot)$ para todo $x \in \mathcal{X}$. Además, se ha deducido una propiedad importante de los núcleos (igualdad 2.3). Si $f \in \mathcal{H}$, entonces

$$f(x) = \langle f, \kappa(x, \cdot)\rangle = \langle f, \phi(x)\rangle, \quad \forall x \in \mathcal{X}.$$

Un ejemplo de núcleo con el que se puede trabajar (por ejemplo, en las fórmulas que se deducirán en la siguiente sección) es el núcleo gaussiano $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$\kappa(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \quad \forall x, y \in \mathbb{R}^d.$$

2.3. MMD

Resumiendo de nuevo el planteamiento del trabajo, en el contexto del aprendizaje justo, existe un atributo protegido binario S que divide al conjunto de datos en dos grupos; P y Q son las leyes de probabilidad que generan los datos de cada grupo. En la sección 1.3.2 del capítulo 1 se explica un método para estimar la distancia en variación total y la parte común entre dos probabilidades P y Q en $\mathbb{R}^d$.

En el paso 3 de dicho algoritmo, el objetivo es contrastar si dos probabilidades son iguales (para ver si fijado un nivel α, los conjuntos $\mathcal{R}_\alpha(P)$ y $\mathcal{R}_\alpha(Q)$ tienen intersección no vacía). Para ello, la estrategia propuesta es usar un test basado en una métrica entre probabilidades. Un tipo de distancias que resultan adecuadas en el paso 3 del algoritmo son las métricas de máxima discrepancia media, que se definen a continuación.

Definición 2.3.1. *Sea $(\mathcal{X}, \mathcal{M})$ un espacio medible y $\mathcal{F}$ una clase de funciones de $\mathcal{X}$ en $\mathbb{R}$. Se define la máxima discrepancia en media entre dos probabilidades p y q en $(\mathcal{X}, \mathcal{M})$, denotada por $MMD[\mathcal{F}, p, q]$, como*

$$MMD[\mathcal{F}, p, q] := \sup_{f\in\mathcal{F}} \left| \int f dp - \int f dq \right| = \sup_{f\in\mathcal{F}} |E_p f - E_q f|.$$

Nota 2.3.2. *Fijada $\mathcal{F}$ una clase de funciones de $\mathcal{X}$ en $\mathbb{R}$. Si p y q son dos probabilidades en $(\mathcal{X}, \mathcal{M})$ para las que $\sup_{f\in\mathcal{F}} \int f dp$ y $\sup_{f\in\mathcal{F}} \int f dq$ son valores finitos, entonces $MMD[\mathcal{F}, p, q]$ está bien definida.*

La máxima discrepancia en media mide diferencias entre probabilidades en un espacio $\mathcal{X}$. Es una pseudo métrica. Solo es una métrica si la clase $\mathcal{F}$ es una clase determinante de la probabilidad (por ejemplo, la clase de todas las funciones continuas y acotadas o la clase de las funciones 1-Lipschitz). Se puede definir de forma general como en la definición 2.3.1, pero hay un tipo de clases de funciones $\mathcal{F}$ que resultan especialmente interesantes: el caso en que $\mathcal{F}$ es la bola unidad de un espacio de Hilbert reproductor de núcleo (RKHS). Por esta razón, se dedica parte de la sección a presentar algunas propiedades interesantes de estas métricas en caso RKHS, así como a desarrollar la teoría necesaria para estimar la métrica a partir de muestras. Los resultados que se prueban en esta sección son una adaptación de [14].

El primer paso es obtener una caracterización de la máxima discrepancia en media, esto se consigue a partir de un núcleo. Sea $\mathcal{X}$ el espacio donde toman valores los datos. Si se considera una función $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ simétrica y finitamente definida positiva (un núcleo), se puede reconstruir, siguiendo los pasos del teorema de Aronszajn 2.2.3, el RKHS $\mathcal{H}$ (un espacio de funciones de $\mathcal{X}$ en $\mathbb{R}$). Los siguientes resultados se enuncian teniendo en cuenta este contexto.

Definición 2.3.3. *Sea p una probabilidad en $(\mathcal{X}, \mathcal{M})$, se denomina inmersión promedio a una función $\mu_p \in \mathcal{H}$ que verifique que $E_p f = \langle f, \mu_p \rangle$ para toda $f \in \mathcal{H}$.*

La existencia de la inmersión promedio no siempre está garantizada. Sin embargo, gracias al teorema de Riesz, si la función núcleo $\sqrt{\kappa(x,x)}$ es integrable respecto de una probabilidad p, entonces sí existe la inmersión promedio μ_p. Además, la inmersión promedio se escribe a partir del núcleo. Esto se prueba a continuación.

Proposición 2.3.4. *Sea p una probabilidad en $(\mathcal{X}, \mathcal{M})$. Si la función núcleo $\kappa(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ es medible en $(\mathcal{X}\times\mathcal{X}, \mathcal{M}\times\mathcal{M})$ y la esperanza (respecto de la probabilidad p) de la función $\sqrt{\kappa(x,x)}$ es finita, entonces se puede garantizar la existencia de la inmersión promedio μ_p. En ese caso,*

$$\mu_p(t) = \int \phi(t)(x)dp(x) = E_p\phi(t).$$

Demostración. El funcional $T : \mathcal{H} \to \mathbb{R}$ que envía cada función $f \in \mathcal{H}$ en $E_p f$ es lineal. Y, teniendo en cuenta que, para todo $x \in \mathcal{X}$, $f(x) = \langle f, \phi(x) \rangle$, visto en 2.3, se prueba que es acotado, y, por lo tanto, continuo:

$$|T(f)| = |\int f(x)dp| \leq \int |f(x)|dp = \int |\langle f, \phi(x)\rangle|dp \leq \int \|f\|_{\mathcal{H}}\|\phi(x)\|_{\mathcal{H}}dp =$$

$$= \int \|f\|_{\mathcal{H}}\sqrt{\langle\phi(x),\phi(x)\rangle_{\mathcal{H}}}dp = \int \|f\|_{\mathcal{H}}\sqrt{\kappa(x,x)}dp = \|f\|_{\mathcal{H}}E_p\sqrt{\kappa(x,x)}.$$

Por el teorema de representación de Riesz, existe una única función μ_p de $\mathcal{H}$ tal que $T(f) = \langle f, \mu_p\rangle_{\mathcal{H}}$.

Además, considerando la función $f = \phi(t) : \mathcal{X} \to \mathbb{R}$, se ve que

$$T(f) = \int \phi(t)dp = \int \kappa(t,\cdot)dp = E_p\kappa(t,\cdot).$$

Por otro lado, se tiene que $T(f) = \langle\phi(t), \mu_p\rangle_{\mathcal{H}} = \mu_p(t)$. Se concluye que

$$\mu_p(t) = E_p\kappa(t,\cdot) = E_p\phi(t).$$

□

La importancia de la inmersión promedio reside en que se puede caracterizar la máxima discrepancia en media entre dos probabilidades a partir de sus respectivas inmersiones promedio, si existen. Y, por lo tanto, como se ha visto que la inmersión promedio está definida a partir de la función núcleo, se deduce una expresión para la máxima discrepancia en media que solo depende del núcleo. Este es el objetivo de los siguientes resultados.

Proposición 2.3.5. *Sean p y q dos probabilidades en $(\mathcal{X}, \mathcal{M})$ para las que se supone que existen las respectivas inmersiones promedio μ_p y μ_q. Sea $\mathcal{F}$ la bola unidad del RKHS $\mathcal{H}$. Entonces, la máxima discrepancia en media entre p y q es la distancia en $\mathcal{H}$ entre las respectivas inmersiones promedio. Es decir,*

$$MMD^2[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|^2.$$

Demostración.

$$\begin{aligned} MMD^2[\mathcal{F}, p, q] &= [\sup_{\|f\|\leq 1} (E_p f - E_q f)]^2 = [\sup_{\|f\|\leq 1} (\langle f, \mu_p\rangle - \langle f, \mu_q\rangle)]^2 \\ &= [\sup_{\|f\|\leq 1} \langle f, \mu_p - \mu_q\rangle]^2 \leq \|\mu_p - \mu_q\|^2. \end{aligned}$$

La última desigualdad es consecuencia de la desigualdad de Cauchy-Schwarz. Se alcanza la igualdad en $f = \frac{\mu_p - \mu_q}{\|\mu_p - \mu_q\|} \in B_{\mathcal{H}}(0, 1)$.

□

Teorema 2.3.6. *Sean p y q dos probabilidades en $(\mathcal{X}, \mathcal{M})$ para las que existen las inmersiones promedio. Sea $\mathcal{F}$ la bola unidad en $\mathcal{H}$. Entonces,*

$$MMD^2[\mathcal{F}, p, q] = \iint \kappa(x, \tilde{x}) d(p \times p) + \iint \kappa(y, \tilde{y}) d(q \times q) - 2 \iint \kappa(x, y) d(p \times q). \tag{2.4}$$

Demostración.

$$\begin{aligned} MMD^2[\mathcal{F}, p, q] &= \|\mu_p - \mu_q\|^2 = \langle \mu_p, \mu_p\rangle + \langle \mu_q, \mu_q\rangle - 2\langle \mu_p, \mu_q\rangle \\ &= \int \mu_p(x) dp(x) + \int \mu_q(y) dq(y) - 2 \int \mu_q(x) dp(x) \\ &= \int \left(\int \kappa(x, \tilde{x}) dp(\tilde{x}) \right) dp(x) + \int \left(\int \kappa(y, \tilde{y}) dq(\tilde{y}) \right) dq(y) \\ &\quad - 2 \int \left(\int \kappa(x, y) dq(y) \right) dp(x) \\ &= \iint \kappa(x, \tilde{x}) d(p \times p) + \iint \kappa(y, \tilde{y}) d(q \times q) - 2 \iint \kappa(x, y) d(p \times q). \end{aligned}$$

□

El teorema anterior, 2.3.6, da una caracterización de la métrica MMD en función de los núcleos. Este teorema muestra las ventajas de trasladar los datos a un espacio de Hilbert. Si se tienen dos probabilidades p y q en $\mathcal{X}$, se considera un núcleo $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ tal que la función $\sqrt{\kappa(x, x)}$ sea integrable respecto de ambas probabilidades. Para calcular la máxima discrepancia en media entre dos probabilidades, basta conocer los valores que toma ese núcleo. No se necesita conocer el espacio $\mathcal{H}$.

A continuación, se prueba que, en un espacio métrico compacto $\mathcal{X}$ en el que se pueda definir una función núcleo $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ que cumpla ciertas propiedades (lo que se definirá como núcleo universal), la máxima discrepancia en media es una distancia entre probabilidades.

Lema 2.3.7. *Si $\mathcal{X}$ es un espacio métrico compacto y la función núcleo κ es continua en $\mathcal{X} \times \mathcal{X}$, entonces el RKHS $\mathcal{H}$ está contenido en el espacio de las funciones continuas de $\mathcal{X}$ en $\mathbb{R}$.*

Demostración. La demostración del teorema de Aronszajn 2.2.3 es constructiva. $\mathcal{H}$ es la compleción de $\mathcal{G}$, ver (2.1). Además, por ser κ continua, es claro que $\mathcal{G} \subseteq \mathcal{C}(\mathcal{X}, \mathbb{R})$. A partir de esta afirmación, se prueba que $\mathcal{H} \subseteq \mathcal{C}(\mathcal{X}, \mathbb{R})$.

Si $f \in \mathcal{H}$, existe $\{f_n\}_{n=1}^{\infty} \subseteq \mathcal{G}$ tal que $\lim_{n\to\infty} f_n = f$ para la topología dada por el producto escalar definido en (2.2). Como la función $\kappa(x, x)$ es continua en $\mathcal{X}$ compacto, existe $M \geq 0$ tal que $\kappa(x, x) \leq M^2$, para todo $x \in \mathcal{X}$. Entonces,

$$|f_n(x) - f(x)| = |\langle f_n - f, \phi(x)\rangle| \leq \|f_n - f\|_{\mathcal{H}} \cdot \|\phi(x)\|_{\mathcal{H}} = \|f_n - f\|_{\mathcal{H}} \cdot \sqrt{\kappa(x, x)}$$

$$\Rightarrow |f_n(x) - f(x)| \leq M\|f_n - f\|_{\mathcal{H}}, \quad \forall x \in \mathcal{X},\ \forall n \in \mathbb{N}.$$

Entonces, $\|f_n - f\|_{\infty} \leq M\|f_n - f\|_{\mathcal{H}}$. Se ha probado que la convergencia en la norma de $\mathcal{H}$ implica la convergencia uniforme. Por lo tanto, f es continua. □

Definición 2.3.8. *Sea $\mathcal{X}$ un espacio métrico compacto. Se dice que el RKHS $\mathcal{H}$ es universal si la función núcleo κ es continua en $\mathcal{X} \times \mathcal{X}$ y $\mathcal{H}$ es denso en $\mathcal{C}(\mathcal{X})$.*

Proposición 2.3.9. *Sea $\mathcal{X}$ un espacio métrico compacto y $\mathcal{H}$ universal. Si $\mathcal{F}$ es la bola unidad en $\mathcal{H}$ y p y q son probabilidades en $\mathcal{X}$, entonces $MMD[\mathcal{F}, p, q] = 0$ si, y solo si, $p = q$. Es decir, la máxima discrepancia en media es una distancia en el espacio de probabilidades en $(\mathcal{X}, \beta)$, donde β es la σ-álgebra generada por los abiertos de $\mathcal{X}$.*

Demostración. Como $\mathcal{H}$ es universal, $\kappa(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es continua y $\sqrt{\kappa(x, x)}$ está acotada en $\mathcal{X}$. En particular, tiene esperanza finita. Por las proposiciones 2.3.4 y 2.3.5, $MMD^2[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|^2$. Esto garantiza la simetría y desigualdad triangular.

Es trivial que si $p = q$, entonces $MMD[\mathcal{F}, p, q] = 0$. Recíprocamente, se supone que $MMD[\mathcal{F}, p, q] = 0$, entonces $\mu_p = \mu_q$. Fijado $\varepsilon > 0$ y $f \in \mathcal{C}(\mathcal{X})$, existe $g \in \mathcal{H}$ tal que $\|f - g\|_{\infty} \leq \varepsilon$.

$$\left|\int f dp - \int f dq\right| \leq \left|\int (f - g) dp\right| + \left|\int g dp - \int g dq\right| + \left|\int (g - f) dq\right|$$
$$\leq 2\ \varepsilon + \langle g, \mu_p - \mu_q\rangle = 2\varepsilon.$$

Se tiene que $\int f dp = \int f dq$, para toda $f \in \mathcal{C}(\mathcal{X})$. Como $\mathcal{X}$ es un espacio métrico compacto, entonces es un espacio polaco. Por el lema B.0.5, $p = q$.

□

Por último, se dan algunos posibles estimadores de la máxima discrepancia en media. En la práctica, se usan estos estimadores para aproximar el valor real de la máxima discrepancia en media.

Sean p y q dos probabilidades en el espacio medible $(\mathcal{X}, \mathcal{M})$. Sea $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ un núcleo, para el cual la función de una variable $\sqrt{\kappa(x, x)}$ es integrable respecto de p y q. Entonces, si $\mathcal{F}$ es la bola unidad del espacio de Hilbert reproductor del

núcleo κ, el teorema 2.3.6 da una caracterización de $MMD[\mathcal{F}, p, q]$. Reemplazando las esperanzas que aparecen en la fórmula 2.4 por la media muestral, se obtiene un estimador insesgado de $MMD[\mathcal{F}, p, q]$. Teniendo en cuenta este contexto, se da siguiente resultado:

Proposición 2.3.10 (Estimadores insesgados)**.** *Sean p y q dos probabilidades en el espacio medible $(\mathcal{X}, \mathcal{M})$ para las cuales se verifica el teorema 2.3.6. Sea $\{x_1, x_2, ..., x_m\}$ una muestra de tamaño m de la distribución p y sea $\{y_1, y_2, ..., y_n\}$ una muestra de tamaño n de la distribución q. Un estimador insesgado de la máxima discrepancia en media al cuadrado entre p y q viene dado por la siguiente fórmula:*

$$MMD_u^2[\mathcal{F}, p, q] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \kappa(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \kappa(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \kappa(x_i, y_j). \tag{2.5}$$

Si se toman muestras de igual tamaño, es decir, $m = n$, entonces se puede usar un estimador insesgado un poco más simple (es diferente del anterior).

$$MMD_u^2[\mathcal{F}, p, q] = \frac{1}{m(m-1)} \left[\sum_{i \neq j}^{m} \kappa(x_i, x_j) + \sum_{i \neq j}^{m} \kappa(y_i, y_j) - 2 \sum_{i \neq j}^{m} \kappa(x_i, y_j) \right]. \tag{2.6}$$

Demostración. Ambos estimadores son U-estadísticos, más detalles se pueden encontrar en [17]. La consistencia de estos estimadores está probada en el teorema A de la sección 5.4. de dicho libro. Además, ambos estimadores son insesgados. Se comprueba para el segundo:

$$\begin{aligned}
&E\Big(MMD_u^2[\mathcal{F}, p, q]\Big) \\
&= \frac{1}{m(m-1)} \left[\sum_{i \neq j}^{m} \iint \kappa(x_i, x_j) dp dp + \sum_{i \neq j}^{m} \iint \kappa(y_i, y_j) dq dq - 2 \sum_{i \neq j}^{m} \iint \kappa(x_i, y_j) dp dq \right] \\
&= \iint \kappa(x, \tilde{x}) d(p \times p) + \iint \kappa(y, \tilde{y}) d(q \times q) - 2 \iint \kappa(x, y) d(p \times q) \\
&= MMD^2[\mathcal{F}, p, q].
\end{aligned}$$

□

Nota 2.3.11. *Aunque la máxima discrepancia en media entre dos probabilidades para las que existen las inmersiones promedio respecto de un núcleo es siempre positiva (ver 2.3.5), el estimador MMD_u (cualquiera de las dos versiones) sí puede ser negativo. Por ejemplo, si $p = q$, entonces $E\Big(MMD_u^2[\mathcal{F}, p, q]\Big) = MMD^2[\mathcal{F}, p, q] = 0$ y $MMD_u^2[\mathcal{F}, p, q]$ tomará valores por debajo de su media (negativos).*

Otro estimador posible para la máxima discrepancia en media entre p y q se deduce de la proposición 2.3.5, donde se prueba que la máxima discrepancia en media entre dos probabilidades es la distancia entre sus respectivas inmersiones promedio (en el RKHS). Por lo tanto, si se sustituyen las inmersiones promedio por sus aproximaciones empíricas a partir de muestras, se llega a un estimador sesgado de MMD.

Proposición 2.3.12 (Estimador sesgado)**.** *Sean p y q dos probabilidades en el espacio medible $(\mathcal{X}, \mathcal{M})$ para las cuales se verifica la proposición 2.3.5. Sea $\{x_1, x_2, ..., x_m\}$ una muestra de tamaño m de la distribución p y sea $\{y_1, y_2, ..., y_n\}$ una muestra de tamaño n de la distribución q. Entonces, un estimador sesgado de la máxima discrepancia en media (al cuadrado) entre p y q es:*

$$MMD_b^2[\mathcal{F}, p, q] = \frac{1}{m^2}\sum_{i,j=1}^{m}\kappa(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n}\kappa(y_i, y_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n}\kappa(x_i, y_j). \quad (2.7)$$

Se explica cómo se ha deducido esa fórmula: Sean $\mu_X = \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$ y $\mu_Y = \frac{1}{n}\sum_{i=1}^{n}\phi(y_i)$ los estimadores asociados a esas muestras de las inmersiones promedio de p y q. Entonces,

$$\begin{aligned}
&\|\mu_X - \mu_Y\|^2 \\
&= \Big\langle \frac{1}{m}\sum_{i=1}^{m}\phi(x_i), \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\Big\rangle + \Big\langle \frac{1}{n}\sum_{i=1}^{n}\phi(y_i), \frac{1}{n}\sum_{i=1}^{n}\phi(y_i)\Big\rangle \\
&- 2\Big\langle \frac{1}{m}\sum_{i=1}^{m}\phi(x_i), \frac{1}{n}\sum_{i=1}^{n}\phi(y_i)\Big\rangle \\
&= MMD_b^2[\mathcal{F}, p, q].
\end{aligned}$$

En general, los tests basados en la métrica MMD se harán usando un estimador empírico. En el caso de este trabajo, se usa un test basado en la máxima discrepancia en media para contrastar si dos probabilidades son iguales (si, y solo si, su máxima discrepancia en media es 0, bajo ciertas condiciones que se han explicado en la sección 2.3). Entonces, se necesita que si dos probabilidades son iguales el estimador empírico de la máxima discrepancia en media que se utilice también sea pequeño.

Las cotas que controlan el error de estos estimadores, asegurando la veracidad de los resultados obtenidos al hacer algún test de este tipo, se pueden encontrar en el artículo [14].

Capítulo 3

El problema de transporte óptimo

Sean $(\mathcal{X}, \sigma_x, P)$ e $(\mathcal{Y}, \sigma_y, Q)$ dos espacios probabilísticos. Se quiere transportar la masa distribuida en el espacio $\mathcal{X}$, según la distribución P, al espacio $\mathcal{Y}$, con distribución de probabilidad Q. El coste de transportar una unidad de masa de $\mathcal{X}$ a $\mathcal{Y}$ depende de cada $x \in \mathcal{X}$ e $y \in \mathcal{Y}$ concretos. Por lo tanto, viene dado por una función $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. El objetivo es encontrar el plan de transporte menos costoso.

Un caso particular es el de espacios finitos,

$$\mathcal{X} = \{x_1, ...x_n\}, \quad \mathcal{Y} = \{y_1, ...y_m\}.$$

Si $i \in \{1, ..., n\}$, la masa de x_i es p_i. Respectivamente, para $j \in \{1, ..., m\}$, la masa de $y_j \in \mathcal{Y}$ es q_j. Obviamente, $\sum_{i=1}^{n} p_i = \sum_{j=1}^{n} q_j = 1$. Se denota por $\pi_{i,j} = \pi(x_i, y_j)$ a la masa transportada de x_i a y_j.

La función $c(x_i, y_j)$ da el coste unitario de transporte de x_i a y_j. La idea es reordenar en $\mathcal{Y}$ la masa de $\mathcal{X}$ para que al final esté distribuida según la ley de probabilidad Q, minimizando el coste.

Matemáticamente, el problema se formula de la siguiente manera:

$$\min_{\sum_{j=1}^{m} \pi_{i,j} = p_i, \sum_{i=1}^{n} \pi_{i,j} = q_j} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{i,j} c(x_i, y_j) = \min_{\pi \in \prod(P,Q)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y).$$

Por ejemplo, si una empresa quiere transportar bienes desde las fábricas a los almacenes, teniendo en cuenta su localización, los costes de transporte dependerán de cada fábrica y cada almacén. Querrá escoger el plan de transporte óptimo, si existe, para minimizar los costes.

Aunque el caso de espacios finitos tenga muchas aplicaciones prácticas, es muy interesante generalizarlo. Con el mismo objetivo de otros capítulos: encontrar distancias entre probabilidades, se estudiará el problema de transporte óptimo para $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, con la σ-álgebra de Borel β^d, y el coste $c(x, y) = \|x - y\|^p$, con $p \geq 1$.

3.1. Formulaciones del problema

La formulación de Kantorovich del problema de transporte óptimo que se acaba de plantear es la siguiente:

Sea $p \geq 1$. Sean P y Q dos probabilidades en $(\mathbb{R}^d, \beta^d)$ con momentos de orden p finitos. Se denota por $\prod(P,Q)$ al conjunto de probabilidades en $(\mathbb{R}^d \times \mathbb{R}^d, \beta^{2d})$ con marginales P y Q. El objetivo es encontrar la distribución de probabilidad conjunta $\pi \in \prod(P,Q)$ que minimice el coste de transporte, denotado por $I(\pi)$. Es decir, hallar:

$$\mathcal{T}_c(P,Q) := \inf_{\pi \in \prod(P,Q)} I(\pi) = \inf_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y).$$

Se supone que P y Q tienen momentos finitos de orden $p \geq 1$, para que la integral $I(\pi)$ esté bien definida: Si

$$\int_{\mathbb{R}^n} \|x\|^p dP(x) + \int_{\mathbb{R}^n} \|y\|^p dQ(y) < +\infty,$$

entonces,

$$\begin{aligned} I(\pi) &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} (\|x\| + \|y\|)^p d\pi(x,y) \\ &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} (2\,\text{máx}\{\|x\|, \|y\|\})^p d\pi(x,y) \leq 2^p \int_{\mathbb{R}^d \times \mathbb{R}^d} (\|x\|^p + \|y\|^p) d\pi(x,y) \\ &= 2^p \int_{\mathbb{R}^n} \|x\|^p dP(x) + 2^p \int_{\mathbb{R}^n} \|y\|^p dQ(y) < +\infty. \end{aligned}$$

El primer paso para abordar el problema de transporte óptimo es probar que la formulación que se ha dado es coherente.

Proposición 3.1.1. *En el problema descrito, el conjunto $\prod(P,Q)$ es no vacío y*

$$\inf_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y)$$

se alcanza para una probabilidad concreta $\pi_0 \in \prod(P,Q)$, es decir, es un mínimo. Se dice que π_0 es un plan de transporte óptimo.

Demostración. La definición de sucesión ajustada, junto con las propiedades que se usarán en esta demostración, y el teorema de Portmanteau se detallan en el apéndice C.

En la σ-álgebra producto $\beta^d \otimes \beta^d = \beta^{2d}$ se define la medida producto

$$(P \times Q)(A) = \int_{\mathbb{R}^d} Q(A_x) dP(x) = \int_{\mathbb{R}^d} P(A_y) dQ(y), \quad \forall A \in \beta^{2d}.$$

Es fácil comprobar que $(P \times Q)$ es una probabilidad cuyas marginales son P y Q.

Como P y Q tienen momentos de orden p finitos, si $\pi \in \prod(P,Q)$, se tiene que $0 \leq I(\pi) < \infty$. Entonces, $0 \leq \inf_{\pi \in \prod(P,Q)} I(\pi) < \infty$. Por definición de inferior, existe una sucesión $\{\pi_n\}_{n=1}^{\infty} \subset \prod(P,Q)$ tal que

$$\lim_{n \to \infty} I(\pi_n) = \inf_{\pi \in \prod(P,Q)} I(\pi).$$

Es conocido que $\forall \varepsilon > 0$, existen K_ε y F_ε compactos de $\mathbb{R}^d$ tales que $P(K_\varepsilon) \geq 1 - \frac{\varepsilon}{2}$ y $Q(F_\varepsilon) \geq 1 - \frac{\varepsilon}{2}$. Por lo tanto, la sucesión $\{\pi_n\}_{n=1}^{\infty}$ es ajustada, ya que $\forall \varepsilon > 0$ y $\forall n \in \mathbb{N}$,

$$\pi_n((K_\varepsilon \times F_\varepsilon)^c) \leq \pi_n(K_\varepsilon^c \times \mathbb{R}^d) + \pi_n(\mathbb{R}^d \times F_\varepsilon^c) = P(K_\varepsilon^c) + P(F_\varepsilon^c) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Esto implica que existe una subsucesión $\{\pi_{n_k}\}_{k=1}^{\infty}$ de $\{\pi_n\}_{n=1}^{\infty}$ que converge hacia una probabilidad π_0. Se ve fácilmente que $\pi_0 \in \prod(P, Q)$. La demostración se concluye probando que

$$I(\pi_0) = \inf_{\pi \in \prod(P,Q)} I(\pi).$$

Se considera una sucesión de funciones $c_l : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ positivas, crecientes, continuas y acotadas, tal que $\lim_{l\to\infty} c_l(x, y) = c(x, y) := \|x - y\|^p$. Por ejemplo, se puede definir

$$c_l(x, y) = \begin{cases} c(x, y) & \text{si } c(x, y) \leq l, \\ l & \text{si } c(x, y) > l. \end{cases}$$

En estas condiciones, el teorema de Portmanteau garantiza que

$$\lim_{k\to\infty} \int c_l(x, y) d\pi_{n_k} = \int c_l(x, y) d\pi_0, \quad \forall l \in \mathbb{N}.$$

Además, por el teorema de la convergencia monótona, $\int c(x, y) d\pi_0 = \lim_{l\to\infty} \int c_l(x, y) d\pi_0$ Por lo tanto,

$$\begin{aligned} \int c(x, y) d\pi_0 &= \lim_{l\to\infty} \int c_l(x, y) d\pi_0 \\ &= \lim_{l\to\infty} \lim_{k\to\infty} \int c_l(x, y) d\pi_{n_k} \\ &\leq \lim_{k\to\infty} \int c(x, y) d\pi_{n_k} = \inf_{\pi \in \prod(P,Q)} I(\pi). \end{aligned}$$

En la última desigualdad se ha usado que $c_l \leq c$, para todo $l \in \mathbb{N}$.

Como $\pi_0 \in \prod(P, Q)$, se tiene la desigualdad contraria, y, por tanto, la igualdad. □

Se puede restringir el conjunto de probabilidades en el que se quiere minimizar el coste de transporte $I(\pi)$. Se considera el conjunto de aplicaciones medibles

$$T : \mathbb{R}^d \to \mathbb{R}^d$$

tales que $T\#P = Q$, es decir, $Q(B) = P(T^{-1}(B))$ para todo $B \in \beta^d$.

La formulación de Monge del problema consiste en hallar

$$\tilde{\mathcal{T}}_c(P, Q) = \inf_{\substack{T:\mathbb{R}^d \to \mathbb{R}^d \\ T\#P=Q}} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - T(x)\|^p dP(x).$$

En el caso discreto, esta formulación tiene un significado claro: la masa de cada punto de $\mathcal{X}$ no se puede dividir entre varios puntos de $\mathcal{Y}$. Es decir, si $\mathcal{X} = \{x_1, ..., x_n\}$ e

$\mathcal{Y} = \{y_1, ..., y_m\}$, se consideran las aplicaciones $T : \mathcal{X} \to \mathcal{Y}$ que reordenan la masa para que la distribución pase de ser la dada por P en $\mathcal{X}$ a la dada por Q en $\mathcal{Y}$. Cada aplicación induce una probabilidad π en el espacio producto $\prod(P, Q)$. De entre todas ellas, se escoge a la que minimice el coste de transporte.

Como el problema de Monge tiene más restricciones que el de Kantorovich, es lógico pensar que

$$\mathcal{T}_c(P, Q) \leq \tilde{\mathcal{T}}_c(P, Q). \tag{3.1}$$

En efecto, si T es una aplicación tal que $T\#P = Q$, Sea π la probabilidad en $(\mathbb{R}^d \times \mathbb{R}^d, \beta^{2d})$ inducida por la aplicación

$$(Id, T) : (\mathbb{R}^d, \beta^d, P) \to (\mathbb{R}^d \times \mathbb{R}^d, \beta^{2d}), \quad x \mapsto (x, T(x)).$$

Es decir, si $A \in \beta^{2d}$, entonces $\pi(A) = P(\{x \in \mathbb{R}^d : (x, T(x)) \in A\})$. Es claro que $\pi \in \prod(P, Q)$. Por el teorema de transferencia de integrales, se tiene que para cualquier T tal que $T\#P = Q$,

$$\min_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\pi(x, y) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\pi(x, y) = \int_{\mathbb{R}^d} c(x, T(x)) dP(x),$$

con lo que se tendría la desigualdad 3.1.

Definición 3.1.2. *Se denomina aplicación de transporte óptimo de P a Q, para el problema descrito anteriormente, a una aplicación $T : \mathbb{R}^d \to \mathbb{R}^d$ medible tal que $T\#P = Q$, es decir, $Q(B) = P(T^{-1}(B))$ para todo $B \in \beta^d$, y que verifique*

$$\min_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) = \int_{\mathbb{R}^d} \|x - T(x)\|^p dP(x).$$

Se probará que si P no da probabilidad a conjuntos de medida nula, existe una aplicación de transporte óptimo para el problema asociado al coste cuadrático, es decir, se alcanza la igualdad en 3.1. Con el fin de llegar a este resultado, la siguiente sección está centrada en el estudio del problema de transporte para el coste cuadrático y su formulación dual.

3.2. El caso cuadrático

En esta sección se va a estudiar el problema del transporte para el coste cuadrático, como se hace en el libro [22].

Sean P y Q dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos, el problema consiste en calcular

$$\mathcal{T}_c(P, Q) = \min_{\pi \in \prod(P,Q)} I(\pi) = \inf_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\pi(x, y).$$

El primer paso es probar una fórmula de dualidad que va a permitir simplificar el estudio del problema. Al final, el objetivo va a ser minimizar una expresión en la que van a ser clave las funciones convexas, cuyas propiedades se estudiarán también en esta sección.

Se enuncia a continuación la fórmula de dualidad de Kantorovich.

Proposición 3.2.1 (Dualidad de Kantorovich)**.** *Sean P y Q dos probabilidades en $(\mathbb{R}^d, \beta^d)$. Sea $c(x,y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ inferiormente semi-continua y positiva, entonces*

$$\inf_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) d\pi(x,y) = \sup_{\substack{(f,g) \in L^1(P) \times L^1(Q) \\ f(x)+g(y) \leq c(x,y)}} \int_{\mathbb{R}^d} f(x) dP + \int_{\mathbb{R}^d} g(y) dQ. \quad (3.2)$$

Demostración. Paso 1: Sean K_1 y K_2 compactos en $\mathbb{R}^d$ y sean P y Q probabilidades concentradas en K_1 y K_2, respectivamente. Se supone que la función coste es continua y positiva en $K_1 \times K_2$. Por el teorema de representación de Riesz, que se explica en el apéndice B.1, el dual topológico de $\mathcal{C}(K_1 \times K_2)$, con la norma infinito, es el espacio de medidas de Borel regulares en $K_1 \times K_2$. Pero, en $K_1 \times K_2 \subset \mathbb{R}^d \times \mathbb{R}^d$ todas las medidas de Borel son regulares (ver B). En otras palabras, si T es un funcional lineal y acotado en $\mathcal{C}(K_1 \times K_2)$, existe una medida de Borel μ en $K_1 \times K_2$ tal que

$$Tf = \int_{K_1 \times K_2} f d\mu, \quad \forall f \in \mathcal{C}(K_1 \times K_2).$$

Se consideran los siguientes funcionales convexos:

$$\Theta : u \in \mathcal{C}(K_1 \times K_2) \mapsto \begin{cases} 0 & \text{si } u(x,y) \geq -c(x,y), \\ +\infty & \text{en otro caso.} \end{cases}$$

$$\Xi : u \in \mathcal{C}(K_1 \times K_2) \mapsto \begin{cases} \int f(x) dP + \int g(y) dQ & \text{si } u(x,y) = f(x) + g(y), \\ +\infty & \text{en otro caso.} \end{cases}$$

Se comprueba fácilmente que el funcional Ξ está bien definido. Además, para la función $v_0 \equiv 1$, se cumple que Θ es continuo en v_0 y

$$\Theta(v_0) = 0 < +\infty, \quad \Xi(v_0) = 1 < +\infty.$$

Se verifica el resultado de dualidad de Fencher-Rockafellar, explicado en el apéndice D.3, entonces:

$$\inf_{u \in \mathcal{C}(K_1 \times K_2)} \{\Theta(u) + \Xi(u)\} = \sup_{\pi \in M(K_1 \times K_2)} \{-\Theta^*(-\pi) - \Xi(\pi)\},$$

donde se denota por $M(K_1 \times K_2)$ al espacio de medidas de Borel en $K_1 \times K_2$. Desarrollando el lado izquierdo de la igualdad, se tiene que:

$$\begin{aligned} \inf_{\mathcal{C}(K_1 \times K_2)} \{\Theta + \Xi\} &= \text{ínf} \Big\{ \int f(x) dP + \int g(y) dQ : f(x) + g(y) \geq -c(x,y) \Big\} \\ &= -\sup \Big\{ \int f(x) dP + \int g(y) dQ : f(x) + g(y) \leq c(x,y) \Big\}. \end{aligned}$$

En cuanto al lado derecho, primero se calculan las transformaciones de Legendre de los funcionales Θ y Ξ:

$$\Theta^*(-\pi) = \sup_{u\in\mathcal{C}(K_1\times K_2)} \Big\{ -\int u(x,y)d\pi : u(x,y) \geq -c(x,y)\Big\}$$
$$= \sup_{u\in\mathcal{C}(K_1\times K_2)} \Big\{ \int u(x,y)d\pi : u(x,y) \leq c(x,y)\Big\}.$$

Si π no es una medida positiva, existe una función $v \in \mathcal{C}(K_1 \times K_2)$ estrictamente negativa tal que $\int v d\pi > 0$. Entonces, para todo $n \in \mathbb{N}$, $nv \in \mathcal{C}(K_1 \times K_2)$ y $\int nv d\pi$ tiende a $-\infty$ cuando n tiende a ∞. Si, por el contrario, la medida π es positiva, $\int u(x,y)d\pi \leq \int c(x,y)d\pi$ para toda $u \in \mathcal{C}(K_1 \times K_2)$. En conclusión,

$$\Theta^*(-\pi) = \begin{cases} \int c(x,y)d\pi & \text{si } \pi \text{ es una medida positiva,} \\ +\infty & \text{si } \pi \text{ no es una medida positiva.} \end{cases}$$

Por otro lado,

$$\Xi^*(\pi) = \sup_{f(x)+g(y)\in\mathcal{C}(K_1\times K_2)} \Big\{ \int [f(x)+g(y)]d\pi - \big(\int f(x)dP + \int g(y)dQ\big)\Big\}.$$

Si $\pi \in \prod(P,Q)$, entonces $\Xi(\pi) = 0$. Si $\pi \notin \prod(P,Q)$, existe $f(x)+g(y) \in \mathcal{C}(K_1\times K_2)$ tal que $\int_{\mathbb{R}^d} f(x)dP + \int_{\mathbb{R}^d} g(y)dQ = 0$ y $\int_{\mathbb{R}^d\times\mathbb{R}^d}[f(x)+g(y)]d\pi \neq 0$, escalando la función $f(x)+g(y)$, se llega a que $\Xi(\pi) = +\infty$. Resumiendo,

$$\Xi^*(-\pi) = \begin{cases} 0 & \text{si } \pi \in \prod(P,Q), \\ +\infty & \text{si } \pi \notin \prod(P,Q). \end{cases}$$

Por lo tanto,

$$\sup_{\pi\in M(\mathbb{R}^d\times\mathbb{R}^d)} \Big\{ -\Theta^*(-\pi) - \Xi(\pi)\}$$
$$= \sup\{-\int c(x,y)d\pi : \pi \text{ es una medida positiva y } \pi \in \prod(P,Q)\Big\}$$
$$= -\,\text{ínf}\{\int c(x,y)d\pi : \pi \text{ es una medida positiva y } \pi \in \prod(P,Q)\Big\}.$$

Con todo esto, se ha probado que

$$\sup\Big\{\int_{\mathbb{R}^d} f(x)dP + \int_{\mathbb{R}^d} g(y)dQ : f(x)+g(y) \leq c(x,y)\Big\}$$
$$= \text{ínf}\{\int c(x,y)d\pi : \pi \text{ es una medida positiva y } \pi \in \prod(P,Q)\Big\}.$$

Paso 2: Se supone que $c(x,y)$ es una función continua, acotada y positiva en $\mathbb{R}^d \times \mathbb{R}^d$. Repitiendo el razonamiento de la demostración de 3.1.1, se prueba que existe una probabilidad $\pi_* \in \prod(P,Q)$ tal que

$$I(\pi_*) := \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\pi_*(x,y) = \inf_{\pi\in\prod(P,Q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 d\pi(x,y).$$

Como se detalla en el apéndice B, para todo $\varepsilon > 0$, existen K_1 y K_2 compactos de $\mathbb{R}^d$ tales que

$$\pi_*\big((K_1 \times K_2)^C\big) \leq \varepsilon.$$

Buscando aplicar el paso 1, se considera la probabilidad π_* restringida al compacto $K_1 \times K_2$, es decir, se define:

$$\pi_{*0}(B) = \frac{1}{\pi_*(K_1 \times K_2)}\pi_*(B), \quad \forall B \subset K_1 \times K_2 \text{ medible.}$$

Si P_0 y Q_0 son las marginales de π_{*0}, tienen soporte en K_1 y K_2, respectivamente. Entonces, si se denota por $\prod_0(P_0, Q_0)$ al conjunto de probabilidades en $K_1 \times K_2$ con marginales P_0 y Q_0, existe una probabilidad π_0 que minimiza el coste de transporte óptimo en $\prod_0(P_0, Q_0)$. La probabilidad $\tilde{\pi}$ definida como:

$$\tilde{\pi}(A) = \pi_*(K_1 \times K_2)\pi_0\Big(A \cap (K_1 \times K_2)\Big) + \pi_*\Big(A \cap (K_1 \times K_2)^C\Big), \quad \forall A \in \beta^d \times \beta^d.$$

tiene marginales P y Q. Se comprueba que la primera marginal es P; para probar que la segunda es Q, se razona de forma análoga.

$$\begin{aligned}
&\tilde{\pi}(C \times \mathbb{R}^d) = \pi_*(K_1 \times K_2)P_0(C \cap K_1) + \pi_*\Big((C \times \mathbb{R}^d) \cap (K_1 \times K_2)^C\Big) \\
&= \pi_*(K_1 \times K_2)\pi_{*0}\Big((C \cap K_1) \times K_2\Big) + \pi_*\Big((C \cap K_1^C) \times \mathbb{R}^d\Big) + \pi_*\Big((C \cap K_1) \times K_2^C\Big) \\
&= \pi_*(K_1 \times K_2)\frac{\pi_*\Big((C \cap K_1) \times K_2\Big)}{\pi_*(K_1 \times K_2)} + \pi_*\Big((C \cap K_1^C) \times \mathbb{R}^d\Big) + \pi_*\Big((C \cap K_1) \times K_2^C\Big) \\
&= \pi_*\Big((C \cap K_1) \times \mathbb{R}^d\Big) + \pi_*\Big((C \cap K_1^C) \times \mathbb{R}^d\Big) = P(C), \quad \forall C \in \beta^d.
\end{aligned}$$

Además, se verifica la siguiente desigualdad:

$$\begin{aligned}
I(\tilde{\pi}) &= \pi_*(K_1 \times K_2)\int_{K_1 \times K_2} c(x,y)d\pi_0 + \int_{(K_1 \times K_2)^C} c(x,y)d\pi_* \\
&\leq \int_{K_1 \times K_2} c(x,y)d\pi_0 + \|c\|_\infty \varepsilon.
\end{aligned}$$

Entonces,

$$\inf_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)d\pi \leq \int_{K_1 \times K_2} c(x,y)d\pi_0 + \|c\|_\infty \varepsilon.$$

Por otro lado, se considera el funcional J_0 definido en $L^1(P_0) \times L^1(Q_0)$:

$$J_0(\varphi_0, \psi_0) = \int_{K_1} \varphi_0 \, dP_0 + \int_{K_2} \psi_0 \, dQ_0.$$

Aplicando el paso 1 de la demostración, se tiene que

$$\int_{K_1 \times K_2} c(x,y)d\pi_0(x,y) = \sup_{(\varphi_0, \psi_0) \in \Phi_0} J_0(\varphi_0, \psi_0),$$

donde $\Phi_0 = \{(\varphi_0, \psi_0) \in L^1(P_0) \times L^1(Q_0) : \varphi_0(x) + \psi_0(y) \leq c(x,y)\}$. Por definición de superior, para todo $\varepsilon > 0$, existe una pareja de funciones $\tilde{\varphi}_0, \tilde{\psi}_0 \in \Phi_0$ tal que

$$J_0(\tilde{\varphi}_0, \tilde{\psi}_0) \geq \sup_{(\varphi_0,\psi_0)\in\Phi_0} J_0(\varphi_0, \psi_0) - \varepsilon.$$

En particular, para $\varepsilon = 1$,

$$J_0(\tilde{\varphi}_0, \tilde{\psi}_0) \geq \sup_{(\varphi_0,\psi_0)\in\Phi_0} J_0(\varphi_0, \psi_0) - 1 \geq J_0(0,0) - 1 \geq -1.$$

Escribiendo

$$J_0(\tilde{\varphi}_0, \tilde{\psi}_0) = \int_{K_1\times K_2} [\tilde{\varphi}_0(x) + \tilde{\psi}_0(y)]\, d\pi_0(x,y),$$

se deduce que existe $(x_0, y_0) \in K_1 \times K_2$ tal que

$$\tilde{\varphi}_0(x_0) + \tilde{\psi}_0(y_0) \geq -1. \tag{3.3}$$

Para todo $t \in \mathbb{R}$, se verifica que $(\varphi_0, \psi_0) \in \Phi_0$ si, y solo si, $(\varphi_0 + t, \psi_0 - t) \in \Phi_0$. Además, $J_0(\varphi_0, \psi_0) = J_0(\varphi_0 + t, \psi_0 - t)$. Escogiendo t de forma adecuada, se garantiza que

$$\tilde{\varphi}_0(x_0) + t \geq -\frac{1}{2} \quad \text{y} \quad \tilde{\psi}_0(y_0) - t \geq -\frac{1}{2}.$$

Para simplificar la notación, se supone que

$$\tilde{\varphi}_0(x_0) \geq -\frac{1}{2} \quad \text{y} \quad \tilde{\psi}_0(y_0) \geq -\frac{1}{2}. \tag{3.4}$$

Por lo tanto,

$$\tilde{\varphi}_0(x) \leq c(x, y_0) - \tilde{\psi}_0(y_0) \leq c(x, y_0) + \frac{1}{2}, \quad \forall x \in K_1, \tag{3.5}$$

$$\tilde{\psi}_0(y) \leq c(x_0, y) - \tilde{\varphi}_0(x_0) \leq c(x_0, y) + \frac{1}{2}, \quad \forall y \in K_2. \tag{3.6}$$

Ahora, se define la función

$$f_0(x) := \inf_{y\in K_2} [c(x,y) - \tilde{\psi}_0(y)], \quad \forall x \in K_1.$$

De la desigualdad $\tilde{\varphi}_0(x) \leq c(x,y) - \tilde{\psi}_0(y)$, se deduce que $\tilde{\varphi}_0 \leq f_0$. Entonces,

$$J(\tilde{\varphi}_0, \tilde{\psi}_0) \leq J(f_0, \tilde{\psi}_0).$$

Además, a partir de las desigualdades 3.6 y 3.4, se tiene una cota superior e inferior de f_0:

$$f_0(x) \geq \inf_{y\in K_2} [c(x,y) - c(x_0,y)] - \frac{1}{2}, \quad \forall x \in K_1, \tag{3.7}$$

$$f_0(x) \leq c(x, y_0) - \tilde{\psi}_0(y_0) \leq c(x, y_0) + \frac{1}{2}, \quad \forall x \in K_1. \tag{3.8}$$

Por último, se define la función

$$g_0(y) := \inf_{x\in K_1}[c(x,y) - f_0(x)], \quad \forall y \in K_2.,$$

Entonces, $(f_0, g_0) \in \Phi_0$ y $g_0 \geq \tilde{\psi}_0$, ya que

$$g_0(y) \geq \inf_{x\in K_1}[c(x,y) - c(x,y) + \tilde{\psi}_0(y)] \geq \tilde{\psi}_0(y), \quad \forall y \in K_2.$$

Por esta razón, se tiene la siguiente cadena de desigualdades:

$$J(f_0, g_0) \geq J(f_0, \tilde{\psi}_0) \geq J(\tilde{\varphi}_0, \tilde{\psi}_0).$$

A partir de las cotas para f_0 (3.8 y 3.4), se deducen cotas para g_0:

$$g_0(y) \geq \inf_{x\in K_1}[c(x,y) - c(x,y_0)] - \frac{1}{2}, \quad \forall y \in K_2, \tag{3.9}$$

$$g_0(y) \leq c(x_0,y) - f_0(x_0) \leq c(x_0,y) - \tilde{\varphi}_0(x_0) \leq c(x_0,y) + \frac{1}{2}, \quad \forall y \in K_2. \tag{3.10}$$

Resumiendo, como $c(x,y) \geq 0$, de 3.7 y 3.9 se concluye que

$$f_0(x) \geq -\|c\|_\infty - \frac{1}{2}, \quad \text{y} \quad g_0(y) \geq -\|c\|_\infty - \frac{1}{2}.$$

De estas desigualdades junto con las cotas superiores 3.10 y 3.8, se deduce que $(f_0, g_0) \in L^1(P_0) \times L^1(Q_0)$. Con estas cotas, y extendiendo las funciones f_0 y g_0 por 0, se puede concluir:

$$\begin{aligned}
&\int_{\mathbb{R}^d} f_0 dP + \int_{\mathbb{R}^d} g_0 dQ = \int_{\mathbb{R}^d\times\mathbb{R}^d} [f_0(x) + g_0(y)] d\tilde{\pi}(x,y) \\
&= \pi_*[K_1 \times K_2] \int_{K_1\times K_2} [f_0(x) + g_0(y)]\, d\pi_0(x,y) + \int_{(K_1\times K_2)^C} [f_0(x) + g_0(y)]\, d\pi_*(x,y) \\
&\geq (1-\varepsilon)\left(\int_{K_1} f_0\, dP_0 + \int_{K_2} g_0\, dQ_0\right) - (2\|c\|_\infty + 1)\pi_*[(K_1\times K_2)^c] \\
&\geq (1-\varepsilon) J_0(f_0, g_0) - (2\|c\|_\infty + 1)\varepsilon \\
&\geq (1-\varepsilon)\left(\inf_{\pi\in\prod(P_0,Q_0)} \int_{K_1\times K_2} c(x,y) d\pi - \varepsilon\right) - (2\|c\|_\infty + 1)\varepsilon \\
&\geq (1-\varepsilon)\left(\inf_{\pi\in\prod(P,Q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y) d\pi - \|c\|_\infty\varepsilon - \varepsilon\right) - (2\|c\|_\infty + 1)\varepsilon.
\end{aligned}$$

Si ε tiende a 0, se concluye que

$$\inf_{\pi\in\prod(P,Q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y) d\pi(x,y) = \sup_{\substack{(f,g)\in L^1(P)\times L^1(Q) \\ f(x)+g(y)\leq c(x,y)}} \int_{\mathbb{R}^d} f(x) dP + \int_{\mathbb{R}^d} g(y) dQ.$$

La desigualdad contraria está garantizada siempre, ya que

$$f(x) + g(y) \leq c(x,y), \quad \forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

Paso 3: Si la función de coste $c(x,y)$ es inferiormente semi-continua y positiva, existe una sucesión $\{c_n\}_{n=1}^{\infty} \subset \mathcal{C}_c(\mathbb{R}^d \times \mathbb{R}^d)$ creciente y de funciones positivas tal que

$$\lim_{n\to\infty} c_n(x,y) = c(x,y), \quad \forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

Si π es una probabilidad con marginales P y Q,

$$I_n(\pi) = \int c_n(x,y) d\pi(x,y).$$

En primer lugar, se va a probar que

$$\inf_{\pi\in\prod(P,Q)} I(\pi) = \sup_{n\in\mathbb{N}} \inf_{\pi\in\prod(P,Q)} I_n(\pi). \tag{3.11}$$

Como $c_n \leq c$ para todo $n \in \mathbb{N}$, entonces $I_n(\pi) \leq I(\pi)$ para toda probabilidad $\pi \in \prod(P,Q)$. De aquí, se deduce que

$$\inf_{\pi\in\prod(P,Q)} I(\pi) \geq \sup_{n\in\mathbb{N}} \inf_{\pi\in\prod(P,Q)} I_n(\pi).$$

Queda probar que

$$\lim_{n\to\infty} \inf_{\pi\in\prod(P,Q)} I_n(\pi) \geq \inf_{\pi\in\prod(P,Q)} I(\pi).$$

Razonando como en la demostración del teorema 3.1.1, se ve que existe una sucesión de probabilidades $\{\pi_n\}_{n=1}^{\infty} \subset \prod(P,Q)$ tal que

$$\inf_{\pi\in\prod(P,Q)} I_n(\pi) = I(\pi_n), \quad \forall n \in \mathbb{N}.$$

Además, la sucesión $\{\pi_n\}_{n=1}^{\infty}$ es ajustada, por estar contenida en $\prod(P,Q)$ (se probó con detalle en la demostración de 3.1.1). Entonces, existe una subsucesión $\{\pi_{n_k}\}_{k=1}^{\infty}$ que converge hacia una probabilidad $\pi_* \in \prod(P,Q)$. Por el teorema de Portmanteau (ver C.0.2),

$$\lim_{k\to\infty} \int c_m d\pi_{n_k} = \int c_m d\pi_*, \quad \forall m \in \mathbb{N}.$$

Además, si $m < n$ entonces, $c_m \leq c_n$ y, se tiene que $I_m(\pi_n) \leq I_n(\pi_n)$. Juntando estas observaciones, se llega a que

$$\lim_{n\to\infty} I_n(\pi_n) \geq \limsup_{n\to\infty} I_m(\pi_n) \geq I_m(\pi_*), \quad \forall m \in \mathbb{N}.$$

Por el teorema de la convergencia monótona, $\lim_{m\to\infty} I_m(\pi_*) = I(\pi_*)$ Entonces,

$$\lim_{n\to\infty} I_n(\pi_n) \geq I(\pi_*) \geq \inf_{\pi\in\prod(P,Q)} I(\pi).$$

Con esto, queda probada la igualdad 3.11. Por el paso 1, para todo $n \in \mathbb{N}$, se tiene que

$$\begin{aligned}\inf_{\pi\in\prod(P,Q)} I_n(\pi) &= \sup\Big\{ \int_{\mathbb{R}^d} f(x)dP + \int_{\mathbb{R}^d} g(y)dQ : f(x)+g(y) \leq c_n(x,y)\Big\} \\ &\leq \sup\Big\{ \int_{\mathbb{R}^d} f(x)dP + \int_{\mathbb{R}^d} g(y)dQ : f(x)+g(y) \leq c(x,y)\Big\}\end{aligned}$$

Entonces, tomando superior en $n \in \mathbb{N}$,

$$\inf_{\pi \in \prod(P,Q)} I(\pi) \leq \sup \Big\{ \int_{\mathbb{R}^d} f(x)dP + \int_{\mathbb{R}^d} g(y)dQ : f(x) + g(y) \leq c(x,y) \Big\}.$$

La desigualdad contraria es evidente. □

Gracias al resultado de dualidad de Kantorovich, el problema de transporte óptimo asociado al coste cuadrático se reduce a calcular el inferior de $\int_{\mathbb{R}^d} \varphi dP + \int_{\mathbb{R}^d} \psi dQ$, en el conjunto de pares de funciones integrables respecto de P y Q, respectivamente, tales que $x \cdot y \leq \varphi(x) + \psi(y)$, para todo $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$. Se comprueba realizando los cálculos:

Proposición 3.2.2. *Si P y Q son dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos, entonces*

$$\max_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d} (x \cdot y) d\pi(x,y) = \inf_{(\varphi,\psi) \in \Phi} \left[\int_{\mathbb{R}^d} \varphi(x) dP(x) + \int_{\mathbb{R}^d} \psi(y) dQ(y) \right], \tag{3.12}$$

donde $\Phi := \{(\varphi,\psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y) \quad \forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d\}$.

Demostración. Por un lado,

$$\begin{aligned}
&\min_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 d\pi(x,y) = \min_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \Big(\|x\|^2 + \|y\|^2 - 2(x \cdot y) \Big) d\pi(x,y) \\
&= \min_{\pi \in \prod(P,Q)} \left[\int_{\mathbb{R}^d} \|x\|^2 dP(x) + \int_{\mathbb{R}^d} \|y\|^2 dQ(y) - 2 \int_{\mathbb{R}^d} (x \cdot y) d\pi(x,y) \right] \\
&= \int_{\mathbb{R}^d} \|x\|^2 dP(x) + \int_{\mathbb{R}^d} \|y\|^2 dQ(y) - 2 \max_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d} (x \cdot y) d\pi(x,y).
\end{aligned} \tag{3.13}$$

Por otro lado, sean $(f,g) \in L^1(P) \times L^1(Q)$ tales que $f(x) + g(y) \leq \|x-y\|^2$. Se escribe

$$\begin{aligned}
f(x) = \|x\|^2 - 2\varphi(x) \Leftrightarrow \varphi(x) = \frac{1}{2}(\|x\|^2 - f(x)), \\
g(y) = \|y\|^2 - 2\psi(y) \Leftrightarrow \psi(y) = \frac{1}{2}(\|y\|^2 - g(y)).
\end{aligned}$$

De aquí se deduce que $f(x) + g(y) \leq \|x-y\|^2 \Leftrightarrow x \cdot y \leq \varphi(x) + \psi(y)$.

Si $\Phi = \{(\varphi,\psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y),\ \forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d\}$. Por dualidad de Kantorovich,

$$(3.13) = \sup_{(\varphi,\psi) \in \Phi} \left[\int_{\mathbb{R}^d} \|x\|^2 dP(x) + \int_{\mathbb{R}^d} \|y\|^2 dQ(y) - 2 \Big(\int_{\mathbb{R}^d} \varphi(x) dP(x) + \int_{\mathbb{R}^d} \psi(y) dQ(y) \Big) \right]$$

si, y solo si,

$$\max_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d} (x \cdot y) d\pi(x,y) = \inf_{(\varphi,\psi) \in \Phi} \left[\int_{\mathbb{R}^d} \varphi(x) dP(x) + \int_{\mathbb{R}^d} \psi(y) dQ(y) \right].$$

□

3.2.1. Convexidad y Transporte Óptimo

Si $(\varphi, \psi) \in \Phi$, entonces $x \cdot y \leq \varphi(x) + \psi(y)$ para todo $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. Por eso, para cualquier probabilidad $\pi \in \prod(P, Q)$, se verifica que

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \Big(\varphi(x) + \psi(y)\Big) d\pi(x, y) \geq \int_{\mathbb{R}^d \times \mathbb{R}^d} x \cdot y \, d\pi(x, y).$$

Ya se ha probado que se alcanza el máximo de la expresión de la derecha, en una probabilidad $\pi_0 \in \prod(P, Q)$. Si también se alcanzase el mínimo de la expresión de la izquierda para un par concreto de funciones $(\varphi_0, \psi_0) \in \Phi$, entonces, por 3.12,

$$\int (\varphi_0(x) + \psi_0(y) - x \cdot y) d\pi_0(x, y) = 0 \Leftrightarrow \varphi_0(x) + \psi_0(y) - x \cdot y = 0 \quad \pi_0 - c.s.$$

Entonces, en el par óptimo (φ_0, ψ_0),

$$\psi_0(y) = \sup_{x \in \mathbb{R}^d} (x \cdot y - \varphi_0(x)) \quad \pi_0 - c.s.$$

$$\varphi_0(x) = \sup_{y \in \mathbb{R}^d} (x \cdot y - \psi_0(y)), \quad \pi_0 - c.s.$$

Esto es lo que se denomina un par de funciones convexas conjugadas. Los resultados de esta sección prueban que existe el mínimo de

$$\int_{\mathbb{R}^d} \varphi(x) dP(x) + \int_{\mathbb{R}^d} \psi(y) dQ(y) \tag{3.14}$$

en el conjunto de pares de funciones

$$\Phi := \{(\varphi, \psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y) \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d\}.$$

Además, se va a formalizar la idea que se acaba de explicar: el mínimo se alcanzará en un par de funciones convexas conjugadas.

Si $(\varphi, \psi) \in \Phi$, en lo que sigue se denota por

$$J(\varphi, \psi) := \int_{\mathbb{R}^d} \varphi(x) dP(x) + \int_{\mathbb{R}^d} \psi(y) dQ(y).$$

Definición 3.2.3. *Se define la convexa conjugada de la función $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ como la función*

$$\varphi^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}, \quad \varphi^*(y) = \sup_{x \in \mathbb{R}^d} [x \cdot y - \varphi(x)].$$

Se dice que φ^* es la transformada de Legendre de φ. Se ha ilustrado la importancia de las funciones que se acaban de definir en el problema del transporte óptimo. Para estudiar sus propiedades, se necesitan algunos resultados sobre funciones convexas.

Definición 3.2.4. *Sea $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ una función convexa. Para cada punto $x \in \mathbb{R}^d$, se define la subdiferencial de φ en x como el conjunto de puntos $y \in \mathbb{R}^d$ que verifican que*

$$\varphi(z) \geq \varphi(x) + y \cdot (z - x), \quad \forall z \in \mathbb{R}^d. \tag{3.15}$$

Se denota por $\partial\varphi(x)$.

La subdiferencial es una generalización del gradiente. Intuitivamente, es claro que si una función convexa es diferenciable en un punto, el gradiente es el único vector que va a verificar la condición 3.15. Se prueba esta idea en los dos siguientes lemas.

Lema 3.2.5. *Sea $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ una función convexa. El dominio de φ es el conjunto de puntos donde φ toma valores reales. Para todo $x \in Int(Dom(\varphi))$, la subdiferencial $\partial\varphi(x)$ es un conjunto no vacío.*

Demostración. La demostración se basa en la segunda forma geométrica del teorema de Hahn-Banach, que permite separar dos convexos no vacíos disjuntos si uno es cerrado y otro compacto, a partir de un hiperplano dado por un funcional lineal.

Sea $x_0 \in \text{Int}(\text{Dom}\varphi)$, existe $r > 0$ tal que $\bar{B}(x_0, r) \in \text{Int}(\text{Dom}\varphi)$. Se define

$$\mathcal{E}(\varphi) \;=\; \big\{(x,t) \in \bar{B}(x_0,r) \times \mathbb{R} : t \geq \varphi(x)\big\}.$$

$\mathcal{E}(\varphi)$ es un conjunto convexo y cerrado. Se comprueba usando que las funciones convexas son semicontinuas inferiormente en los puntos del interior del dominio (ver D.1.1).

Como $\varphi(x_0) \neq +\infty$, para cualquier $\varepsilon > 0$ se tiene que $z := (x_0, \varphi(x_0) - \varepsilon) \notin \mathcal{E}(\varphi)$. Se aplica la segunda forma geométrica del teorema de Hahn-Banach (ver D.2.1) a los conjuntos $\{z\}$ y $\mathcal{E}(\varphi)$. Entonces, existe un funcional lineal $L : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ y $\gamma \in \mathbb{R}$ tal que

$$L(x,t) < \gamma \leq L(z), \quad \forall\, (x,t) \in \mathcal{E}(\varphi). \tag{3.16}$$

Por el teorema de representación de Riesz, existe $\alpha \in \mathbb{R}^d$ y $\beta \in \mathbb{R}$ tal que

$$L(x,t) = \langle \tilde{\alpha}, x\rangle + \beta\, t.$$

$\beta \neq 0$, pues si no, no se verificaría la desigualdad para el punto $(x_0, \varphi(x_0)) \in \mathcal{E}(\varphi)$. Como los puntos de la forma $(x_0, t) \in \mathcal{E}(\varphi)$ para valores arbitrariamente grandes de t, se deduce que $\beta < 0$.

El funcional $\frac{1}{|\beta|}L(x,t) = \langle \alpha, x\rangle - t$, con $\alpha = \frac{\tilde{\alpha}}{|\beta|}$, seguirá verificando 3.16. Por lo tanto, se tienen las siguientes desigualdades:

$$\gamma \leq \langle \alpha, x_0\rangle - (\varphi(x_0) - \varepsilon) \quad \text{y} \quad \langle \alpha, x\rangle - t < \gamma, \quad \forall\, (x,t) \in \mathcal{E}(\varphi).$$

De la primera desigualdad, se obtiene $\varphi(x_0) - \langle \alpha, x_0\rangle - \varepsilon \leq -\gamma$. Evaluando la segunda desigualdad en los puntos $(x, \varphi(x)) \in \mathcal{E}(\varphi)$, con $x \in \bar{B}(x_0, r)$, se tiene que $\varphi(x) \geq \langle \alpha, x\rangle - \gamma$. Por lo tanto, se concluye que

$$\varphi(x) \geq \langle \alpha, x\rangle + \varphi(x_0) - \langle \alpha, x_0\rangle - \varepsilon = \varphi(x_0) + \langle \alpha, x - x_0\rangle - \varepsilon, \quad \forall x \in \bar{B}(x_0, r).$$

Haciendo tender ε a 0, se tiene que

$$\varphi(x) \geq \varphi(x_0) + \langle \alpha, x - x_0\rangle, \quad \forall x \in \bar{B}(x_0, r).$$

Si esta desigualdad se verifica para los puntos de $\bar{B}(x_0, r)$, se verifica para cualquier punto de $\mathbb{R}^d$. Se razona por reducción al absurdo: Sea $x \in \mathbb{R}^d$ tal que

$$\varphi(x) < \varphi(x_0) + \langle \alpha, x - x_0\rangle.$$

Entonces, el segmento que une $\varphi(x)$ con $\varphi(x_0)$ está por debajo del hiperplano $\varphi(x_0)+\langle\alpha, x-x_0\rangle$. Esto se puede ver multiplicando la desigualdad anterior por t y operando:

$$t\varphi(x) + (1-t)\varphi(x_0) < \varphi(x_0) + \langle\alpha, tx + (1-t)x_0 - x_0\rangle, \quad \forall t \in (0,1).$$

Se llega a una contradicción, puesto que para algún t suficientemente pequeño,

$$tx + (1-t)x_0 \in \bar{B}(x_0, r),$$

$$\varphi(tx + (1-t)x_0) \leq t\varphi(x) + (1-t)\varphi(x_0) < \varphi(x_0) + \langle\alpha, tx + (1-t)x_0 - x_0\rangle.$$

Esto es absurdo. Por lo tanto,

$$\varphi(x) \geq \varphi(x_0) + \langle\alpha, x - x_0\rangle, \quad \forall x \in \mathbb{R}^d.$$

Se tiene que $\alpha \in \partial\varphi(x_0)$. □

Lema 3.2.6. *Sea $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ una función convexa. Si φ es diferenciable en el punto $x \in \mathbb{R}^d$, entonces el conjunto $\partial\varphi(x)$ está formado por un único punto, el gradiente de φ en x, es decir,*

$$\partial\varphi(x) = \{\nabla\varphi(x)\}.$$

Demostración. Primero, se prueba que $\nabla\varphi(x) \in \partial\varphi(x)$. Por convexidad, para todo $y \in \mathbb{R}^d$, se tiene que

$$\varphi\big(x + t(y-x)\big) \;\leq\; (1-t)\,\varphi(x) + t\,\varphi(y), \quad \forall t \in (0,1).$$

Reescribiéndolo,

$$\frac{\varphi\big(x + t(y-x)\big) - \varphi(x)}{t} \;\leq\; \varphi(y) - \varphi(x).$$

El límite de la expresión de la izquierda, cuando t tiende a 0, es la derivada direccional de φ en x en la dirección $y-x$, y se denota por $D\varphi(x;\, y-x)$. Como φ es diferenciable en x, ese límite existe y se cumple

$$D\varphi(x;\, y-x) \;=\; \big\langle\nabla\varphi(x),\, y-x\big\rangle, \quad \forall y \in \mathbb{R}^d.$$

Entonces,

$$\big\langle\nabla\varphi(x),\, y-x\big\rangle \;\leq\; \varphi(y) - \varphi(x), \quad \forall y \in \mathbb{R}^d.$$

Se concluye que $\nabla\varphi(x) \in \partial\varphi(x)$.

Queda probar que si otro punto $\alpha \in \mathbb{R}^d$ pertenece al gradiente de φ en x, entonces $\alpha = \nabla\varphi(x)$. Si $\alpha \in \partial\varphi(x)$, entonces

$$\varphi(y) \;\geq\; \varphi(x) + \langle\alpha,\, y-x\rangle, \quad \forall\, y \in \mathbb{R}^d.$$

En particular, si $v \in \mathbb{R}^d$ y $t > 0$, tomando $y = x + tv$ en la desigualdad anterior, se tiene que

$$\frac{\varphi(x+tv) - \varphi(x)}{t} \;\geq\; \langle\alpha, v\rangle, \quad \forall\, v \in \mathbb{R}^d,\; t > 0.$$

Tomando límite cuando t tiende a 0, se deduce que

$$\langle\nabla\varphi(x),\, v\rangle \;\geq\; \langle\alpha,\, v\rangle, \quad \forall\, v \in \mathbb{R}^d \Leftrightarrow \langle\nabla\varphi(x) - \alpha,\, v\rangle \geq 0, \quad \forall\, v \in \mathbb{R}^d.$$

Tomando $v = \alpha - \nabla\varphi(x)$, se tiene que $-\|\nabla\varphi(x) - \alpha\|^2 \geq 0$ si, y solo si, $\nabla\varphi(x) = \alpha$. □

Se enuncian, en la siguiente proposición, propiedades básicas de la convexa conjugada de una función. En el apéndice D.1 se define una función inferiormente semicontinua (ver definición D.1.1).

Proposición 3.2.7. *Sea $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Entonces su transformada de Legendre φ^* (ver definición en 3.2.3) es convexa e inferiormente semi-continua. Además, si φ no es idénticamente $+\infty$, entonces φ^* tampoco es idénticamente $+\infty$.*

Demostración. Primero se prueba que φ^* es convexa. Para cada $x \in \mathbb{R}^d$, se define la función afín $\tilde{\varphi}_x(y) := x \cdot y - \varphi(x)$. Entonces, sean $y_1, y_2 \in \mathbb{R}^d$,

$$\tilde{\varphi}_x(ty_1 + (1-t)y_2) = t\tilde{\varphi}_x(y_1) + (1-t)\tilde{\varphi}_x(y_2), \quad \forall t \in (0,1).$$

Tomando superior en $x \in \mathbb{R}^d$,

$$\begin{aligned}\varphi^*(y) &:= \sup_{x\in\mathbb{R}^d} \tilde{\varphi}_x(y) = \sup_{x\in\mathbb{R}^d} \left[t\tilde{\varphi}_x(y_1) + (1-t)\tilde{\varphi}_x(y_2)\right] \\ &\leq t \sup_{x\in\mathbb{R}^d} \tilde{\varphi}_x(y_1) + (1-t) \sup_{x\in\mathbb{R}^d} \tilde{\varphi}_x(y_2) =: t\varphi^*(y_1) + (1-t)\varphi^*(y_2).\end{aligned}$$

El segundo paso es probar que φ^* es inferiormente semicontinua. Para cada $y_0 \in \mathbb{R}^d$, sea $\{y_n\}_{n=1}^{\infty}$ una sucesión tal que $\lim_{n\to\infty} y_n = y_0$. Entonces, $\lim_{n\to\infty} x \cdot y_n - \varphi(x) = x \cdot y_0 - \varphi(x)$, para cada $x \in \mathbb{R}^d$. Como $\varphi^*(y_n) \geq x \cdot y_n - \varphi(x)$, para todo $x \in \mathbb{R}^d$, entonces

$$\liminf_{n\to\infty} \varphi^*(y_n) \geq x \cdot y_0 - \varphi(x), \quad \forall x \in \mathbb{R}^d.$$

Por lo tanto, $\liminf_{n\to\infty} \varphi^*(y_n) \geq \sup_{x\in\mathbb{R}^d} x \cdot y_0 - \varphi(x) =: \varphi^*(y_0)$.

Por último, si φ no es idénticamente $+\infty$, existe $x_0 \in \mathbb{R}^d$ tal que $\varphi(x_0) < +\infty$. Como se ha visto en el lema 3.2.5, la subdiferencial de φ en x_0 es no vacía, es decir, existe $y \in \partial\varphi(x_0)$. Por definición de subdiferencial,

$$\varphi(x) \geq \varphi(x_0) + \langle y, x - x_0\rangle, \quad \forall x \in \mathbb{R}^d.$$

si, y solo si,

$$\langle y, x\rangle - \varphi(x) \leq \langle y, x_0\rangle - \varphi(x_0), \quad \forall x \in \mathbb{R}^d.$$

Entonces, tomando superior en x,

$$\varphi^*(y) = \sup_x\{\langle y, x\rangle - \varphi(x)\} \leq \langle y, x_0\rangle - \varphi(x_0) < +\infty.$$

□

A continuación, se da una caracterización de la subdiferencial de una función a partir de su convexa conjugada.

Proposición 3.2.8. *Sea $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ una función convexa e inferiormente semicontinua no idénticamente $+\infty$. Entonces, para todo $x, y \in \mathbb{R}^d$,*

$$x \cdot y = \varphi(x) + \varphi^*(y) \Leftrightarrow y \in \partial\varphi(x).$$

Demostración. De la definición de convexa conjugada es fácil ver que se cumple que $x \cdot y \leq \varphi(x) + \varphi^*(y)$, para todo $x, y \in \mathbb{R}^d$. Por lo tanto,

$$x \cdot y = \varphi(x) + \varphi^*(y) \Leftrightarrow x \cdot y \geq \varphi(x) + \varphi^*(y).$$

Como $\varphi^*(y) = \sup_{z \in \mathbb{R}^d} [z \cdot y - \varphi(z)]$,

$$\begin{aligned} x \cdot y \geq \varphi(x) + \varphi^*(y) &\Leftrightarrow x \cdot y \geq \varphi(x) + z \cdot y - \varphi(z), && \forall z \in \mathbb{R}^d \\ &\Leftrightarrow \varphi(z) \geq \varphi(x) + y \cdot (z - x), && \forall z \in \mathbb{R}^d \\ &\Leftrightarrow y \in \partial\varphi(x). \end{aligned}$$

□

El objetivo del siguiente resultado es llegar a una caracterización de las funciones convexas conjugadas φ^*, será clave para el estudio del problema dual de transporte óptimo. Se ha visto, en la proposición 3.2.7, que la convexa conjugada de una función φ es convexa e inferiormente semicontinua. Se prueba, a continuación, el recíproco.

Proposición 3.2.9. *Sea $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ una función no idénticamente $+\infty$. Son equivalentes:*

1. *φ es convexa e inferiormente semi-continua.*
2. *Existe una función $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, no idénticamente $+\infty$, tal que $\varphi = \psi^*$.*
3. *$\varphi^{**} = \varphi$.*

Demostración. 3) $\Rightarrow$ 2) es claro, tomando $\psi := \varphi^*$, que no es idénticamente $+\infty$, como se ha visto en la proposición 3.2.7. 2) $\Rightarrow$ 1) está ya probado en 3.2.7. Queda probar 1) $\Rightarrow$ 3).

Sea φ convexa e inferiormente semicontinua. La desigualdad $\varphi(x) \geq \varphi^{**}(x)$, se da siempre. Para todo $y \in \mathbb{R}^d$, $x \cdot y - \varphi(x) \leq \sup_{x \in \mathbb{R}^d} [x \cdot y - \varphi(x)] =: \varphi^*(y)$. Entonces,

$$\varphi^{**}(x) := \sup_{y \in \mathbb{R}^d} [x \cdot y - \varphi^*(y)] \leq \varphi(x), \quad \forall x \in \mathbb{R}^d.$$

Caso 1: Se supone que $\varphi \geq 0$. Para probar que $\varphi \leq \varphi^{**}$ se razona por reducción al absurdo. Se supone que existe $x_0 \in \mathbb{R}^d$ tal que

$$\varphi(x_0) > \varphi^{**}(x_0).$$

Como φ es inferiormente semi-continua, el conjunto convexo

$$\text{epi}(\varphi) \ = \ \big\{(x,t) \in \mathbb{R}^d \times \mathbb{R} : t \geq \varphi(x)\big\}$$

es cerrado. Se aplica la segunda forma geométrica del teorema de Hahn Banach (ver D.2.1) a los conjuntos epi(φ) y $\{(x_0, \varphi^{**}(x_0))\}$. Existe un funcional lineal $L : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ y $\gamma \in \mathbb{R}$ tal que

$$L(x_0, \varphi^{**}(x_0)) < \gamma \leq L(x,t), \quad \forall\, (x,t) \in \text{epi}(\varphi). \tag{3.17}$$

Por el teorema de representación de Riesz, existe $\alpha \in \mathbb{R}^d$ y $\beta \in \mathbb{R}$ tal que

$$L(x,t) = \langle \alpha, x \rangle + \beta\, t.$$

Existe un punto $x \in \mathrm{Dom}\varphi$, por ser φ propia. Como los puntos de la forma $(x,t) \in \mathrm{epi}(\varphi)$ para valores arbitrariamente grandes de t, se deduce que $\beta \geq 0$.

Para cualquier $\varepsilon > 0$, como $\varphi \geq 0$, se verifica:

$$\langle \alpha, x \rangle + (\beta + \varepsilon)\varphi(x) \geq \langle \alpha, x \rangle + \beta\varphi(x) \geq \gamma, \quad \forall x \in \mathbb{R}^d$$

si, y solo si,

$$\langle -\frac{\alpha}{\beta+\varepsilon}, x \rangle - \varphi(x) \leq -\frac{\gamma}{(\beta+\varepsilon)}, \quad \forall x \in \mathbb{R}^d,$$

de donde se deduce que

$$\varphi^*(-\frac{\alpha}{\beta+\varepsilon}) \leq -\frac{\gamma}{(\beta+\varepsilon)}.$$

Por lo tanto,

$$\varphi^{**}(x_0) \geq \langle -\frac{\alpha}{\beta+\varepsilon}, x_0 \rangle - \varphi^*(-\frac{\alpha}{\beta+\varepsilon}) \geq \langle -\frac{\alpha}{\beta+\varepsilon}, x_0 \rangle + \frac{\gamma}{(\beta+\varepsilon)}.$$

Se llega una contradicción, puesto que

$$\langle \alpha, x_0 \rangle + (\beta + \varepsilon)\varphi^{**}(x_0) \geq \gamma, \quad \forall \varepsilon > 0.$$

Caso general: Como φ no es idénticamente $+\infty$, entonces φ^* tampoco. Sea $y_0 \in \mathbb{R}^d$ tal que $\varphi^*(y_0) < +\infty$. Se considera la función

$$\bar{\varphi}(x) = \varphi(x) - \langle x, y_0 \rangle + \varphi^*(y_0), \quad \forall x \in \mathbb{R}^d.$$

$\bar{\varphi}$ es convexa, inferiormente semicontinua y mayor o igual que 0. Entonces, por el caso 1, se deduce que $\bar{\varphi} = \bar{\varphi}^{**}$. Desarrollando primero la expresión de φ^*, se tiene que

$$\begin{aligned}
\bar{\varphi}^*(y) &= \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - \varphi(x) + \langle x, y_0 \rangle - \varphi^*(y_0)\} \\
&= \sup_{x \in \mathbb{R}^d} \{\langle x, y + y_0 \rangle - \varphi(x)\} - \varphi^*(y_0) \\
&= \varphi^*(y + y_0) - \varphi^*(y_0), \quad \forall y \in \mathbb{R}^d.
\end{aligned}$$

En cuanto a φ^{**}, se tiene que:

$$\begin{aligned}
\bar{\varphi}^{**}(x) &= \sup_{y \in \mathbb{R}^d} \{\langle x, y \rangle - \varphi^*(y + y_0)\} + \varphi^*(y_0) \\
&= \sup_{y \in \mathbb{R}^d} \{\langle x, y + y_0 \rangle - \varphi^*(y + y_0)\} - \langle x, y_0 \rangle + \varphi^*(y_0) \\
&= \varphi^{**}(x) - \langle x, y_0 \rangle + \varphi^*(y_0).
\end{aligned}$$

Se concluye que $\varphi^{**} = \varphi$. □

3.2.2. Estudio del problema dual

En el contexto del problema de transporte óptimo que se ha planteado en esta sección, se ha estudiado la expresión $\int \varphi dP + \int \psi dQ$, y se ha visto que la clave para minimizar esa suma son los pares de funciones convexas conjugadas. Por eso, en el apartado anterior, se estudiaron las propiedades de la transformada de Legendre. Ahora ya se dispone de las herramientas necesarias para probar que existe un par de funciones óptimas en el problema dual y que son inferiormente semicontinuas conjugadas.

Lema 3.2.10 (Doble convexificación)**.** *Sean P y Q dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos. Sea*

$$\Phi = \{(\varphi, \psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y), \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d\}$$

Para cada $(\varphi, \psi) \in \Phi$, existe $a \in \mathbb{R}$ tal que el par de funciones de Φ definido de la siguiente forma:

$$(\bar{\varphi}, \bar{\psi}) := (\varphi^{**} - a, \varphi^* + a)$$

verifica que $J(\bar{\varphi}, \bar{\psi}) \leq J(\varphi, \psi)$ y satisface, para todo $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\bar{\varphi}(x) \geq -\frac{\|x\|^2}{2}, \quad \bar{\psi}(y) \geq -\frac{\|y\|^2}{2}. \tag{3.18}$$

Demostración. Sea $(\varphi, \psi) \in \Phi$, es fácil comprobar que el par $(\varphi^{**}, \varphi^*)$ verifica que:

$$\varphi^{**}(x) + \varphi^*(y) \geq x \cdot y, \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

Además, $J(\varphi^{**}, \varphi^*) \leq J(\varphi, \psi)$. Esta afirmación se demuestra a continuación: Para todo $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, $x \cdot y \leq \varphi(x) + \psi(y)$. Se deduce que

$$\varphi^*(y) := \sup_{x \in \mathbb{R}^d} [x \cdot y - \varphi(x)] \leq \psi(y).$$

Por otro lado, siguiendo el razonamiento que ya se ha hecho en otras demostraciones, se verifica que $x \cdot y - \varphi(x) \leq \sup_{x \in \mathbb{R}^d} [x \cdot y - \varphi(x)] =: \varphi^*(y)$. Entonces,

$$\varphi^{**}(x) := \sup_{y \in \mathbb{R}^d} [x \cdot y - \varphi^*(y)] \leq \varphi(x).$$

Por ser φ integrable respecto de P, no es idénticamente $+\infty$. Esto implica que $\varphi^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ tampoco es idénticamente $+\infty$ (por la proposición 3.2.7). Además, φ^* está acotada inferiormente por una función lineal: existe $x_0 \in \mathbb{R}^d$ tal que $\varphi(x_0) < +\infty$, entonces,

$$x_0 \cdot y + \varphi(x_0) \leq \varphi^*(y) \quad \forall y \in \mathbb{R}^d.$$

Sea $a = \inf_{y \in \mathbb{R}^d} [\varphi^*(y) + \frac{\|y\|^2}{2}]$. Como φ^* toma valores reales para algún punto, $a < +\infty$. Acotando la función que se quiere minimizar por otra que alcanza el mínimo absoluto, se ve que $a \in \mathbb{R}$:

$$\varphi^*(y) + \frac{\|y\|^2}{2} \geq x_0 \cdot y + \varphi(x_0) + \frac{\|y\|^2}{2} = \frac{\|y + x_0\|^2}{2} - \frac{\|x_0\|^2}{2} + \varphi(x_0), \quad \forall y \in \mathbb{R}^d.$$

Se considera el par de funciones $(\bar{\varphi}, \bar{\psi}) := (\varphi^{**} + a, \varphi^* - a)$. Para todo $y \in \mathbb{R}^d$, $\varphi^*(y) + \frac{\|y\|^2}{2} - a \geq 0$, (por ser a el inferior de esa función). Entonces $\bar{\varphi}(y) \geq -\frac{\|y\|^2}{2}$, para todo $y \in \mathbb{R}^d$. En cuanto a la función $\bar{\psi}$, se tiene la siguiente cadena de desigualdades:

$$\bar{\varphi}(x) + \frac{\|x\|^2}{2} = \bar{\psi}^*(x) + \frac{\|x\|^2}{2} = \sup_{y \in \mathbb{R}^d} \Big[x \cdot y - \bar{\psi}(y) + \frac{\|x\|^2}{2} \Big]$$

$$\geq \sup_{y \in \mathbb{R}^d} \Big[-\bar{\psi}(y) - \frac{\|y\|^2}{2} \Big] = -\inf_{y \in \mathbb{R}^d} \Big[\bar{\psi}(y) + \frac{\|y\|^2}{2} \Big] = 0, \quad \forall x \in \mathbb{R}^d.$$

Para probar que $(\bar{\varphi}, \bar{\psi}) \in \Phi$, solo falta ver que $(\bar{\varphi}, \bar{\psi}) \in L^1(P) \times L^1(Q)$, porque ya se ha probado que

$$\bar{\varphi}(x) + \bar{\psi}(y) = \varphi^{**}(x) + \varphi^*(y) \geq x \cdot y, \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

Se cumple que $J(\bar{\varphi}, \bar{\psi}) = J(\varphi^{**}, \psi^*) \leq J(\varphi, \psi) < +\infty$. Entonces, se tiene que $\int \bar{\varphi} dP + \int \bar{\psi} dQ < +\infty$. Además, como ambas funciones ($\bar{\varphi}$ y $\bar{\psi}$) están acotadas inferiormente por funciones integrables ($-\frac{\|x\|^2}{2}$ y $-\frac{\|y\|^2}{2}$), se deduce que

$$(\bar{\varphi}, \bar{\psi}) \in L^1(P) \times L^1(Q).$$

□

De la demostración del lema 3.2.10, se deduce que si $(\varphi, \psi) \in \Phi$, entonces $(\varphi^{**}, \varphi^*) \in \Phi$ y se cumple que $J(\varphi^{**}, \varphi^*) \leq J(\varphi, \psi)$. Por lo tanto, para buscar el mínimo de $J(\varphi, \psi)$ en Φ hay que restringirse a los pares de funciones convexas conjugadas del tipo $(\varphi^{**}, \varphi^*)$.

Corolario 3.2.11. *Sean P y Q dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos. Minimizar la expresión $J(\varphi, \psi) := \int \varphi dP + \int \psi dQ$ en el conjunto*

$$\Phi = \{(\varphi, \psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y), \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d\}$$

es equivalente a calcular $\inf_{\varphi \in L^1(Q)} J(\varphi^{**}, \varphi^*)$.

A partir del corolario 3.2.11 y de la caracterización de las funciones convexas conjugadas dada en la proposición 3.2.9, se tiene el siguiente resultado:

Proposición 3.2.12. *Sean P y Q dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos. Minimizar la expresión $J(\varphi, \psi) := \int \varphi dP + \int \psi dQ$ en el conjunto*

$$\Phi = \{(\varphi, \psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y), \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d\}$$

es equivalente a calcular

$$\inf_{\varphi \in \mathcal{C}} J(\varphi, \varphi^*),$$

donde $\mathcal{C}$ es el conjunto de funciones $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ de $L^1(P)$ convexas e inferiormente semicontinuas, no idénticamente $+\infty$.

Lo único que queda por demostrar es que se alcanza el mínimo de $J(\varphi, \psi)$ en un par de funciones concretas, esto es lo que se prueba en el siguiente teorema.

Teorema 3.2.13. *Sean P y Q dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos. Sea $\Phi = \{(\varphi,\psi) \in L^1(P)\times L^1(Q) : x\cdot y \leq \varphi(x)+\psi(y), \quad \forall (x,y)\in\mathbb{R}^d\times\mathbb{R}^d\}$. Entonces, existe un par (φ,φ^*) de funciones convexas e inferiormente semicontinuas conjugadas propias definidas en $\mathbb{R}^d$ tal que*

$$\inf_{(\varphi,\psi)\in\Phi} J(\varphi,\psi) = J(\varphi,\varphi^*).$$

Demostración. Sea $\{(\varphi_k,\psi_k)\}_{k=1}^{\infty}$ una sucesión de elementos de Φ tal que

$$\lim_{k\to\infty} J(\varphi_k,\psi_k) = \inf_{(\varphi,\psi)\in\Phi} J(\varphi,\psi).$$

Por el lema 3.2.10, se puede suponer que

$$0 \leq \varphi_k(x) + \frac{\|x\|^2}{2}, \quad 0 \leq \psi_k(y) + \frac{\|y\|^2}{2}.$$

Buscando la convergencia débil de alguna subsucesión de $\{(\varphi_k,\psi_k)\}_{k=1}^{\infty}$, se truncan las funciones para conseguir una acotación por arriba (y garantizar la integrabilidad uniforme). En el apéndice E, se definen estos conceptos y se explican las propiedades que se van a usar a continuación.

Para cada $n\in\mathbb{N}$, se define

$$\varphi_k^n(x) = \text{mín}\{\varphi_k(x)+\frac{\|x\|^2}{2}, n\} - \frac{\|x\|^2}{2}, \quad \psi_k^n(y) = \text{mín}\{\psi_k(y)+\frac{\|y\|^2}{2}, n\} - \frac{\|y\|^2}{2}.$$

Se tiene, para todo $k\in\mathbb{N}$,

$$\left\|\varphi_k^n + \frac{\|x\|^2}{2}\right\|_{L^1(P)} \leq n, \quad \left\|\psi_k^n + \frac{\|y\|^2}{2}\right\|_{L^1(Q)} \leq n.$$

Además, fijado $\varepsilon>0$, si $E\in\beta^d$ tal que $P(E) < \frac{\varepsilon}{n}$

$$\int_E \left|\varphi_k^n(x) + \frac{\|x\|^2}{2}\right| dP \leq nP(E) < \varepsilon, \quad \forall k\in\mathbb{N}.$$

Análogamente, si $F\in\beta^d$ tal que $Q(F) < \frac{\varepsilon}{n}$

$$\int_F \left|\psi_k^n(y) + \frac{\|y\|^2}{2}\right| dQ \leq nQ(E) < \varepsilon, \quad \forall k\in\mathbb{N}.$$

Se acaba de probar que, para cada $n\in\mathbb{N}$ fijo, las sucesiones $\{\varphi_k^n + \frac{\|x\|^2}{2}\}_{k=1}^{\infty}$ y $\{\psi_k^n + \frac{\|y\|^2}{2}\}_{k=1}^{\infty}$ verifican la condición de integrabilidad uniforme (ver E.1). Por lo tanto, por el teorema E.1.2, para cada $n\in\mathbb{N}$ fijo, existe una subsucesión de índices $\{k_j\}_{j=1}^{\infty}$ tales que $\{\varphi_{k_j}^n + \frac{\|x\|^2}{2}\}_{j=1}^{\infty}$ converge débilmente en $L^1(P)$ (equivalentemente $\{\varphi_{k_j}^n\}_{j=1}^{\infty}$ converge débilmente) y otra subsucesión de índices $\{k_i\}_{i=1}^{\infty}$ tales que $\{\psi_{k_i}^n + \frac{\|y\|^2}{2}\}_{i=1}^{\infty}$ converge débilmente en $L^1(Q)$ (equivalentemente $\{\psi_{k_i}^n\}_{i=1}^{\infty}$ converge débilmente). Se denota al límite por $(\varphi^n,\psi^n)\in L^1(P)\times L^1(Q)$. Sin pérdida de generalidad, se supone que es la misma subsucesión de índices.

Si se extraen las subsucesiones de índices de forma recursiva, (es decir, para cada $n \in \mathbb{N} \setminus \{1\}$ se extrae una nueva subsucesión de la subsucesión que se había extraído para $n-1$), se tiene que, para cada $n \in \mathbb{N}$, existe una subsucesión de índices $\{k_l\}_{l=1}^{\infty}$ tales que los pares de funciones $(\varphi_{k_l}^n, \psi_{k_l}^n)$ convergen $\forall m \leq n$. Al formar la sucesión de índices quedándose con el elemento n-ésimo de la subsucesión que elegida para n, se construye una subsucesión de $\{(\varphi_k, \psi_k)\}_{k=1}^{\infty}$ cuyas truncaciones convergen, para cualquier $n \in \mathbb{N}$. Para simplificar la notación, se supone que es la propia sucesión la que cumple esta propiedad.

La convergencia débil implica, en particular, que

$$\lim_{k\to\infty} \int \varphi_k^n(x) dP(x) = \int \varphi^n(x) dP(x), \quad \lim_{k\to\infty} \int \psi_k^n(y) dQ(y) = \int \psi^n(y) dQ(y).$$

Por lo tanto,

$$J(\varphi^n, \psi^n) = \lim_{k\to\infty} J(\varphi_k^n, \psi_k^n) \leq \lim_{k\to\infty} J(\varphi_k, \psi_k) = \inf_{(\varphi,\psi)\in\Phi} J(\varphi, \psi), \quad \forall n \in \mathbb{N}.$$

La última desigualdad se debe a la monotonía de la integral, ya que, fijado $k \in \mathbb{N}$, $\varphi_k^n \leq \varphi_k$ y $\psi_k^n \leq \psi_k$, para todo $n \in \mathbb{N}$.

Para cada $k \in \mathbb{N}$ fijo, las sucesiones $\{\psi_k^n\}_{n=1}^{\infty}$ y $\{\varphi_k^n\}_{n=1}^{\infty}$ son crecientes. Tomando límites en la convergencia débil, la monotonía de estas sucesiones se traslada a las sucesiones $\{\psi^n\}_{n=1}^{\infty}$ y $\{\varphi^n\}_{n=1}^{\infty}$, aunque son desigualdades casi seguro (con respecto a las medidas P y Q, respectivamente). Esto se ha probado en E.0.3. Por lo tanto, existen los límites puntuales, definidos para casi todo punto, ya que la unión numerable de conjuntos de medida nula es de medida nula.

$$\varphi_0 := \lim_{n\to\infty} \varphi^n : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}, \quad \psi_0 := \lim_{n\to\infty} \psi^n : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}.$$

En los puntos de $\mathbb{R}^d$ que no se cumplen las desigualdades $\varphi^n(x) \leq \varphi^{n+1}(x)$ para todo $n \in \mathbb{N}$, se define $\varphi_0(x) = +\infty$. Análogamente, en los puntos que la sucesión $\{\psi^n(y)\}_n$ no es creciente, se define $\psi_0(y) = +\infty$.

El siguiente paso es comprobar que $(\varphi_0, \psi_0) \in \Phi$. Como $(\varphi_k, \psi_k) \in \Phi$ para todo $k \in \mathbb{N}$, se tiene que:

$$\varphi_k(x) + \frac{\|x\|^2}{2} + \psi_k(y) + \frac{\|y\|^2}{2} \geq \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} + x \cdot y = \frac{\|x+y\|^2}{2}.$$

Fijado $n \in \mathbb{N}$,

$$\varphi_k^n(x) + \psi_k^n(y) + \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} \geq \text{mín}\left\{n, \frac{\|x+y\|^2}{2}\right\}, \quad \forall k \in \mathbb{N}.$$

Tomando límite, cuando k tiende a ∞, y después, haciendo tender n a ∞:

$$\varphi^n(x) + \psi^n(y) + \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} \geq \text{mín}\left\{n, \frac{\|x+y\|^2}{2}\right\},$$

$$\varphi_0(x) + \psi_0(y) + \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} \geq \frac{\|x+y\|^2}{2} \Leftrightarrow \varphi_0(x) + \psi_0(y) \geq x \cdot y.$$

Por cómo se habían construido las funciones φ_0 y ψ_0, la última desigualdad se verifica para todo $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, (aunque fijado $n \in \mathbb{N}$, la respectiva desigualdad para (φ^n, ψ^n) pueda no verificarse para los x de un conjunto de probabilidad P nula, y los y de un conjunto de probabilidad Q nula).

Las sucesiones de funciones $\{\psi^n\}_{n=1}^{\infty}$ y $\{\varphi^n\}_{n=1}^{\infty}$ son crecientes y acotadas inferiormente por funciones integrables:

$$-\frac{\|x\|^2}{2} \leq \varphi^n(x), \quad -\frac{\|y\|^2}{2} \leq \psi^n(y).$$

Por el teorema de la convergencia monótona

$$\int \varphi_0 dP = \lim_{n\to\infty} \int \varphi^n dP, \quad \int \psi_0 dQ = \lim_{n\to\infty} \int \psi^n dQ.$$

Por lo tanto,

$$J(\varphi_0, \psi_0) = \lim_{n\to\infty} J(\varphi^n, \psi^n) \leq \inf_{(\varphi,\psi)\in\Phi} J(\varphi, \psi).$$

De esta cadena de desigualdades, también se deduce que $(\varphi_0, \psi_0) \in L^1(P) \times L^1(Q)$, puesto que $-\infty < -\int \frac{\|x\|^2}{2} dP \leq \int \varphi_0 dP$, $-\infty < -\int \frac{\|y\|^2}{2} dQ \leq +\int \psi_0 dQ$ y la suma de ambas integrales verifica que:

$$\int \varphi_0 dP + \int \varphi_0 dP < +\infty.$$

Entonces, se verifica la desigualdad contraria $J(\varphi_0, \psi_0) \geq \inf_{(\varphi,\psi)\in\Phi} J(\varphi, \psi)$ y, por lo tanto, la igualdad.

□

3.2.3. Caracterización del plan de transporte óptimo

Por último, para cerrar la sección, se prueban dos teoremas que caracterizan los planes de transporte óptimos, garantizando unicidad bajo ciertas condiciones.

Teorema 3.2.14 (Criterio de optimalidad)**.** *Sean P y Q dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos. La probabilidad $\pi \in \prod(P, Q)$ es un plan de transporte óptimo para el coste cuadrático $c(x, y) = \|x - y\|^2$ si, y solo si, existe una función $\varphi : \mathbb{R}^d \to \mathbb{R}$ convexa e inferiormente semicontinua tal que*

$$y \in \partial\varphi(x), \quad \pi - c.s.$$

Si esto ocurre, el par (φ, φ^*) minimiza la expresión

$$J(\varphi, \psi) := \int_{\mathbb{R}^d} \varphi dP + \int_{\mathbb{R}^d} \psi dQ$$

en el conjunto $\Phi = \{(\varphi, \psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y), \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d\}$

Demostración. Se ha probado en el teorema 3.2.13 que existe una función φ convexa e inferiormente semicontinua tal que el par (φ, φ^*) minimiza $J(\varphi, \psi)$ en el conjunto Φ. Gracias a la fórmula de dualidad 3.12, se tiene que una probabilidad $\pi \in \prod(P, Q)$ es un plan de transporte óptimo para el coste cuadrático si, y solo si,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} x \cdot y d\pi(x, y) = \int_{\mathbb{R}^d} \varphi dP + \int_{\mathbb{R}^d} \varphi^* dQ = \int_{\mathbb{R}^d \times \mathbb{R}^d} \Big[\varphi(x) + \varphi^*(y)\Big] d\pi(x, y),$$

si, y solo si,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \Big[\varphi(x) + \varphi^*(y) - x \cdot y\Big] d\pi(x, y) = 0.$$

Como $\varphi(x) + \varphi^*(y) - x \cdot y \geq 0$ para todo $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, que la integral sea nula es equivalente a que

$$\varphi(x) + \varphi^*(y) = x \cdot y, \quad \pi - c.s.$$

es decir, es equivalente a que $y \in \partial\varphi(x), \quad \pi - c.s.$ □

Se enuncia, a continuación, el teorema que garantiza que si P no da probabilidad a conjuntos de medida nula, entonces existe una aplicación de transporte óptimo de P a Q. Además, el plan de transporte inducido por esta aplicación es el único plan de transporte óptimo.

Lema 3.2.15. *Sean P y Q dos probabilidades en $\mathbb{R}^d$. Sea $T : \mathbb{R}^d \to \mathbb{R}^d$ una aplicación medible. Si $\pi \in \prod(P, Q)$, entonces son equivalentes:*

1. $\pi = (Id \times T)\#P$.
2. $y = T(x)$, $\pi - c.s.$

Además, si se verifica alguna de estas condiciones, entonces $T\#P = Q$.

Demostración. Si $\pi = (Id \times \nabla\varphi)\#P$, entonces

$$\pi(C) = P\Big(\{x \in \mathbb{R}^d : (x, T(x)) \in C\}\Big), \quad \forall C \in \beta^{2d}.$$

Con esta definición, es claro que el conjunto $\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : y \neq T(x)\}$ tiene probabilidad π nula.

Para probar la implicación contraria, se define la probabilidad R en $\mathbb{R}^d \times \mathbb{R}^d$ de la siguiente forma:

$$R(C) = P\Big(\{x \in \mathbb{R}^d : (x, T(x)) \in C\}\Big), \quad \forall C \in \beta^{2d}.$$

Para comprobar que $\pi = R$ basta con comprobar que $\pi(A \times B) = R(A \times B)$ para todo $A, B \in \beta^d$, ya que los conjuntos de este tipo son una clase determinante de la probabilidad. Si se supone que $y = T(x)$, $\pi - c.s.$, entonces

$$\begin{aligned} R(A \times B) &= P\Big(\{x \in \mathbb{R}^d : (x, T(x)) \in A \times B\}\Big) = P\Big(A \cap T^{-1}(B)\Big) \\ &= \pi\Big(A \cap T^{-1}(B) \times \mathbb{R}^d\Big) = \pi\Big(\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : x \in A, T(x) \in B\}\Big) \\ &= \pi(A \times B). \end{aligned}$$

Por último, se comprueba que si $\pi = (Id \times T)\#P$, entonces $T\#P = Q$ ya que:

$$Q(B) = \pi(\mathbb{R}^d \times B) = P\Big(\{x \in \mathbb{R}^d : T(x) \in B\}\Big) = P(T^{-1}(B)), \quad \forall B \in \beta^d.$$

□

Teorema 3.2.16. *Sean P y Q dos probabilidades en $\mathbb{R}^d$ con momentos de orden 2 finitos. Si P no da masa a conjuntos de medida nula, entonces el plan de transporte óptimo π, para el problema del coste cuadrático, es único. Y, además, se verifica que*

$$\pi = (Id \times \nabla\varphi)\#P,$$

donde $\nabla\varphi$ es el único (salvo un conjunto de P-probabilidad nula) gradiente de una función convexa tal que $\nabla\varphi\#P = Q$.

Demostración. Primero, se probará que cualquier plan de transporte óptimo π cumple las condiciones descritas en el teorema. Sea φ la función convexa que minimiza

$$\int \varphi dP + \int \varphi^* dQ$$

en el conjunto de funciones convexas e inferiormente semicontinuas. Como $\varphi \in L^1(P)$, entonces solo puede tomar el valor $+\infty$ en un conjunto de medida nula. Entonces, $P(\text{Dom}\varphi) = 1$. Además, la frontera de $\text{Dom}\varphi$ es de medida nula, por ser un conjunto convexo. Entonces, $P(\text{Int}(\text{Dom}\varphi)) = 1$ Por el teorema de Rademacher (ver D.1.4), una función convexa definida en un abierto de $\mathbb{R}^d$ es diferenciable en casi todo punto. Como P no da masa a conjuntos de medida nula, φ restringida al interior de su dominio es diferenciable P-c.s. y su gradiente está bien definido, salvo en un conjunto de P-probabilidad nula. Por lo tanto, el conjunto donde φ es diferenciable, tiene $P-$probabilidad 1. Además, como una de las marginales de π es P, el conjunto de puntos $(x, y) \in \mathbb{R}^d$ tales que φ no es diferenciable en x es de π-probabilidad nula. Entonces, por el lema 3.2.6, el único vector que pertenece a la subdiferencial de φ en (x, y) es el gradiente de φ, salvo en un conjunto de puntos de π-probabilidad nula.

Por lo tanto, si π es un plan de transporte óptimo, por el criterio de optimalidad 3.2.14, se verifica que $y = \nabla\varphi(x)$, π-c.s. Esta condición es equivalente a que $\pi = (Id \times \nabla\varphi)\#P$. Se tiene también que $P\#\nabla\varphi = Q$.

En segundo lugar, queda probar la unicidad del plan de transporte. A la vez se probará que solo existe un gradiente de una función convexa $\nabla\varphi$ tal que $\nabla\varphi\#P = Q$.

Sea $\bar{\varphi}$ otra función convexa tal que $\nabla\bar{\varphi}\#P = Q$. El objetivo es ver que $\nabla\bar{\varphi} = \nabla\varphi$ (salvo en un conjunto de P-probabilidad nula). Se considera la probabilidad

$$\pi_0 = (Id \times \nabla\bar{\varphi})\#P \in \prod(P, Q).$$

Por el lema previo 3.2.15, $y = \nabla\bar{\varphi}(x)$, $\pi_0 - c.s.$. En consecuencia, por el criterio de optimalidad 3.2.14, π_0 es un plan de transporte óptimo y el par $(\bar{\varphi}, \bar{\varphi}^*)$ minimiza la expresión

$$J(\varphi, \psi) := \int_{\mathbb{R}^d} \varphi dP + \int_{\mathbb{R}^d} \psi dQ$$

en el conjunto $\Phi = \{(\varphi, \psi) \in L^1(P) \times L^1(Q) : x \cdot y \leq \varphi(x) + \psi(y),\ \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d\}$. Se tiene, por la fórmula de dualidad de la proposición 3.2.2, que

$$\int_{\mathbb{R}^d} \bar{\varphi}(x) dP + \int_{\mathbb{R}^d} \bar{\varphi}^*(y) dQ = \int_{\mathbb{R}^d} \varphi(x) dP + \int_{\mathbb{R}^d} \varphi^*(y) dQ = \int_{\mathbb{R}^d \times \mathbb{R}^d} x \cdot y \, d\pi(x, y).$$

Reescribiéndolo,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \Big[\bar{\varphi}(x) + \bar{\varphi}^*(y) - x \cdot y\Big] d\pi(x, y).$$

Usando que $\pi = (Id \times \nabla\varphi)\#P$, por el teorema de transferencia de integrales,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \Big[\bar{\varphi}(x) + \bar{\varphi}^*(\nabla\varphi(x)) - x \cdot \nabla\varphi(x)\Big] dP(x)$$

si, y solo si,

$$\bar{\varphi}(x) + \bar{\varphi}^*(\nabla\varphi(x)) - x \cdot \nabla\varphi(x) = 0, \quad \pi - c.s.$$

Esta última equivalencia se debe a que la función del integrando es mayor o igual que 0 siempre. Por la proposición 3.2.8,

$$\nabla\varphi(x) \in \partial(\bar{\varphi}(x)), \quad P - c.s.$$

Y, como $\bar{\varphi}$ es diferenciable salvo en un conjunto de probabilidad P nula (por un razonamiento explicado al principio de esta demostración),

$$\nabla\varphi(x) = \nabla(\bar{\varphi}(x)), \quad P - c.s.$$

De esta igualdad, se concluye que existe un único gradiente de una función convexa que minimiza

$$\int \varphi dP + \int \varphi^* dQ$$

en el conjunto de funciones convexas e inferiormente semicontinuas y, por lo tanto, un único plan de transporte óptimo.

□

3.3. Métricas de Wasserstein

Se ha estudiado el problema de transporte óptimo, para el coste $c(x, y) = \|x - y\|^p$ con $p \geq 1$, porque la expresión que minimiza el coste define una métrica en la clase de probabilidades de $\mathbb{R}^d$ con momentos de orden p finitos, denominada métrica de Wasserstein:

$$\mathcal{W}_p(P, Q) = \left(\min_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y)\right)^{\frac{1}{p}}, \quad p \geq 1. \tag{3.19}$$

Volviendo al contexto del Aprendizaje Justo, para buscar subgrupos comunes en una población (que está dividida en clases dependiendo del atributo protegido $S \in \{0, 1\}$), se trabaja con los recortes. Ya se ha visto que el primer paso para entrenar reglas justas es encontrar la parte común a varios conjuntos de datos diferentes.

Formalmente, se trata de buscar un elemento común P_0 a los conjuntos de recortes de cada una de las leyes de probabilidad que generan los datos de cada grupo.

$$P := \mathcal{L}(X|S=0), \quad Q := \mathcal{L}(X|S=1).$$

En el capítulo 1 se ha propuesto un algoritmo para estimar la parte común entre P y Q que utiliza distancias entre probabilidades. En el paso 2 de dicho algoritmo, se calculan los recortes óptimos, es decir,

$$(P_\alpha, Q_\alpha) := \underset{\substack{R\in\mathcal{R}_\alpha(P)\\ S\in\mathcal{R}_\alpha(Q)}}{\operatorname{argmin}} d(R,S),$$

para cierta distancia d entre probabilidades en $\mathbb{R}^d$. Como ya se adelantó, una opción es considerar la métrica de Wasserstein.

A continuación, se demuestra que la expresión 3.19 define una distancia. Pero, para probar que verifica la desigualdad triangular, se necesita un resultado auxiliar: el lema de pegado, que se puede encontrar en [21].

Lema 3.3.1 (Lema de pegado)**.** *Sean $(\mathcal{X}_1, \mathcal{M}_1, P_1)$, $(\mathcal{X}_2, \mathcal{M}_2, P_2)$, $(\mathcal{X}_3, \mathcal{M}_3, P_3)$ tres espacios probabilísticos. Se supone, además, que $\mathcal{X}_1$, $\mathcal{X}_2$ y $\mathcal{X}_3$ son espacios métricos, completos y separables. Entonces, fijadas las probabilidades $\pi_{1,2} \in \prod(P_1, P_2)$ y $\pi_{2,3} \in \prod(P_2, P_3)$, existe una probabilidad π en $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ tal que la distribución marginal de π sobre $\mathcal{X}_1 \times \mathcal{X}_2$ es $\pi_{1,2}$ y su distribución marginal sobre $\mathcal{X}_2 \times \mathcal{X}_3$ es $\pi_{2,3}$.*

Demostración. Para poder construir la probabilidad descrita, se recurre a la desintegración de medidas, que se explica en el apéndice F.

Como $\pi_{1,2}$ es una probabilidad en $\prod(P_1, P_2)$, existe una función $F : \mathcal{M}_1 \times \mathcal{X}_2 \to \mathbb{R}$ de forma que

- Para cada $y \in \mathcal{X}_2$ fijo, $F_y := F(\cdot, y)$ es una probabilidad en $(\mathcal{X}_1, \mathcal{M}_1)$.
- La función $y \mapsto F(A, y)$ es medible en $(\mathcal{X}_2, \mathcal{M}_2)$ para todo $A \in \mathcal{M}_1$.

verificando que

$$\pi_{1,2}(E) = \int_{\mathcal{X}_2} F(E_y, y)\, dP_2(y), \quad \forall E \in \mathcal{M}_1 \otimes \mathcal{M}_2.$$

Análogamente, para $\pi_{2,3}$, existe una función $G : \mathcal{X}_2 \times \mathcal{M}_3 \to \mathbb{R}$ de forma que

- Para cada $y \in \mathcal{X}_2$ fijo, $G_y := G(y, \cdot)$ es una probabilidad en $(\mathcal{X}_3, \mathcal{M}_3)$.
- La función $y \mapsto G(y, C)$ es medible en $(\mathcal{X}_2, \mathcal{M}_2)$ para todo $C \in \mathcal{M}_3$.

verificando que

$$\pi_{2,3}(F) = \int_{\mathcal{X}_2} G(y, F_y)\, dP_2(y), \quad \forall F \in \mathcal{M}_2 \otimes \mathcal{M}_3.$$

Primero, se da una idea intuitiva para construir una probabilidad π en $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ que cumpla las condiciones del lema: Se fija $y \in \mathcal{X}_2$. Si $D \in \mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \mathcal{M}_3$,

la probabilidad que se asigna a la $y-$sección de D es la probabilidad producto $F_y \otimes G_y\,(D_y)$. Después, se integra en $\mathcal{X}_2$, respecto de la probabilidad P_2.

Formalizando lo anterior, si $D \in \mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \mathcal{M}_3$, se define

$$\begin{aligned}\pi(D) &:= \int_{\mathcal{X}_2} F_y \otimes G_y\,(D_y)\,dP_2(y)\\ &= \int_{\mathcal{X}_2} \Big(\int_{\mathcal{X}_1} G_y((D_y)_x)\,dF_y(x) \Big)\,dP_2(y) = \int_{\mathcal{X}_2} \Big(\int_{\mathcal{X}_3} F_y((D_y)_z)\,dG_y(z) \Big)\,dP_2(y).\end{aligned}$$

Entonces, si $U \in \mathcal{M}_1 \otimes \mathcal{M}_2$,

$$\begin{aligned}\pi(U \times \mathcal{X}_3) &= \int_{\mathcal{X}_2} F_y \otimes G_y\,(U_y \times \mathcal{X}_3)\,dP_2(y) = \int_{\mathcal{X}_2} F_y(U_y) G_y(\mathcal{X}_3)\,dP_2(y)\\ &= \int_{\mathcal{X}_2} F_y(U_y)\,dP_2(y) = \pi_{1,2}(U).\end{aligned}$$

Análogamente, si $V \in \mathcal{M}_2 \otimes \mathcal{M}_3$,

$$\begin{aligned}\pi(\mathcal{X}_1 \times V) &= \int_{\mathcal{X}_2} F_y \otimes G_y\,(\mathcal{X}_1 \times V_y)\,dP_2(y) = \int_{\mathcal{X}_2} F_y(\mathcal{X}_1) G_y(V_y)\,dP_2(y)\\ &= \int_{\mathcal{X}_2} G_y(V_y)\,dP_2(y) = \pi_{2,3}(U).\end{aligned}$$

□

Proposición 3.3.2. *Sea $p \geq 1$. Se considera el espacio de probabilidades*

$$\mathcal{F}_p = \{\mu \text{ probabilidad en } \mathbb{R}^d : \int_{\mathbb{R}^d} \|x\|^p d\mu(x) < +\infty\}.$$

Se denomina

$$\mathcal{T}_p(P,Q) = \min_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y)$$

al coste de transporte óptimo entre dos probabilidades P y Q de $\mathcal{F}_p$. Entonces,

$$\mathcal{W}_p : \mathcal{F}_p \times \mathcal{F}_p \to \mathbb{R}, \quad \mathcal{W}_p(P,Q) = (\mathcal{T}_p(P,Q))^{\frac{1}{p}}$$

es una distancia.

Demostración. Como el coste $c(x,y) = \|x-y\|^p$ es una función simétrica y positiva, $\mathcal{W}_p$ también lo es. Se demuestra, a continuación que $\mathcal{W}_p(P,Q) = 0$ si, y solo si, $P = Q$.

Se ha probado que existe un plan de transporte óptimo $\pi_0 \in \prod(P,Q)$. Es decir, existe $\pi_0 \in \prod(P,Q)$ tal que

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi_0(x,y) = \min_{\pi \in \prod(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y).$$

Si $\mathcal{W}_p(P,Q) = 0$, entonces

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi_0(x,y) = 0.$$

Por lo tanto, $\|x-y\|^p = 0,\ \pi_0$-c.s. $\Leftrightarrow x = y,\ \pi_0$-c.s.. A partir de esta igualdad, se comprueba que $P = Q$. Para todo $A \in \mathbb{R}^d$,

$$P(A) = \pi_0\big(A \times \mathbb{R}^d\big) = \pi_0\big(A \times \mathbb{R}^d \cap \triangle(\mathbb{R}^d \times \mathbb{R}^d)\big) = \pi_0\big(\{(x,x) : x \in A\}\big),$$

$$Q(A) = \pi_0\big(\mathbb{R}^d \times A\big) = \pi_0\big(\mathbb{R}^d \times A \cap \triangle(\mathbb{R}^d \times \mathbb{R}^d)\big) = \pi_0\big(\{(y,y) : y \in A\}\big).$$

Recíprocamente, si $P = Q$, se define la probabilidad

$$\pi_0(C) = P\big(\mathrm{pr}(C \cap \triangle(\mathbb{R}^d \times \mathbb{R}^d)\big) = P\big(\{x \in \mathbb{R}^d : (x,x) \in C\}\big) \quad \forall C \in \beta^{2d}.$$

π_0 es una probabilidad bien definida.

- $\pi_0(A) \geq 0$ para todo $A \in \beta^{2d}$ y $\pi_0(\mathbb{R}^d \times \mathbb{R}^d) = P(\mathbb{R}^d) = 1$.
- Si $\{C_n\}_n$ es una sucesión disjunta de elementos de β^{2d}, entonces

$$\begin{aligned}\pi_0\big(\sum_{n=1}^{\infty} C_n\big) &= P\big(\{x \in \mathbb{R}^d : (x,x) \in \sum_{n=1}^{\infty} C_n\}\big) = P\big(\sum_{n=1}^{\infty}\{x \in \mathbb{R}^d : (x,x) \in C_n\}\big) \\ &= \sum_{n=1}^{\infty} P\big(\mathrm{pr}(C_n \cap \triangle)\big) = \sum_{n=1}^{\infty} \pi_0(C_n).\end{aligned}$$

Por la definición de π_0, sus marginales son P y Q respectivamente. El soporte de esta probabilidad está contenido en la diagonal, ya que $\pi_0(\triangle) = P(\mathbb{R}^d) = 1$. Por lo tanto, $x = y,\ \pi_0$-c.s. y $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi_0(x,y) = 0$. Entonces, π_0 es la distribución que alcanza el mínimo y $\mathcal{T}_p(P,Q) = 0$.

Queda probar la desigualdad triangular. Si π es una probabilidad en $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, en el espacio $L^p(\pi)$, se tiene la desigualdad de Minkowski:

$$\|x-z\|_p \leq \|x-y\|_p + \|y-z\|_p.$$

Reescribiéndolo,

$$\begin{aligned}&\Big(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-z\|^p d\pi_{1,3}(y,z)\Big)^{\frac{1}{p}} = \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|x-z\|^p d\pi(x,y,z)\Big)^{\frac{1}{p}} \\ &\leq \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y,z)\Big)^{\frac{1}{p}} + \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|y-z\|^p d\pi(x,y,z)\Big)^{\frac{1}{p}} \\ &= \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi_{1,2}(x,y)\Big)^{\frac{1}{p}} + \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y-z\|^p d\pi_{2,3}(y,z)\Big)^{\frac{1}{p}},\end{aligned}$$

donde $\pi_{1,2}$ y $\pi_{2,3}$ son las marginales $1,2$ y $2,3$ de π, respectivamente.

Sean μ, ν, ρ tres probabilidades de $\mathcal{F}_p$. Se denota por $\pi_{1,2}$ a la distribución que minimiza el coste de transporte óptimo para μ y ν. Análogamente, $\pi_{2,3}$ es la probabilidad que minimiza el coste de transporte óptimo para ν y ρ. Es decir,

$$\mathcal{W}_p^p(\mu,\nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi_{1,2}(x,y),$$

$$\mathcal{W}_p^p(\nu,\rho) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y-z\|^p d\pi_{2,3}(y,z).$$

El lema de pegado 3.3.1 garantiza que existe $\pi \in \prod(\mu, \nu, \rho)$ cuyas marginales $1, 2$ y $2, 3$ son $\pi_{1,2}$ y $\pi_{2,3}$. Por lo tanto,

$$\mathcal{W}_p(\mu, \rho) \leq \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - z\|^p d\pi_{1,3}(y, z) \Big)^{\frac{1}{p}} \leq \mathcal{W}_p(\mu, \nu) + \mathcal{W}_p(\nu, \rho).$$

□

3.3.1. Aproximaciones empíricas

Hasta ahora, se han estudiado las propiedades teóricas de la distancia de Wasserstein. Pero, para poder usar esta distancia en el algoritmo descrito en el capítulo 1, se necesita un método para hallarla. En la práctica, la forma de calcular la distancia de Wasserstein es a partir de aproximaciones discretas.

Sean R y S dos probabilidades en $\mathbb{R}^d$ con momentos de orden $p \geq 1$ finitos. Sean $n, m \in \mathbb{N}$, se toma una muestra de cada distribución de n y m elementos, respectivamente. Es decir, se toman $X_1, X_2, ..., X_n$ vectores aleatorios independientes con distribución R y, análogamente, $Y_1, Y_2, ..., Y_m$ vectores aleatorios independientes con distribución S. Las probabilidades empíricas dadas por estas muestras son:

$$R_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}, \qquad S_m = \frac{1}{m} \sum_{j=1}^{m} \delta_{y_j}.$$

Calcular la distancia de Wasserstein entre R_n y S_m es un problema de investigación operativa.

$$\mathcal{W}_p^p(R_n, S_m) = \text{mín} \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - y_j\|^2 \pi_{i,j},$$

sujeto a $\pi_{i,j} \geq 0$; $\sum\limits_{j=1}^{n} \pi_{i,j} = \frac{1}{n}$; $\sum\limits_{i=1}^{n} \pi_{i,j} = \frac{1}{n}$.

El algoritmo para resolver este problema tiene un coste alto, pero se ha reducido el problema original a uno discreto, calculable en la práctica. Si la distancia de Wasserstein empírica (entre R_n y S_m) tiende hacia la distancia de Wasserstein ente R y S, cuando n y m tienden a ∞, entonces se tiene un método práctico para estimar la distancia de Wasserstein entre dos probabilidades. No es cierto que se verifique esta propiedad para cualquier distancia entre probabilidades. Por ejemplo, ya se explicó en la sección 1.3.2 del capítulo 1 que no se verifica para la distancia en variación total. Sin embargo, en el caso de la distancia de Wasserstein, sí se puede aproximar a partir de las distancias empíricas. Se enuncia, a continuación, un resultado que caracteriza la convergencia en la distancia de Wasserstein; su demostración se puede encontrar en el lema 8.3 de [5].

Proposición 3.3.3. *Sean R y $\{R_n\}_{n=1}^{\infty}$ probabilidades en $\mathcal{F}_p(\mathbb{R}^d)$, $p \geq 1$. Entonces,*

$$\lim_{n \to \infty} \mathcal{W}_p(R_n, R) = 0$$

si, y solo si, la sucesión $\{R_n\}_{n=1}^{\infty}$ converge débilmente a R y, además,

$$\lim_{n \to \infty} \int \|x\|^p \, dR_n = \int \|x\|^p \, dR.$$

Demostración. Si P y Q son dos probabilidades en $\mathcal{F}_p(\mathbb{R}^d)$, entonces

$$\mathcal{W}_p^p(P,Q) = \min_{\pi\in\prod(P,Q)} \int \|x-y\|^p d\pi(x,y) = \min_{(X,Y)\sim(P,Q)} E\|X-Y\|^p.$$

$\Rightarrow$) Se denota por π_n a la probabilidad de $\prod(R_n, R)$ que minimiza el coste de transporte. Como se detalla en el apéndice F sobre desintegración de medidas, para cada $y \in \mathbb{R}^d$, existe una medida $R_n(\,\cdot\,|y)$ tal que

$$\pi_n(A\times B) = \int_B R_n(A|y)dR(y).$$

Sea X un vector aleatorio con distribución R. Condicionalmente dado $X = y$, se toman $\{X_n\}_n$ vectores aleatorios independientes con $X_n \sim R_n(\,\cdot\,|y)$. Entonces,

$$(X_n, X) \sim \pi_n.$$

Por lo tanto,

$$E\|X_n - X\|^p = \mathcal{W}_p^p(R_n, R).$$

Si se supone que $\lim_{n\to\infty} \mathcal{W}_p(R_n, R) = 0$, para todo $\varepsilon > 0$, aplicando la desigualdad de Markov,

$$P(\|X_n - X\|^p > \varepsilon) \le \frac{E\|X_n - X\|^p}{\varepsilon} \xrightarrow[n\to\infty]{} 0.$$

De aquí se deduce que X_n converge en probabilidad a X. Esta convergencia es más fuerte que la convergencia en distribución. Entonces, se concluye que

$$R_n \underset{w}{\to} R, \quad \text{cuando } n \to \infty.$$

Por otro lado, aplicando la desigualdad triangular de la norma en $L^p(\pi_n)$, para cada $n \in \mathbb{N}$,

$$\left|\Big(E\|X_n\|^p\Big)^{\frac{1}{p}} - \Big(E\|X\|^p\Big)^{\frac{1}{p}}\right| \le \Big(E\|X_n - X\|^p\Big)^{\frac{1}{p}} \xrightarrow[n\to\infty]{} 0.$$

Entonces,

$$E\|X_n\|^p \xrightarrow[n\to\infty]{} E\|X\|^p,$$

es decir, $\lim_{n\to\infty} \int \|x\|^p \, dR_n = \int \|x\|^p \, dR$.

$\Leftarrow$) Recíprocamente, por el teorema de Representación de Skorokhod (ver C.0.3), existen X y $\{X_n\}_n$ vectores aleatorios tales que

$$X_n \sim R_n, \quad X \sim R \quad \text{y} \quad X_n \underset{c.s.}{\longrightarrow} X.$$

Además, por hipótesis,

$$E\|X_n\|^p \xrightarrow[n\to\infty]{} E\|X\|^p < \infty.$$

Se concluye, aplicando el teorema de Vitali (ver E.1.3), que

$$\mathcal{W}_p^p(R_n, R) \le E\|X_n - X\|^p \xrightarrow[n\to\infty]{} 0.$$

□

Usando la proposición anterior, se demuestra que la distancia de Wasserstein entre las probabilidades empíricas tiende a la distancia de Wasserstein entre las respectivas probabilidades reales.

Proposición 3.3.4. *Sea R una probabilidad en $\mathcal{F}_p(\mathbb{R}^d)$, con $p \geq 1$. Sea $\{X_n\}_n$ una sucesión de vectores aleatorios independientes e igualmente distribuidos con ley de probabilidad R, definidos en el espacio probabilístico $(\Omega, \mathcal{F}, \mathbb{P})$. Si $w \in \Omega$, la probabilidad empírica R_n^ω asociada a la muestra $(X_1(\omega), \ldots, X_n(\omega))$ es*

$$R_n^\omega = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)},$$

donde δ_{X_i} es la medida de Dirac en $X_i(\omega)$. Entonces, cuando n tiende a $+\infty$,

$$W_2(R_n^\omega, R) \to 0, \quad \textit{para casi todo } \omega \in \Omega.$$

Demostración. Sea F la función de distribución de la probabilidad R. Teniendo en cuenta que $X_i \leq x$ significa $X_{i,1} \leq x_1, \ldots, X_{i,d} \leq x_d$, sea

$$F_n^\omega(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i(\omega) \leq x\}}$$

la función de distribución empírica asociada a la muestra $(X_1(\omega), \ldots, X_n(\omega))$. Por el teorema de Glivenko-Cantelli,

$$\sup_{x \in \mathbb{R}^d} |F_n^\omega(x) - F(x)| \xrightarrow{c.s.} 0.$$

En particular, se tiene convergencia puntual para los puntos de continuidad de F. Esto garantiza que $\{R_n^\omega\}_n$ converge débilmente a R para casi todo $\omega \in \Omega$.

Por otro lado, como R tiene momento cuadrático finito, la ley de los grandes números asegura que:

$$\int \|x\|^p \, dR_n^\omega(x) = \frac{1}{n} \sum_{i=1}^{n} \|X_i(\omega)\|^p \to \int \|x\|^p \, dR(x), \quad \text{c.s.}$$

Se verifican las hipótesis de la proposición 3.3.3, por lo tanto,

$$W_2(R_n^\omega, R) \to 0, \quad \text{para casi todo } \omega \in \Omega.$$

□

Corolario 3.3.5. *Sean R y S dos probabilidades en $\mathcal{F}_p(\mathbb{R}^d)$, con $p \geq 1$. Sea $\{X_n\}_n$ una sucesión de vectores aleatorios independientes e igualmente distribuidos con ley de probabilidad R y sea $\{Y_m\}_m$ una sucesión de vectores aleatorios independientes e igualmente distribuidos con ley de probabilidad S. Ambas sucesiones son independientes y están definidas en el espacio probabilístico $(\Omega, \mathcal{F}, \mathbb{P})$. Si $w \in \Omega$, la probabilidad empírica R_n^ω asociada a la muestra $(X_1(\omega), \ldots, X_n(\omega))$ es*

$$R_n^\omega = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}.$$

Análogamente, la probabilidad empírica S_m^ω *asociada a la muestra* $(Y_1(\omega), \ldots, Y_m(\omega))$ *es*

$$S_m^\omega = \frac{1}{m} \sum_{j=1}^{m} \delta_{Y_j(\omega)}.$$

Entonces, cuando $\min(n, m) \to +\infty$,

$$W_2(R_n^\omega, S_m^\omega) \to 0, \quad \textit{para casi todo } \omega \in \Omega.$$

Demostración. Se comprueba fácilmente a partir de la desigualdad triangular y de la proposición anterior:

$$W_2(R_n^\omega, S_m^\omega) \leq W_2(R_n^\omega, R) + W_2(R, S) + W_2(R_n^\omega, S_m^\omega).$$

□

Se concluye este apartado, probando que se alcanza el mínimo de la distancia de Wasserstein entre los conjuntos de recortes de dos probabilidades P_1 y P_2. Es decir, se garantiza que existe un minimizador de $\mathcal{W}_p(R, S)$, donde $R \in \mathcal{R}_\alpha(P_1)$ y $S \in \mathcal{R}_\alpha(P_2)$. Este resultado es importante para poder aplicar el paso 2 del algoritmo que se explica en la sección 1.3.2.

En el algoritmo propuesto en esa sección, fijado un nivel $\alpha \in [0, 1]$, se usa la distancia de Wasserstein para encontrar el par de recortes óptimos de nivel α entre las dos probabilidades que se quieren comparar (P_1 y P_2), es decir, los dos recortes "más parecidos" entre sí, en el sentido de que su distancia de Wasserstein sea la más pequeña posible. Si son iguales (se denota por P_0), quiere decir que P_0 verifica que

$$\begin{aligned} P_1 &= (1-\alpha)P_0 + \alpha R_1, \\ P_2 &= (1-\alpha)P_0 + \alpha R_2, \end{aligned} \tag{3.20}$$

con R_1 y R_2 probabilidades

Cuando α sea el máximo para el cual $\mathcal{R}_\alpha(P_1) \cap R_\alpha(P_2) \neq \emptyset$, la probabilidad P_0 que verifique 3.3.1 será la parte común entre P_1 y P_2.

Primero se prueba un lema necesario para la demostración del resultado principal.

Lema 3.3.6. *Sea* $p \geq 1$ *y sea* $P \in \mathcal{F}_2(\mathbb{R}^d)$. *Entonces* $\mathcal{R}_\alpha(P)$ *es compacto para la distancia* $\mathcal{W}_p$.

Demostración. Se prueba que cualquier sucesión $\{R_n\}_{n=1}^\infty \subset \mathcal{R}_\alpha(P)$ admite una subsucesión convergente para la distancia $\mathcal{W}_p$.

Si $\{R_n\}_{n=1}^\infty \subset \mathcal{R}_\alpha(P)$, entonces es ajustada (ver el apéndice C.1): Fijado $\varepsilon > 0$, sea K compacto en $\mathbb{R}^d$ tal que $P(K^C) < (1-\alpha)\varepsilon$. Entonces, $P_n(K^C) < \frac{1}{(1-\alpha)}(1-\alpha)\varepsilon = \varepsilon$ para todo $n \in \mathbb{N}$. Por lo tanto, se puede extraer una subsucesión $\{R_{n_j}\}_{j=1}^\infty$ que converge débilmente hacia una probabilidad R en $\mathbb{R}^d$. Además, se comprueba que $R \in \mathcal{R}_\alpha(P)$:

Si $A \in \beta^d$, por la regularidad exterior de la probabilidad P (ver B),

$$P(A) = \inf\{P(U) : U \text{ abierto}, \ A \subset U, \ U \in \beta^d\}.$$

Además, si $U \in \beta^d$ es un abierto que contiene a A, por el teorema de Portmanteau (ver en los apéndices C.0.2), se tiene que

$$R(A) \leq R(U) \leq \liminf_{n\to\infty} R_n(U) \leq \frac{1}{1-\alpha} P(U).$$

Entonces,

$$R(A) \leq \frac{1}{1-\alpha} \text{ínf}\{P(U) : U \text{ abierto}, \ A \subset U, \ U \in \beta^d\} = \frac{1}{1-\alpha} P(A).$$

Falta comprobar que $\lim_{j\to\infty} \mathcal{W}_p(R_{n_j}, R) = 0$. Por la proposición 3.3.3, basta probar que

$$\lim_{j\to\infty} \int \|x\|^p \, dR_{n_j} = \int \|x\|^p \, dR.$$

Como para todo $j \in \mathbb{N}$, R_{n_j} y R son recortes de nivel α de P, se verifica, para todo $M \in \mathbb{N}$, que

$$\int_{\|x\|^p \geq M} \|x\|^p \, dR_{n_j} \leq \frac{1}{1-\alpha} \int_{\|x\|^p \geq M} \|x\|^p \, dP, \quad \forall j \in \mathbb{N};$$
$$\int_{\|x\|^p \geq M} \|x\|^p \, dR \leq \frac{1}{1-\alpha} \int_{\|x\|^p \geq M} \|x\|^p \, dP.$$

Por el teorema de la convergencia dominada, es claro que $\lim_{M\to\infty} \int_{\|x\|^p \geq M} \|x\|^p \, dP = 0$. Por lo tanto, fijado $\varepsilon > 0$, existe $M_0 \in \mathbb{N}$, tal que $\frac{1}{1-\alpha} \int_{\|x\|^p \geq M_0} \|x\|^p \, dP < \frac{\varepsilon}{3}$. Por otro lado, por el teorema de Portmanteau, se tiene que:

$$\lim_{j\to\infty} \int \text{mín}(\|x\|^p, M) \, dR_{n_j} = \int \text{mín}(\|x\|^p, M) \, dR, \quad \forall M \in \mathbb{N}.$$

Entonces, fijado $\varepsilon > 0$, existe un $j_0 \in \mathbb{N}$ tal que, para todo $j \geq j_0$,

$$\left| \int \text{mín}(\|x\|^p, M) \, dR_{n_j} - \int \text{mín}(\|x\|^p, M) \, dR \right| \leq \frac{\varepsilon}{3}.$$

Por lo tanto, si $j \geq j_0$,

$$\begin{aligned}
&\left| \int \|x\|^p \, dR_{n_j} - \int \|x\|^p \, dR \right| \\
&\leq \left| \int_{\|x\|^p \leq M_0} \|x\|^p \, dR_{n_j} - \int_{\|x\|^p \leq M_0} \|x\|^p \, dR \right| \\
&+ \int_{\|x\|^p > M_0} \|x\|^p \, dR_{n_j} + \int_{\|x\|^p > M_0} \|x\|^p \, dR \\
&\leq \left| \int \text{mín}(\|x\|^p, M_0) \, dR_{n_j} - \int \text{mín}(\|x\|^p, M_0) \, dR \right| \\
&+ \int_{\|x\|^p > M_0} \|x\|^p \, dR_{n_j} + \int_{\|x\|^p > M_0} \|x\|^p \, dR \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.
\end{aligned}$$

Con lo que se concluye la prueba. □

Proposición 3.3.7. *Sea $p \geq 1$. Se consideran $P_1, P_2 \in \mathcal{F}_p(\mathbb{R}^d)$ y $\alpha > 0$. Entonces, existe $(\tilde{R}_1, \tilde{R}_2) \in \mathcal{R}_\alpha(P_1) \times \mathcal{R}_\alpha(P_2)$ tal que*

$$(\tilde{R}_1, \tilde{R}_2) = \underset{\substack{R_1 \in \mathcal{R}_\alpha(P_1) \\ R_2 \in \mathcal{R}_\alpha(P_2)}}{\operatorname{argmin}} \mathcal{W}_p(R_1, R_2).$$

Esto quiere decir que la función

$$\mathcal{W}_p(\cdot, \cdot) : \mathcal{F}_p(\mathbb{R}^d) \times \mathcal{F}_p(\mathbb{R}^d) \to \mathbb{R}$$

alcanza un mínimo absoluto en el conjunto $\mathcal{R}_\alpha(P_1) \times \mathcal{R}_\alpha(P_2)$.

Demostración. Por el lema 3.3.6, se sabe que $\mathcal{R}_\alpha(P_1)$ y $\mathcal{R}_\alpha(P_2)$ son compactos para la distancia $\mathcal{W}_p$. Entonces, $\mathcal{R}_\alpha(P_1) \times \mathcal{R}_\alpha(P_2)$ es compacto en $\mathcal{F}_p(\mathbb{R}^d) \times \mathcal{F}_p(\mathbb{R}^d)$ con la topología producto, que es la dada por la distancia:

$$\mathcal{R}_\alpha(P_1) \times \mathcal{R}_\alpha(P_2) \to \mathbb{R}; \qquad \big((\mu_1, \mu_2), (\nu_1, \nu_2)\big) \mapsto \mathcal{W}_p(\mu_1, \nu_1) + \mathcal{W}_p(\mu_2, \nu_2).$$

Además, $\mathcal{W}_p(\cdot, \cdot) : \mathcal{F}_p(\mathbb{R}^d) \times \mathcal{F}_p(\mathbb{R}^d) \to \mathbb{R}$ es continua con esta topología. Si $(\mu_1, \mu_2), (\nu_1, \nu_2) \in \mathcal{R}_\alpha(P_1) \times \mathcal{R}_\alpha(P_2)$, se cumple:

$$\begin{aligned} &\Big|W_p(\mu_1, \mu_2) - \mathcal{W}_p(\nu_1, \nu_2)\Big| \\ &\leq \Big|W_p(\mu_1, \mu_2) - W_p(\nu_1, \mu_2)\Big| + \Big|W_p(\nu_1, \mu_2) - \mathcal{W}_p(\nu_1, \nu_2)\Big| \\ &\leq W_p(\mu_1, \nu_1) + W_p(\mu_2, \nu_2). \end{aligned}$$

Con esta cota se deduce que la función es uniformemente continua, y, por lo tanto, alcanza el mínimo absoluto en un compacto.

□

3.3.2. Distancia de Wasserstein en $\mathbb{R}$

Para concluir este capítulo, se prueba un último resultado, que caracteriza la distancia de Wasserstein para el coste cuadrático en $\mathbb{R}$. La distancia de Wasserstein para $p = 2$ entre dos probabilidades es la distancia L^2 entre los cuantiles de las respectivas distribuciones de probabilidad.

Proposición 3.3.8. *Sean μ y ν dos probabilidades en $\mathbb{R}$. Si μ tiene función de distribución F y ν tiene función de distribución G, entonces*

$$\mathcal{W}_2^2(\mu, \nu) = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt.$$

Demostración. Caso 1: Se supone que μ tiene densidad, entonces no da masa a conjuntos de medida nula. Por el teorema (3.2.16), existe una única distribución de probabilidad óptima, $\pi = (Id \times \nabla\varphi)\#\mu$, donde $\nabla\varphi$ es el único gradiente de una función convexa que verifica $\nabla\varphi\#\mu = \nu$. En dimensión 1, las derivadas de funciones convexas son las funciones crecientes.

Si $X \sim \mu$, entonces $F(X) \sim U(0,1)$ y $G^{-1} \circ F(X) \sim \nu$, estas propiedades se prueban en la sección C.2 de los apéndices. Como F y G^{-1} son crecientes, $G^{-1} \circ F$ también. Entonces, $G^{-1} \circ F$ es la única aplicación de transporte óptimo de μ a ν.

$$\mathcal{W}_2^2(\mu,\nu) = \int_{\mathbb{R}} (x - G^{-1} \circ F(x)) dF(x) = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt.$$

Caso general: Si μ es una probabilidad en $\mathbb{R}^d$, siempre existe una sucesión de probabilidades con densidad $\{\mu_n\}_{n=1}^{\infty}$ que converge débilmente hacia μ. Esta sucesión se construye regularizando mediante convoluciones la distribución μ, se detalla este procedimiento en el apéndice C, concretamente en C.3.1. Además, escogiendo una sucesión adecuada, se prueba en ese mismo apéndice que se tiene la convergencia de momentos de orden 2 de μ_n hacia μ. Entonces, si F es la función de distribución de μ, y F_n es la función de distribución de μ_n para cada $n \in \mathbb{N}$, por definición de convergencia débil, se tiene que $\lim_{n\to\infty} F_n(x) = F(x)$ para todo $x \in \mathbb{R}$ en los que F es continua. Por el teorema de Skorohod (C.2.5), $\lim_{n\to\infty} F_n^{-1}(t) = F(t)$ para casi todo $t \in [0,1]$. Por lo tanto, aplicando el lema de Fatou (válido para funciones positivas),

$$\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt \leq \liminf_{n\to\infty} \int_0^1 (F_n^{-1}(t) - G^{-1}(t))^2 dt = \liminf_{n\to\infty} \mathcal{W}_2^2(\mu_n,\nu).$$

Se ha probado en la proposición 3.3.3 que

$$\lim_{n\to\infty} \mathcal{W}_2^2(\mu_n,\nu) = \mathcal{W}_2^2(\mu,\nu).$$

Por lo tanto,

$$\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt \leq \mathcal{W}_2^2(\mu,\nu).$$

La desigualdad contraria se deduce considerando la probabilidad $\pi = (Id \times G^{-1} \circ F)\#\mu$. Se tiene que $\pi \in \prod(\mu,\nu)$ y, por lo tanto,

$$\mathcal{W}_2^2(\mu,\nu) \leq \int_{\mathbb{R}} (x - G^{-1} \circ F(x)) dF(x) = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt.$$

□

Capítulo 4

Transporte entrópico

El problema, que se ha planteado en el capítulo anterior, de minimizar el coste de transporte tiene un coste computacional alto. La forma de calcular la distancia de Wasserstein entre dos probabilidades R y S en $\mathbb{R}^d$, con momentos finitos del orden que sea necesario, es a partir de aproximaciones empíricas R_n y S_n, para cierto $n \in \mathbb{N}$. El coste de esos cálculos es del orden de n^3.

$$\mathcal{W}_2(R_n, S_n) = \text{mín} \sum_{i=1}^{n}\sum_{j=1}^{n} \|x_i - y_j\|^2 \pi_{i,j}, \text{ sujeto a } \pi_{i,j} \geq 0; \sum_{j=1}^{n} \pi_{i,j} = \frac{1}{n}; \sum_{i=1}^{n} \pi_{i,j} = \frac{1}{n}.$$

En esta sección se da una alternativa a la distancia de Wasserstein: el transporte entrópico. Si la expresión a minimizar en el problema de transporte óptimo se contamina con la divergencia de Kullback-Leibler, el problema de minimización resultante se puede resolver con una iteración de punto fijo. Se denomina coste de transporte entrópico a:

$$\mathcal{W}^2_{2,\varepsilon}(P, Q) = \inf_{\pi \in \prod(P,Q)} \left[\int \frac{\|x - y\|^2}{2} d\pi(x, y) + \varepsilon D(\pi | P \otimes Q) \right],$$

donde $D(\pi|P \otimes Q)$ es la divergencia de Kulback-Leibler de π respecto de la medida producto $P \otimes Q$ y $\varepsilon > 0$ determina cuánto se va a modificar el funcional lineal en π:

$$I(\pi) = \int \frac{\|x - y\|^2}{2} d\pi(x, y).$$

Primero, se estudiarán las propiedades de la divergencia de Kullback-Leibler, que mide la diferencia entre dos probabilidades, buscando minimizadores en conjuntos con ciertas propiedades. Después, se formulará el problema de transporte entrópico.

4.1. Divergencia de Kullback-Leibler

Si $P \ll Q$ son dos probabilidades, la divergencia de Kullback-Leibler de P respecto de Q da una idea de cuánto se parecen ambas probabilidades. Si son muy similares, su divergencia será prácticamente 0, mientras que si se diferencian mucho, la divergencia de Kullback-Leibler tomará valores altos. Esto ocurre porque la divergencia de Kullback-Leibler se define a partir de una integral de la función $\log \frac{dP}{dQ}$.

Si P y Q son parecidas, la derivada de Radon-Nikodym de P respecto de Q será prácticamente 1, y, por tanto, su logaritmo será casi 0.

En el apéndice A, se recuerdan las definiciones y teoremas sobre las medidas absolutamente continuas y la derivada de Radon-Nikodym que se usarán continuamente a lo largo de este capítulo.

Los resultados que se prueban en esta sección son una adaptación de [10]. En el artículo, solo se habla de la divergencia de Kullback-Leibler de probabilidades. En este trabajo, se generaliza a cualquier medida finita y positiva, porque es necesario, más adelante, para probar que existe un minimizador de

$$\mathcal{W}_{2,\varepsilon}^2(P,Q) = \inf_{\pi\in\prod(P,Q)}\left[\int \frac{\|x-y\|^2}{2}d\pi(x,y) + \varepsilon D(\pi|P\otimes Q)\right].$$

A continuación, se da la definición formal de la divergencia de Kullback-Leibler.

Definición 4.1.1. *Sean P y Q medidas finitas y positivas en un espacio medible $(\mathcal{X},\mathcal{M})$ tales que $P \ll Q$, es decir, P es absolutamente continua respecto de Q. Se denota por $\frac{dP}{dQ} : \mathcal{X} \to [0,\infty]$ a la derivada de Radon-Nikodym de P respecto de Q. Se define la divergencia de Kullback-Leibler como*

$$D(P|Q) := \int \log\frac{dP}{dQ}dP = \int \frac{dP}{dQ}\log\frac{dP}{dQ}dQ.$$

Si P no es absolutamente continua respecto de Q, entonces se define

$$D(P|Q) := +\infty.$$

Como P es una medida finita y positiva, se puede suponer que la derivada de Radon-Nikodym de P respecto de Q es una función $\frac{dP}{dQ} : \mathcal{X} \to [0,\infty)$, ya que como mucho toma el valor $+\infty$ con $Q-$medida 0 (y se podría redefinir la función en esos puntos). Además, la definición de la divergencia de Kullback-Leibler es correcta porque el conjunto $A = \{x \in \mathcal{X} : \frac{dP}{dQ}(x) = 0\}$ cumple, obviamente, que $P(A) = 0$. Por lo tanto, la función $\log\frac{dP}{dQ} : \mathcal{X} \to (0,\infty)$ únicamente no está bien definida en un conjunto de $P-$medida nula. Se entiende que

$$\int \log\frac{dP}{dQ}dP = \int_{A^C} \log\frac{dP}{dQ}dP.$$

Si se acuerda que $\log 0 = -\infty$ y que $0\cdot(\pm\infty) = 0$, no hace falta hacer las consideraciones anteriores. En los resultados sucesivos, se tendrá en cuenta este comentario. Se supone también que $\log\frac{r}{0} = +\infty$, para $r > 0$. Con este convenio, se tiene una caracterización de la divergencia de Kullback-Leibler, dada por el lema siguiente. Si se tiene que las medidas P y Q son ambas absolutamente continuas respecto de otra medida R, entonces la divergencia de P respecto de Q se puede escribir a partir de las derivadas de Radon-Nikodym de P y Q respecto de R.

Lema 4.1.2. *Si P, Q y R son medidas finitas y positivas en un espacio medible $(\mathcal{X},\mathcal{M})$ tales que $P \ll R$ y $Q \ll R$, entonces*

$$D(P|Q) = \int \log\frac{\frac{dP}{dR}}{\frac{dQ}{dR}}dP = \int \frac{dP}{dR}\log\frac{\frac{dP}{dR}}{\frac{dQ}{dR}}dR.$$

Demostración. Si $P \ll Q$, entonces $\frac{dP}{dR} = \frac{dP}{dQ}\frac{dQ}{dR}$, $R - c.s.$. Entonces, el conjunto de puntos donde se anula $\frac{dQ}{dR}$ está contenido en el conjunto de puntos donde se anula $\frac{dP}{dR}$, que tiene medida P nula. Por lo tanto, la integral está bien definida. Se entiende que

$$\int \log \frac{dP}{dQ} dP = \int_{\{x \in \mathcal{X}: \frac{dP}{dQ}(x) \neq 0\}^C} \log \frac{dP}{dQ} dP = \int_{\{x \in \mathcal{X}: \frac{dP}{dQ}(x) \neq 0\}^C} \log \frac{\frac{dP}{dR}}{\frac{dQ}{dR}} dP.$$

Si P no es absolutamente continua respecto de Q, existe un conjunto $A \in \mathcal{M}$ tal que $P(A) \neq 0$ y $Q(A) = 0$. Entonces, $P(\{x \in A : \frac{dP}{dR}(x) \neq 0\}) > 0$ mientras que $Q(\{x \in A : \frac{dP}{dR}(x) \neq 0\}) = 0$. Con el convenio adoptado,

$$\int_{\{x \in A: \frac{dP}{dR}(x) \neq 0\}} \log \frac{\frac{dP}{dR}}{\frac{dQ}{dR}} dP = P(\{x \in A : \frac{dP}{dR}(x) \neq 0\}) \cdot (+\infty) = +\infty.$$

□

Proposición 4.1.3. *Se considera el espacio medible $(\mathcal{X}, \mathcal{M})$. La divergencia de Kullback-Leibler de una probabilidad P respecto de una medida finita y positiva Q es mayor o igual que 0. Y, además, $D(P|Q) = 0$ si, y solo si, $P = Q$.*

Demostración. En el caso, $P \ll Q$, como la función $f : [0, \infty) \to \mathbb{R}$, dada por $f(x) = x \log x$ es convexa, aplicando la desigualdad de Jensen se tiene que

$$\int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ = Ef(\frac{dP}{dQ}) \geq f(E\frac{dP}{dQ}) = \int \frac{dP}{dQ} dQ \cdot \log(\int \frac{dP}{dQ} dQ) = f(1) = 0.$$

Como $f(x) = x \log x$ es estrictamente convexa, la desigualdad anterior se alcanza si, y solo si, $\frac{dP}{dQ} = 1$, $Q - c.s.$. Esto es equivalente a que $P = Q$. □

Es importante darse cuenta de que si P no fuese una probabilidad, sino que fuese una medida finita y positiva no nula, es decir, $0 < P(\mathcal{X}) < \infty$, la divergencia de Kullback-Leibler puede ser negativa (si la medida del total es menor que 1), pero está acotada inferiormente por $f(P(\mathcal{X}))$, siendo $f(x) = x \log x$.

Es fácil ver que la divergencia de Kullback-Leibler no es simétrica, y, en consecuencia, no es una distancia en el espacio de probabilidades. Por ejemplo, si en $\mathcal{X} = \{0, 1\}$ se definen las probabilidades P y Q dadas por $P(0) = \frac{1}{2}$, $P(1) = \frac{1}{2}$ y $Q(0) = \frac{1}{4}$, $Q(1) = \frac{3}{4}$. Se comprueba, sin dificultad, que $D(P|Q) \neq D(Q|P)$.

Aun así, las siguientes proposiciones recuerdan los resultados que se tienen para las distancias, y tienen una idea geométrica clara.

4.1.1. Resultados geométricos para la divergencia de K.-L.

Se define el concepto de bola (análogo al conocido para distancias) y de proyección de una medida finita positiva R sobre un conjunto de probabilidades.

Definición 4.1.4. *Sea R una medida finita y positiva (no nula) en un espacio medible $(\mathcal{X}, \mathcal{M})$ y sea $\rho \in (0, \infty]$. Se define la bola, dada por la divergencia de Kullback-Leibler, centrada en R y de radio ρ al conjunto*

$$B(R, \rho) = \{P \text{ probabilidad en } (\mathcal{X}, \mathcal{M}) : D(P|R) < \rho\}.$$

La bola de centro R y radio ρ es el conjunto de probabilidades cuya divergencia de Kullback-Leibler respecto de R es menor que ρ.

Definición 4.1.5. *Sea R una medida finita y positiva (no nula) en $(\mathcal{X}, \mathcal{M})$. Si $\mathcal{B}$ es un subconjunto del espacio de probabilidades en $\mathcal{X}$ tal que $\mathcal{B} \cap B(R, \infty) \neq \emptyset$, se denomina proyección de R sobre $\mathcal{B}$ a una probabilidad Q que satisface*

$$D(Q|R) = \min_{P \in \mathcal{B}} D(P|R).$$

La siguiente proposición garantiza la existencia y unicidad de la proyección en un conjunto de probabilidades convexo y cerrado para la distancia en variación total.

En la demostración de ese resultado, se necesita una desigualdad que relaciona la distancia en variación total con la divergencia de Kullback-Leibler, la desigualdad de Pinsker. La demostración que se da en este trabajo se puede encontrar en [20].

Lema 4.1.6 (Desigualdad de Pinsker)**.** *Sean μ y ν y dos probabilidades en $(\mathcal{X}, \mathcal{M})$. Entonces, se verifica la siguiente desigualdad:*

$$d_{TV}(\mu, \nu) \leq \sqrt{\frac{1}{2} D(\mu|\nu)}. \tag{4.1}$$

Demostración. Si μ no es absolutamente continua respecto de ν, entonces $D(\mu|\nu) = +\infty$ y la desigualdad es obvia. Se estudia el caso en el que $\mu \ll \nu$.

Teniendo en cuenta el convenio $0 \log 0 = 0$, se define la función auxiliar

$$\psi(x) = x \log x - x + 1, \quad x \geq 0.$$

Entonces, se verifica que

$$(x-1)^2 \leq (\frac{4}{3} + \frac{2}{3}x)\psi(x), \quad x \geq 0.$$

Para $x = 0$, es obvio. Para $x > 0$, se comprueba derivando la función $g(x) = (x-1)^2 - (\frac{4}{3} + \frac{2}{3}x)\psi(x)$. Se verifica que $g(1) = 0$, $g'(1) = 0$ y $g''(x) = -\frac{4\psi(x)}{3x} < 0$ para todo $x > 0$. Escribiendo el desarrollo de Taylor de orden 1 de la función g, se tiene que para todo $x > 0$ y $x \neq 1$, existe $\xi \in \mathbb{R}$ con $|\xi - 1| < |x - 1|$ tal que

$$g(x) = g(1) + g'(1)(x-1) + \frac{g''(\xi)}{2}(x-1)^2 = -\frac{4\psi(\xi)}{3\xi}(x-1)^2 \leq 0.$$

Usando el lema 1.2.2,

$$\begin{aligned} d_{TV}(\mu, \nu) &= \frac{1}{2} \int |\frac{d\mu}{d\nu} - 1| d\nu \leq \frac{1}{2} \int \sqrt{(\frac{4}{3} + \frac{2}{3}\frac{d\mu}{d\nu})\psi(\frac{d\mu}{d\nu})}\, d\nu \\ &\leq \frac{1}{2} \sqrt{\int (\frac{4}{3} + \frac{2}{3}\frac{d\mu}{d\nu}) d\nu} \sqrt{\int \psi(\frac{d\mu}{d\nu}) d\nu} = \frac{1}{2}\sqrt{2} \sqrt{\int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu} \\ &= \sqrt{\frac{1}{2} D(\mu|\nu)}. \end{aligned}$$

En la segunda desigualdad, se ha usado la desigualdad de Cauchy-Schwarz. □

Proposición 4.1.7. *Sea R una medida finita y positiva (no nula) en $(\mathcal{X}, \mathcal{M})$. Sea $\mathcal{B}$ un conjunto de probabilidades en $(\mathcal{X}, \mathcal{M})$ tal que $\mathcal{B} \cap B(R, \infty) \neq \emptyset$. Si $\mathcal{B}$ es convexo y cerrado para la distancia en variación total, entonces existe la proyección de R sobre $\mathcal{B}$. Además, la proyección es única.*

Demostración. Por definición de inferior, existe una sucesión de probabilidades $\{P_n\}_n$ de $\mathcal{B}$ tal que $\lim_{n\to\infty} D(P_n|R) = \inf_{P\in\mathcal{B}} D(P|R)$. El objetivo es probar que la sucesión $\{P_n\}_n$ admite una subsucesión $\{P_{n_k}\}_k$ convergente, para la distancia en variación total, hacia una probabilidad $Q \in \mathcal{B}$, que será la proyección de R sobre $\mathcal{B}$.

Se verifica la siguiente igualdad, para todo $n, m \in \mathbb{N}$:

$$D(P_m|R)+D(P_n|R) = 2D(\frac{P_m+P_n}{2}|R)+D(P_m|\frac{P_n+P_m}{2})+D(P_n|\frac{P_n+P_m}{2}). \quad (4.2)$$

Esto se comprueba usando las propiedades de la divergencia de Kullback-Leibler:

$$\begin{aligned}\int \frac{dP_m}{dR}\log\frac{dP_m}{dR}dR+\int \frac{dP_n}{dR}\log\frac{dP_n}{dR}dR&\\ =\int(\frac{dP_m}{dR}+\frac{dP_n}{dR})\log(\frac{1}{2}\frac{dP_m}{dR}+\frac{1}{2}\frac{dP_n}{dR})dR&\\ +\int\frac{dP_m}{dR}\log\frac{dP_m}{dR}dR-\int\frac{dP_m}{dR}\log(\frac{1}{2}\frac{dP_m}{dR}+\frac{1}{2}\frac{dP_n}{dR})dR&\\ +\int\frac{dP_n}{dR}\log\frac{dP_n}{dR}dR-\int\frac{dP_n}{dR}\log(\frac{1}{2}\frac{dP_m}{dR}+\frac{1}{2}\frac{dP_n}{dR})dR.&\end{aligned}$$

Debido a que la función $x\log x$ es convexa, la divergencia de Kullback-Leibler respecto de una probabilidad R también lo es. En particular,

$$D(\frac{P_m+P_n}{2}|R) \leq \frac{1}{2}D(P_m|R)+\frac{1}{2}D(P_n|R).$$

Tomando límites, cuando m y n tienden a ∞:

$$\lim_{n\to\infty} D(\frac{P_m+P_n}{2}|R) \leq \frac{1}{2}\inf_{P\in\mathcal{B}} D(P|R)+\frac{1}{2}\inf_{P\in\mathcal{B}} D(P|R) = \inf_{P\in\mathcal{B}} D(P|R).$$

Por ser $\mathcal{B}$ un conjunto convexo, $\frac{P_m+P_n}{2} \in \mathcal{B}$. Se deduce, de la igualdad 4.2 y del hecho de que la divergencia de una probabilidad respecto de una medida positiva y finita es mayor o igual que 0, que $\lim_{n,m\to\infty} D(P_m|\frac{P_n+P_m}{2}) = 0$ y $\lim_{n,m\to\infty} D(P_n|\frac{P_n+P_m}{2}) = 0$. Usando la desigualdad de Pinsker (4.1), se prueba que la sucesión $\{P_n\}_n$ es de Cauchy, con la distancia en variación total, debido a que se verifica que:

$$\begin{aligned}d_{TV}(P_n,P_m) &\leq d_{TV}(P_m,\frac{P_n+P_m}{2})+d_{TV}(P_n,\frac{P_n+P_m}{2})\\ &\leq \sqrt{\frac{1}{2}D(P_m|\frac{P_n+P_m}{2})}+\sqrt{\frac{1}{2}D(P_n|\frac{P_n+P_m}{2})}.\end{aligned}$$

La convergencia de probabilidades en $(\mathcal{X}, \mathcal{M})$ absolutamente continuas respecto de R, con la distancia en variación total, no es más que la convergencia en $L^1(R)$, un

espacio completo. Esto se debe a la caracterización de la distancia en variación total probada en 1.2.2,

$$d_{TV}(P_n, P_m) = \frac{1}{2}\int |\frac{dP_n}{dR} - \frac{dP_m}{dR}|dR.$$

Por lo tanto, existe el límite en $L^1(R)$ de la sucesión de funciones $\left\{\frac{dP_n}{dR}\right\}_{n=1}^{\infty}$, que es una función no negativa (por serlo todas las funciones de la sucesión). Se denota por h. Equivalentemente, existe una probabilidad Q (absolutamente continua respecto de R) límite de la sucesión $\{P_n\}_{n=1}^{\infty}$, definida de la siguiente forma:

$$Q(A) = \int_A h dR, \quad \forall A \in \mathcal{M}.$$

Además, $Q \in \mathcal{B}$, por ser un conjunto cerrado para la distancia en variación total. La sucesión de funciones $\{\frac{dP_n}{dR}\}_{n=1}^{\infty}$ converge en $L^1(R)$, lo que implica que se puede extraer una subsucesión $\left\{\frac{dP_{n_k}}{dR}\right\}_{k=1}^{\infty}$ que converge hacia $\frac{dQ}{dR}$ puntualmente $R-c.s.$. Las funciones $\frac{dP_{n_k}}{dR}\log\frac{dP_{n_k}}{dR}$ están acotadas inferiormente, para todo $k \in \mathbb{N}$, y convergen $R-c.s.$ hacia $\frac{dQ}{dR}\log\frac{dQ}{dR}$. Se concluye la demostración aplicando el lema de Fatou:

$$D(Q|R) = \int \frac{dQ}{dR}\log\frac{dQ}{dR}dR \leq \lim_{n\to\infty}\int \frac{dP_{n_k}}{dR}\log\frac{dP_{n_k}}{dR}dR = \inf_{P\in\mathcal{B}} D(P|R).$$

La unicidad de la proyección se debe a la convexidad estricta de la divergencia de Kullback-Leibler: Si existen dos proyecciones P y P' de R sobre $\mathcal{B}$ tales que $P \neq P'$, entonces, existe un conjunto A de R-medida estrictamente positiva en el que

$$\frac{dP}{dR}(x) \neq \frac{dP'}{dR}(x), \quad \forall x \in A.$$

Entonces, en A,

$$\left(\frac{1}{2}\frac{dP}{dR} + \frac{1}{2}\frac{dP'}{dR}\right)\log\left(\frac{1}{2}\frac{dP}{dR} + \frac{1}{2}\frac{dP'}{dR}\right) < \frac{1}{2}\frac{dP}{dR}\log\frac{dP}{dR} + \frac{1}{2}\frac{dP'}{dR}\log\frac{dP'}{dR}.$$

Por lo tanto, para la probabilidad $\frac{1}{2}P + \frac{1}{2}P' \in \mathcal{B}$, se tiene que

$$D(\frac{1}{2}P + \frac{1}{2}P'|R) < \inf_{P\in\mathcal{B}} D(P|R).$$

Esto es absurdo. □

El siguiente teorema recuerda al teorema de Pitágoras y al concepto de ortogonalidad que se tiene para distancias. Da una caracterización importante de la proyección en un conjunto convexo.

Teorema 4.1.8. *Sea $\mathcal{B}$ un subconjunto convexo del espacio de probabilidades en $(\mathcal{X}, \mathcal{M})$ tal que $\mathcal{B} \cap B(R, \infty) \neq \emptyset$, y sea R una medida finita y positiva (no nula) en $(\mathcal{X}, \mathcal{M})$. Una probabilidad $Q \in \mathcal{B}$ es la proyección de R sobre $\mathcal{B}$ si, y solo si, para toda probabilidad $P \in \mathcal{B}$,*

$$D(P|R) \geq D(P|Q) + D(Q|R). \tag{4.3}$$

Además, para cada probabilidad $P \in \mathcal{B}$ para la cual existe otra probabilidad $P' \in \mathcal{B}$ y $\alpha \in (0,1]$ tal que

$$Q = \alpha P + (1-\alpha)P',$$

la desigualdad 4.3 es una igualdad. La última condición significa que P pertenece al conjunto de recortes de nivel $1-\alpha$ de Q, tomando como espacio total de probabilidades el conjunto $\mathcal{B}$.

Demostración. $\Rightarrow$) Si Q es la proyección de R sobre $\mathcal{B}$, entonces $D(Q|R) \leq D(P|R)$, para toda $P \in \mathcal{B}$. Sea probabilidad $P \in \mathcal{B}$ fija. Si $D(P|R) = \infty$, la desigualdad se cumple. Si no, se tiene que $P \ll R$. Como $D(Q|R) < \infty$, también se tiene que $Q \ll R$. Se considera el segmento que une P y Q, que está contenido en $\mathcal{B}$ por ser convexo, es decir, se consideran las probabilidades

$$P_\alpha = \alpha P + (1-\alpha)Q, \quad 0 \leq \alpha \leq 1.$$

La derivada de Radon-Nikodym de P_α respecto de R es la combinación lineal de las respectivas derivadas de Radon-Nikodym de P y Q:

$$\frac{dP_\alpha}{dR} = \alpha \frac{dP}{dR} + (1-\alpha)\frac{dQ}{dR}.$$

Como P y Q son probabilidades, se puede suponer que sus respectivas derivadas de Radon-Nikodym toman valores en $[0,\infty)$. Para cada $x \in \mathcal{X}$, la función

$$h : [0,1] \to [0,\infty), \quad h(\alpha) = \frac{dP_\alpha}{dR}(x) \log \frac{dP_\alpha}{dR}(x)$$

está bien definida en $[0,1]$ (se ha definido $\log 0 = -\infty$ y $0 \cdot (-\infty) = 0$) y es derivable. Además, es convexa por ser composición de $\frac{dP_\alpha}{dR}$ (una función lineal en α y, por tanto, convexa) y de la función convexa $t \log t$ definida en $[0,\infty)$. Por lo tanto, los cocientes incrementales de h decrecen si $\alpha \to 0$ y se cumple:

$$\lim_{\alpha \to 0} \frac{1}{\alpha}[h(\alpha) - h(0)] = h'(0) = (\frac{dP}{dR} - \frac{dQ}{dR})(\log \frac{dQ}{dR} + 1).$$

Se tiene que

$$\int [h(1) - h(0)]dR = \int \Big[\frac{dP}{dR} \log \frac{dP}{dR} - \frac{dQ}{dR} \log \frac{dQ}{dR}\Big] dR = D(P|R) - D(Q|R).$$

Como $P, Q \in B(R,\infty)$, entonces $\int [h(1) - h(0)]dR < \infty$. Por el teorema de la convergencia monótona (los cocientes incrementales son funciones decrecientes y acotadas inferiormente, para las cuales la integral del primer cociente incremental, cuando $\alpha = 1$, es finita).

$$\begin{aligned}
\frac{d}{d\alpha} D(P_\alpha|R)\Big|_{\alpha=0} &= \lim_{\alpha \to 0} \int \Big[\frac{dP_\alpha}{dR} \log \frac{dP_\alpha}{dR} - \frac{dQ}{dR} \log \frac{dQ}{dR}\Big] dR \\
&= \int (\frac{dP}{dR} - \frac{dQ}{dR})(\log \frac{dQ}{dR} + 1) dR \\
&= \int \frac{dP}{dR} \log \frac{dQ}{dR} dR + \int \frac{dP}{dR} dR - \int \frac{dQ}{dR} \log \frac{dQ}{dR} dR - \int \frac{dQ}{dR} dR \\
&= \int \log \frac{dQ}{dR} dP + 1 - D(Q|R) - 1 = \int \log \frac{dQ}{dR} dP - D(Q|R).
\end{aligned}$$

Se verifica que $\int \log \frac{dQ}{dR} dP - D(Q|R) \geq 0$, porque si $\int \log \frac{dQ}{dR} dP < D(Q|R)$, entonces la derivada en 0 de $D(P_\alpha|R)$ sería negativa, y esa función sería decreciente en 0, es decir, existiría $\alpha_0 > 0$ tal que

$$D(P_{\alpha_0}|R) < D(P_0|R) = D(Q|R).$$

Esto es absurdo, por ser Q la proyección de R sobre $\mathcal{B}$.

Para concluir la prueba de esta implicación, basta darse cuenta de que

$$\begin{aligned} D(P|R) - D(P|Q) &= \int \frac{dP}{dR} \log \frac{dP}{dR} dR - \int \frac{dP}{dR} \log \frac{\frac{dP}{dR}}{\frac{dQ}{dR}} dR = \int \frac{dP}{dR} \log \frac{dQ}{dR} dR \\ &= \int \log \frac{dQ}{dR} dP. \end{aligned} \tag{4.4}$$

Como se había supuesto $P \in B(R, \infty)$, la resta anterior de las divergencias de Kulback-Leibler está bien definida.

$\Leftarrow$) Recíprocamente, si $D(P|R) \geq D(P|Q) + D(Q|R)$, para toda $P \in \mathcal{B}$, como la divergencia de Kullback-Leibler de una probabilidad P respecto de otra probabilidad Q es siempre positiva, $D(P|R) \geq D(Q|R)$ para toda $P \in \mathcal{B}$. Y, por tanto, Q es la proyección de R en $\mathcal{B}$.

Por último, sea $P \in \mathcal{B}$, tal que, para cierto $\alpha \in (0, 1]$, existe $P' \in \mathcal{B}$ verificando $Q = \alpha P + (1 - \alpha)P'$. Entonces $P \ll Q \ll R$. Por lo tanto, $P \in B(R, \infty)$. Si se supone que $\int \log \frac{dQ}{dR} dP > D(Q|R)$, entonces,

$$\begin{aligned} D(Q|R) &= \int \log \frac{dQ}{dR} dQ \\ &= \int \log \frac{dQ}{dR} d(\alpha P + (1 - \alpha)P') = \alpha \int \log \frac{dQ}{dR} dP + (1 - \alpha) \int \log \frac{dQ}{dR} dP' \\ &> \alpha D(Q|R) + (1 - \alpha) D(Q|R) = D(Q|R). \end{aligned}$$

Se llega a una contradicción. □

4.1.2. Minimizadores de la divergencia de K.-L.

El objetivo de este apartado es caracterizar la proyección de una medida R en un espacio convexo concreto. Primero, se enunciará un teorema general para espacios definidos a partir de una serie de restricciones. Después, se verá que, fijadas dos probabilidades P_1 y P_2 en $\mathbb{R}^d$, el conjunto de probabilidades en el espacio producto cuyas marginales son P_1 y P_2 es un caso particular de los espacios considerados en el teorema. Entonces, se tendrá caracterizada la proyección de cualquier medida finita positiva R sobre $\prod(P_1, P_2)$, que se sabe que existe (porque $\prod(P_1, P_2)$ es convexo y cerrado para la distancia en variación total).

Teorema 4.1.9. *Sea $\{f_i\}_{i \in I}$ un conjunto de funciones medibles de $(\mathcal{X}, \mathcal{M})$ en $\mathbb{R}$ y $\{a_i\}_{i \in I} \subset \mathbb{R}$. Sea Γ el conjunto de probabilidades en el espacio medible $(\mathcal{X}, \mathcal{M})$ para las cuales f_i es integrable para todo $i \in I$, y se cumple que $\int f_i dP = a_i$. Entonces,*

si una medida finita y positiva (no nula) R en $(\mathcal{X}, \mathcal{M})$ tiene proyección Q sobre Γ, la derivada de Radon-Nikodym de Q respecto de R es de la forma

$$\frac{dQ}{dR}(x) = \begin{cases} \exp g(x) & si\ x \notin N, \\ 0 & si\ x \in N, \end{cases} \tag{4.5}$$

donde N es un conjunto tal que $P(N) = 0$, para toda probabilidad $P \in \Gamma \cap B(R, \infty)$, y $g : \mathcal{X} \to \mathbb{R}$ es una función que pertenece a la adherencia del subespacio vectorial de $L^1(Q)$ generado por por la familia de funciones $\{f_i\}_{i\in I}$ y por la función constante 1.

Recíprocamente, si $Q \in \Gamma$ es absolutamente continua respecto de R, con densidad de la forma 4.5, donde g pertenece al subespacio lineal de $L^1(Q)$ generado por $\{f_i\}_{i\in I} \cup \{1\}$, sin tomar la adherencia, entonces Q es la proyección de R sobre Γ.

Demostración. Γ es un conjunto convexo. Si Q es la proyección de R sobre Γ, entonces, $D(Q|R) < \infty$ y, en particular, $Q \ll R$. Sea $N = \{x \in \mathcal{X} : \frac{dQ}{dR}(x) = 0\}$. Si $P \in \Gamma \cap B(R, \infty)$, la desigualdad 4.3, garantiza que $D(P|Q) < \infty$ y, en consecuencia, $P \ll Q$. Como $Q(N) = 0$, entonces $P(N) = 0$.

Además, $\frac{dQ}{dR}(x) = +\infty$ a lo sumo en un conjunto de $R-$probabilidad nula, pudiéndose redefinir la función en este conjunto para que tome valores reales mayores o iguales que 0. Entonces, la función $\log \frac{dQ}{dR} : \mathcal{X} - N \to \mathbb{R}$ está bien definida. Se concluye la demostración de esta parte del teorema probando que la función

$$g(x) := \begin{cases} \log \frac{dQ}{dR}(x) & \text{si } x \notin N, \\ 0 & \text{si } x \in N \end{cases}$$

pertenece a la adherencia del subespacio vectorial de $L^1(Q)$ generado por la familia $\{f_i\}_{i\in I}$ y por la función constante 1, es decir, $g \in \overline{\langle 1, \{f_i\}_{i\in I}\rangle}$. En estas condiciones, $g : \mathcal{X} \to R$ verifica la igualdad 4.5.

Se razona por reducción al absurdo. Si $g \notin \overline{\langle 1, \{f_i\}_{i\in I}\rangle}$, por el teorema de Hahn-Banach, existe un funcional lineal y continuo de $T : L^1(Q) \to \mathbb{R}$ tal que $T|_{\langle 1,\{f_i\}_{i\in I}\rangle} \equiv 0$ y $T(g) \neq 0$. Como Q es una probabilidad (medida finita), el dual de $L^1(Q)$ está formado por los funcionales

$$T : L^1(Q) \to \mathbb{R}, \quad T(f) = \int f h dQ, \ \text{ con } h \in L^\infty(Q).$$

Por lo tanto, existe una función $h \in L^\infty(Q)$ tal que $\int ghdQ \neq 0$, $\int hdQ = 0$ y para todo $i \in I$, $\int f_i h dQ = 0$.

Por otro lado, se define al probabilidad

$$\tilde{P}(A) = \int_A (1 + \frac{h(x)}{\|h\|_\infty}) dQ, \quad \forall A \in \mathcal{M}.$$

Es fácil comprobar que $\tilde{P}$ es una probabilidad: $\tilde{P}$ es medida por construcción y se cumple:

- $\tilde{P}(\mathcal{X}) = \int dQ + \frac{1}{\|h\|_\infty} \int hdQ = 1 + 0 = 1.$

- $1 + \frac{h(x)}{\|h\|_\infty} \geq 0$, $Q - c.s. \Rightarrow \tilde{P}(A) \geq 0$, $\forall A \in \mathcal{M}$.

Como $\int f_i d\tilde{P} = \int f_i(1 + \frac{h(x)}{\|h\|_\infty})dQ = \int f_i dQ + \frac{1}{\|h\|_\infty}\int h(x)dQ = a_i$, para todo $i \in I$, entonces $\tilde{P} \in \Gamma$. Análogamente, la probabilidad $\tilde{P}'$ definida por

$$\tilde{P}'(A) = \int_A (1 - \frac{h(x)}{\|h\|_\infty})dQ, \quad \forall A \in \mathcal{M}.$$

también pertenece a Γ. Como $Q = \frac{1}{2}(\tilde{P}+\tilde{P}')$, teniendo en cuenta que Γ es un conjunto convexo, se puede aplicar el teorema 4.1.8 y se tiene que $D(\tilde{P}|R) - D(\tilde{P}|Q) = D(Q|R)$. Como $\tilde{P} \ll Q \ll R$, la igualdad anterior equivale a

$$\int \log \frac{dQ}{dR} d\tilde{P} = \int \log \frac{dQ}{dR} dQ \Leftrightarrow \int \log \frac{dQ}{dR}(1 + \frac{h(x)}{\|h\|_\infty})dQ = \int \log \frac{dQ}{dR} dQ$$

si, y solo si, $\int \log \frac{dQ}{dR} h dQ = 0$, es decir, $\int ghdQ = 0$. Esto es absurdo.

Para probar el recíproco, se supone que $Q \in \Gamma$ es absolutamente continua respecto de R y su derivada de Radon-Nikodym tiene la forma 4.5. Entonces, $g(x) = \lambda + \sum_{j=1}^{n} f_{i_j}(x)$, con $\lambda \in \mathbb{R}$. Se considera $P \in \Gamma \cap B(R, \infty)$. Se tiene que $\log \frac{dQ}{dR} = g$, $P - c.s.$ Entonces,

$$\int \log \frac{dQ}{dR} dP = \int g dP = \lambda + \sum_{j=1}^{n} \int f_{i_j} dP = \lambda + \sum_{j=1}^{n} a_{i_j} = c \in \mathbb{R}, \quad \forall P \in \Gamma \cap B(R, \infty).$$

En particular, $D(Q|R) = \int \log \frac{dQ}{dR} dQ = c$. Repitiendo el razonamiento de 4.4,

$$D(P|R) - D(P|Q) = \int \log \frac{dQ}{dR} dP = c = D(Q|P).$$

Aplicando el teorema 4.1.8, se concluye que Q es la proyección de R sobre Γ.

□

El espacio de probabilidades en $\mathbb{R}^d \times \mathbb{R}^d$ cuyas marginales son dos probabilidades P_1 y P_2 fijas es un caso particular de los espacios considerados en el teorema anterior. El último corolario de esta sección garantiza la existencia de probabilidades en $\prod(P_1, P_2)$ con una forma concreta (en términos de su derivada de Radon-Nikodym respecto de la medida producto $P_1 \otimes P_2$). En el problema de transporte entrópico, que se formalizará en la siguiente sección, se necesita la existencia de estas probabilidades para construir la probabilidad $\pi \in \prod(P_1, P_2)$ que minimiza el coste de transporte entrópico.

Primero, se prueba un lema que se usará en la demostración del corolario.

Lema 4.1.10. *Sean P_1 y P_2 probabilidades en $(\mathbb{R}^d, \beta^d)$. Se supone que $\pi \in \prod(P_1, P_2)$. Entonces, el subespacio*

$$V := \{f \in L^1(\pi) : f(x, y) = f(x)\} + \{g \in L^1(\pi) : g(x, y) = g(y)\}$$

es cerrado en $L^1(\pi)$.

Demostración. Como se ha visto en el apéndice F sobre desintegración de medidas, existe un sistema de probabilidades $\{\pi_{Y|X=x}\}_{x\in\mathbb{R}^d}$ en $\mathbb{R}^d$ tal que

$$\pi(B) = \int_{\mathbb{R}^d} \pi_{Y|X=x}(B_x)\, dP_1(x), \quad \forall B \in \beta^d \times \beta^d.$$

Si $f \in L^1(\pi)$, se define la esperanza condicionada dado $X = x$ a la función de $\mathbb{R}^d$ en $\mathbb{R}$ definida como:

$$E(f \mid X)(x) := \int_{\mathbb{R}^d} f(x,y)\, d\pi_{Y|X=x}(y), \quad \forall x \in \mathbb{R}^d.$$

Entonces, el funcional lineal

$$T_X : L^1(\pi) \to \{f \in L^1(\pi) : f(x,y) = f(x)\}, \qquad T_X(f) = E(f \mid X)$$

está bien definido y es continuo, ya que

$$\begin{aligned}
\|E(f \mid X)\|_1 &\leq \int_{\mathbb{R}^d} |E(f \mid X)(x)| d\pi(x,y) = \int_{\mathbb{R}^d} |E(f \mid X)(x)| dP_1(x) \\
&= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} f(x,y')\, d\pi_{Y|X=x}(y') \right| dP_1(x) \\
&\leq \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |f(x,y')|\, d\pi_{Y|X=x}(y') \right) dP_1(x) \\
&= \int_{\mathbb{R}^d} |f(x,y)| d\pi(x,y) = \|f\|_1.
\end{aligned}$$

Análogamente, existe un sistema de probabilidades $\{\pi_{X|Y=y}\}_{y\in\mathbb{R}^d}$ en $\mathbb{R}^d$ tal que

$$\pi(B) = \int_{\mathbb{R}^d} \pi_{X|Y=y}(B_y)\, dP_1(y), \quad \forall B \in \beta^d \times \beta^d.$$

Si $g \in L^1(\pi)$, se define la esperanza condicionada dado $Y = y$ a la función de $\mathbb{R}^d$ en $\mathbb{R}$ definida como:

$$E(g \mid Y)(y) := \int_{\mathbb{R}^d} g(x,y)\, d\pi_{X|Y=y}(x), \quad \forall y \in \mathbb{R}^d.$$

Entonces, el funcional lineal

$$T_Y : L^1(\pi) \to \{g \in L^1(\pi) : g(x,y) = g(y)\}, \qquad T_Y(g) = E(g \mid Y)$$

está bien definido y es continuo.

Sea $u \in L^1(\pi)$ tal que existe una sucesión de funciones $\{\tilde{f}_n(x)+\tilde{g}_n(y)\}_{n=1}^{\infty} \subset L^1(\pi)$ con límite la función u (con la norma de $L^1(\pi)$). Si $f_n(x) := \tilde{f}_n(x) + \int g_n(y) dP_2(y)$ y $g_n(y) := \tilde{g}_n(y) - \int g_n(y) dP_2(y)$, se verifica que

$$f_n(x) + g_n(y) = \tilde{f}_n(x) + \tilde{g}_n(y) \xrightarrow[n\to\infty]{} u(x,y) \quad \text{en } L^1(\pi).$$

Entonces, utilizando la continuidad de T_X,

$$\begin{aligned} T_X(f_n(x)+g_n(y)) &= f_n(x) + \int g_n(y)\, d\pi_{Y|X=x}(y) = f_n(x) + \int g_n(y) d\pi(x,y) \\ &= f_n(x) \xrightarrow[n\to\infty]{} T_X(u)(x) \quad \text{en } L^1(\pi). \end{aligned}$$

Usando la continuidad de T_Y,

$$\begin{aligned} T_Y(f_n(x)+g_n(y)) &= \int f_n(x)\, d\pi_{X|Y=y}(x) + g_n(y) = \int f_n(x)\, d\pi(x,y) + g_n(y) \\ &= \int f_n(x)\, dP_1(x) + g_n(y) \\ &= a_n + g_n(y) \xrightarrow[n\to\infty]{} T_Y(u)(y) \quad \text{en } L^1(\pi). \end{aligned}$$

Por otro lado,

$$g_n(y) \xrightarrow[n\to\infty]{} u(x,y) - T_X(u)(x) \quad \text{en } L^1(\pi)$$

Entonces $\{a_n\}_{n=1}^{\infty} \subset \mathbb{R}$ converge hacia cierto $a \in \mathbb{R}$. Y, se tiene que

$$g_n(y) \xrightarrow[n\to\infty]{} T_Y(u)(y) - a \quad \text{en } L^1(\pi).$$

Por la unicidad del límite, se concluye que

$$u(x,y) = T_X(u)(x) + (T_Y(u)(y) - a),$$

con $T_X(u)(x) \in L^1(\pi)$ y $(T_Y(u)(y) - a) \in L^1(\pi)$.

□

Corolario 4.1.11. *Sean P_1 y P_2 probabilidades en $(\mathbb{R}^d, \beta^d)$. Se denota por $P_1 \otimes P_2$ a la probabilidad producto en $(\mathbb{R}^d \times \mathbb{R}^d, \beta^d \otimes \beta^d)$. Se considera una función $\tilde{c}(x,y) \in L^1(P_1 \otimes P_2)$ tal que $\tilde{c}(x,y) > 0$ para todo $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$. Entonces, existe una probabilidad π en el espacio producto con marginales P_1 y P_2, tal que $\pi \ll P_1 \otimes P_2$, que verifica:*

$$\frac{d\pi}{d(P_1 \otimes P_2)}(x,y) = \tilde{c}(x,y) \exp[f(x) + g(y)] \tag{4.6}$$

donde $(f,g) \in L^1(P_1) \times L^1(P_2)$.

Demostración. Se denota por Γ al espacio de probabilidades en $(\mathbb{R}^d \times \mathbb{R}^d, \beta^d \otimes \beta^d)$ con marginales P_1 y P_2. Entonces $\pi \in \Gamma$ si, y solo si, verifica las restricciones siguientes:

$$\int f(x) d\pi = \int f(x) dP_1, \quad \forall f \in L^1(P_1),$$

$$\int g(y) d\pi = \int g(y) dP_2, \quad \forall g \in L^1(P_2).$$

Se define una medida finita y positiva de la siguiente forma:

$$R(E) = \int_E \tilde{c}(x,y) d(P_1 \otimes P_2), \quad \forall E \in \beta^d \otimes \beta^d.$$

Por la proposición 4.1.7, existe la proyección π de R sobre este espacio, y, aplicando el teorema 4.1.9, se deduce que $\pi \in \prod(P_1, P_2)$ es absolutamente continua respecto de R y su derivada de Radon-Nikodym respecto de R es de la forma 4.5.

$$\frac{d\pi}{dR}(x) = \begin{cases} \exp h(x,y) & \text{si } x \notin N, \\ 0 & \text{si } x \in N. \end{cases}$$

donde N es un subconjunto tal que $\mu(N) = 0$, para toda $\mu \in \prod(P,Q) \cap B(R, \infty)$ y $h(x,y)$ pertenece a la adherencia del subespacio lineal de $L^1(\pi)$ generado por las funciones de $L^1(\pi)$ de una sola variable, es decir, si

$$\begin{aligned} V :&= \langle \{f \in L^1(\pi) : f(x,y) = f(x)\} \cup \{g \in L^1(\pi) : g(x,y) = g(y)\} \rangle \\ &= \{f \in L^1(\pi) : f(x,y) = f(x)\} + \{g \in L^1(\pi) : g(x,y) = g(y)\}, \end{aligned}$$

entonces, $h \in \bar{V}$. Pero el subespacio vectorial V es cerrado en $L^1(\pi)$, se probó en el lema 4.1.10. Por lo tanto, $h(x,y) = f(x) + g(y)$, para ciertas funciones $(f,g) \in L^1(P_1) \times L^1(P_2)$. Entonces,

$$\frac{d\pi}{d(P_1 \otimes P_2)}(x,y) = \tilde{c}(x,y) \exp[f(x) + g(y)], \quad \forall (x,y) \notin N.$$

Se sabe que $R(N) = 0$. A partir de la definición de R y del hecho de que la función $c : \mathbb{R}^d \times \mathbb{R}^d \to (0, +\infty)$ es estrictamente positiva, se tiene que $R(N) = 0$ si, y solo si, $P_1 \otimes P_2(N) = 0$. Entonces, (como la derivada de Radon-Nikodym π es única salvo conjuntos de $P_1 \otimes P_2$-probabilidad nula) se puede concluir que

$$\frac{d\pi}{d(P_1 \otimes P_2)}(x,y) = \tilde{c}(x,y) \exp[f(x) + g(y)], \quad \forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

□

4.2. Problema del transporte entrópico

Se formula el problema de transporte entrópico para el coste cuadrático:

Sea $\varepsilon > 0$. Se consideran P y Q dos probabilidades en $(\mathbb{R}^d, \beta^d)$ con momentos de orden 2 finitos. El problema de transporte entrópico para el coste cuadrático consiste en calcular

$$\mathcal{W}_{2,\varepsilon}^2(P,Q) = \inf_{\pi \in \prod(P,Q)} \left[\int \frac{\|x-y\|^2}{2} d\pi(x,y) + \varepsilon D(\pi | P \otimes Q) \right]. \tag{4.7}$$

Al contaminar la distancia de Wasserstein $\mathcal{W}_2$ con la divergencia de Kullback-Leibler, deja de ser una métrica. Esto se debe a que la divergencia de Kullback-Leibler no es simétrica. Además, $\mathcal{W}_{2,\varepsilon}^2(P,P) \neq 0$.

En este trabajo, se estudia el coste entrópico con el fin de comparar si dos recortes de probabilidades distintas son iguales. Por eso, perder la simetría de la distancia de Wasserstein no es relevante. Por el contrario, sí se puede adaptar la expresión del coste de transporte entrópico $\mathcal{W}_{2,\varepsilon}^2(P,Q)$ (modificándola) para que sea igual a 0 si, y solo si, $P = Q$.

Además, como ya se ha mencionado en el capítulo 1, el problema de transporte entrópico presenta ciertas ventajas respecto al problema de transporte estudiado en el capítulo 3. En la práctica, calcular la distancia de Wasserstein entre dos probabilidades es costoso y se calcula aproximando esa distancia por la empírica (a partir de una muestra). Esto hace que no sea resistente a la maldición de la dimensionalidad. En cambio, en esta sección se va a deducir una iteración de punto fijo para calcular el coste de transporte entrópico entre dos probabilidades, lo que elimina la dependencia de una muestra en los cálculos y hace que el problema de transporte entrópico no se vea afectado por la maldición de la dimensionalidad.

Por estas razones, el transporte entrópico es una buena alternativa para comparar dos probabilidades P y Q. Primero, se deduce una fórmula de dualidad para este problema, que permite probar la existencia de una probabilidad $\pi_0 \in \prod(P,Q)$ tal que

$$\mathcal{W}_{2,\varepsilon}^2(P,Q) = \int \frac{\|x-y\|^2}{2} d\pi_0(x,y) + \varepsilon D(\pi_0 | P \otimes Q).$$

Para esto, se usan los resultados que se han probado previamente sobre la divergencia de Kullback-Leibler. La siguiente proposición es la versión de la dualidad de Kantorovich para el transporte entrópico.

Teorema 4.2.1. *Sea $\varepsilon > 0$. Se consideran P y Q dos probabilidades en $(\mathbb{R}^d, \beta^d)$ con momentos de orden 2 finitos. Existe una probabilidad $\tilde{\pi}$ en $(\mathbb{R}^d \times \mathbb{R}^d, \beta^d \times \beta^d)$ cuyas marginales son P y Q que minimiza la expresión 4.7, es decir,*

$$\mathcal{W}_{2,\varepsilon}^2(P,Q) = \int \frac{\|x-y\|^2}{2} d\tilde{\pi}(x,y) + \varepsilon D(\tilde{\pi} | P \otimes Q).$$

Además, se verifica la fórmula de dualidad

$$\mathcal{W}_{2,\varepsilon}^2(P,Q) = \sup_{(f,g)\in L^1(P)\times L^1(Q)} \left[\int f\, dP + \int g\, dQ - \varepsilon \int \gamma(x,y) dP(x) dQ(y) \right] + \varepsilon, \tag{4.8}$$

con

$$\gamma(x,y) = \exp\Big(- \frac{\frac{1}{2}\|x-y\|^2 + f(x) + g(y)}{\varepsilon} \Big).$$

Demostración. Como $D(\tilde{\pi}|P\otimes Q) = \infty$ si π no es absolutamente continua respecto de $P \otimes Q$, basta con minimizar la expresión del enunciado en el subconjunto de probabilidades de $\prod(P,Q)$ que sean absolutamente continuas respecto de la medida producto $P \otimes Q$.

Por lo tanto, se supone que $\pi(x,y) = r(x,y)dP(x)dQ(y)$, es decir, la derivada de Radon-Nikodym de π respecto de $P \otimes Q$ es la función $r : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty]$. Como π es una medida finita, se puede suponer que r toma valores reales. Se tiene que

$$\begin{aligned}
&\int \frac{\|x-y\|^2}{2} d\pi(x,y) + \varepsilon D(\pi | P\otimes Q) \\
&= \int \left[\frac{1}{2}\|x-y\|^2 r(x,y) + \varepsilon r(x,y) \log r(x,y) \right] dP(x) dQ(y) \\
&= \varepsilon \int r(x,y) \log \frac{r(x,y)}{\exp(-\frac{\frac{1}{2}\|x-y\|^2}{\varepsilon})} dP(x) dQ(y).
\end{aligned}$$

Se comprueba, a continuación, que se verifica una fórmula de dualidad 4.8. Si $f \in L^1(P)$ y $g \in L^1(Q)$, entonces

$$\int f(x)dP(x) = \int f(x)d\pi(x,y) = \int f(x)r(x,y)dP(x)dQ(y),$$

$$\int g(y)dQ(y) = \int g(y)d\pi(x,y) = \int g(y)r(x,y)dP(x)dQ(y).$$

Por lo tanto,

$$\mathcal{W}^2_{2,\varepsilon}(P,Q) - \int f dP - \int g dQ = \varepsilon \int r(x,y) \log \frac{r(x,y)}{\exp(-\frac{\frac{1}{2}\|x-y\|^2+f(x)+g(y)}{\varepsilon})} dP(x)dQ(y). \tag{4.9}$$

Para escribirlo de forma más compacta, se había denotado

$$\gamma(x,y) = \exp\Big(-\frac{\frac{1}{2}\|x-y\|^2 + f(x) + g(y)}{\varepsilon}\Big).$$

Es conocido que $s \log s \geq s-1$, $\forall x > 0$, ya que la función $h(s) = s\log s$ definida en $(0,\infty)$ es estrictamente convexa y la recta tangente a h en $s_0 = 1$ viene dada por la ecuación $s-1$. Se da la igualdad solo cuando $s = 1$. Entonces,

$$\begin{aligned}
(4{,}9) &= \varepsilon \int \gamma(x,y)\frac{r(x,y)}{\gamma(x,y)} \log \frac{r(x,y)}{\gamma(x,y)} dP(x)dQ(y) \\
&\geq \varepsilon \int \gamma(x,y)\big(\frac{r(x,y)}{\gamma(x,y)} - 1\big) dP(x)dQ(y) \\
&= \varepsilon \int r(x,y)dP(x)dQ(y) - \varepsilon \int \gamma(x,y)dP(x)dQ(y) \\
&= \varepsilon - \varepsilon \int \exp\Big(-\frac{\frac{1}{2}\|x-y\|^2 + f(x) + g(y)}{\varepsilon}\Big) dP(x)dQ(y).
\end{aligned}$$

Entonces,

$$\mathcal{W}^2_{2,\varepsilon}(P,Q) \geq \varepsilon + \int f dP + \int g dQ - \varepsilon \int \gamma(x,y)dP(x)dQ(y),\ \forall (f,g) \in L^1(P)\times L^1(Q).$$

Tomando superior en $(f,g) \in L^1(P) \times L^1(Q)$,

$$\mathcal{W}^2_{2,\varepsilon}(P,Q) \geq \sup_{(f,g)\in L^1(P)\times L^1(Q)} \Big[\int f\, dP + \int g\, dQ - \varepsilon \int \gamma(x,y)dP(x)dQ(y) + \varepsilon\Big].$$

Sea $\tilde{\pi} \in \prod(P,Q)$ tal que $d\tilde{\pi}(x,y) = r(x,y)dP(x)dQ(y)$ con

$$r(x,y) = \exp\Big(-\frac{\frac{1}{2}\|x-y\|^2 + \tilde{f}(x) + \tilde{g}(y)}{\varepsilon}\Big),$$

para ciertas funciones $\tilde{f} \in L^1(P)$ y $\tilde{g} \in L^1(Q)$. La existencia de esta probabilidad se ha probado en el corolario 4.1.11.

Con la probabilidad $\tilde{\pi} \in \prod(P,Q)$ y las funciones $(\tilde{f},\tilde{g}) \in L^1(Q)$ se alcanza la igualdad de la fórmula de dualidad 4.8. Y, por lo tanto, $\tilde{\pi}$ es la probabilidad en la que se alcanza el mínimo del coste de transporte. □

En la demostración del teorema anterior se ha deducido una caracterización de la probabilidad óptima que minimiza el coste de transporte entrópico (en el caso cuadrático). Esto se recoge en el siguiente corolario.

Corolario 4.2.2. *Sea $\varepsilon > 0$. Se consideran P y Q dos probabilidades en $(\mathbb{R}^d, \beta^d)$ con momentos de orden 2 finitos. Sea $\pi \in \prod(P,Q)$ absolutamente continua respecto de $P \otimes Q$. Es decir, $d\pi(x,y) = r(x,y)dP(x)dQ(y)$, donde se denota por $r(x,y)$ a la derivada de Radon-Nikodym de π respecto de $P \otimes Q$. Entonces, π es óptima para el problema de transporte entrópico con el coste cuadrático, es decir,*

$$\mathcal{W}_{2,\varepsilon}^2(P,Q) = \inf_{\pi \in \prod(P,Q)} \left[\int \frac{\|x-y\|^2}{2} d\pi(x,y) + \varepsilon D(\pi | P \otimes Q) \right] \tag{4.10}$$

si, y solo si,

$$r(x,y) = \exp\Big(- \frac{\frac{1}{2}\|x-y\|^2 + f_0(x) + g_0(y)}{\varepsilon} \Big),$$

donde $f_0 \in L^1(P)$ y $g_0 \in L^1(Q)$.

Para finalizar el capítulo, y el trabajo, se estudia la forma de calcular la probabilidad óptima del problema de transporte entrópico planteado. Con la caracterización del corolario 4.2.2, bastaría con hallar un par de funciones $(f_0, g_0) \in L^1(P) \times L^1(Q)$ que verifique que

$$r(x,y) = \exp\Big(- \frac{\frac{1}{2}\|x-y\|^2 + f_0(x) + g_0(y)}{\varepsilon} \Big)$$

es la derivada de Radon-Nikodym de cierta probabilidad π respecto de $P \otimes Q$.

Esas funciones se pueden calcular mediante una iteración de punto fijo. Como ya se ha mencionado, este es el resultado que hace que el coste de transporte entrópico tenga ventajas importantes, en la práctica, con respecto a la distancia de Wasserstein. Se enuncia en la siguiente proposición.

Proposición 4.2.3. *Sea $\varepsilon > 0$. Se consideran P y Q dos probabilidades en $(\mathbb{R}^d, \beta^d)$ con momentos de orden 2 finitos. Un par de funciones $(f_0, g_0) \in L^1(P) \times L^1(Q)$ es óptimo para el problema dual de transporte entrópico (con el coste cuadrático), es decir,*

$$\begin{aligned} &\int f_0\, dP + \int g_0\, dQ - \varepsilon \int \exp\Big(- \frac{\frac{1}{2}\|x-y\|^2 + f_0(x) + g_0(y)}{\varepsilon} \Big) dP(x)dQ(y) + \varepsilon \\ &= \sup_{\mathcal{L}} \left[\int f\, dP + \int g\, dQ - \varepsilon \int \exp\Big(- \frac{\frac{1}{2}\|x-y\|^2 + f(x) + g(y)}{\varepsilon} \Big) dP(x)dQ(y) + \varepsilon \right], \end{aligned}$$

donde $\mathcal{L} = \{(f,g) : f \in L^1(P), g \in L^1(Q)\}$, si, y solo si, (f_0, g_0) verifica

$$\begin{cases} f(x) = -\varepsilon \log\Big(\displaystyle\int \exp(\frac{g(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}) dQ(y) \Big), \\ g(x) = -\varepsilon \log\Big(\displaystyle\int \exp(\frac{f(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon}) dP(x) \Big). \end{cases} \tag{4.11}$$

Demostración. Sea $d\pi(x,y) = r(x,y)dP(x)dQ(y)$ con

$$r(x,y) = \exp\Big(\frac{-\frac{1}{2}\|x-y\|^2 + f_0(x) + g_0(y)}{\varepsilon}\Big),$$

para ciertas funciones $(f_0, g_0) \in L^1(P) \times L^1(Q)$. Se verifica que:

$$\begin{cases} \int r(x,y)dQ(y) = 1, \\ \int r(x,y)dP(x) = 1. \end{cases} \tag{4.12}$$

Teniendo esta propiedad, es fácil deducir la iteración del enunciado. Se prueba la segunda igualdad:

$$\begin{aligned} \int_B dQ(y) = Q(B) = \pi(\mathbb{R}^d \times B) &= \int_{\mathbb{R}^d \times B} r(x,y)dP(x)dQ(y) \\ &= \int_B \Big(\int_{\mathbb{R}^d} r(x,y)dP(x)\Big)dQ(y), \quad \forall B \in \beta^d. \end{aligned}$$

De aquí se deduce que $\int r(x,y)dP(x) = 1$. Análogamente, $\int r(x,y)dQ(y) = 1$. Por lo tanto, reescribiendo 4.12,

$$\begin{cases} \exp(\frac{f(x)}{\varepsilon}) \int \exp(\frac{g(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dQ(y) = 1, \\ \exp(\frac{g(y)}{\varepsilon}) \int \exp(\frac{f(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dP(x) = 1. \end{cases}$$

Entonces,

$$\begin{cases} f(x) = -\varepsilon \log\Big(\int \exp(\frac{g(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dQ(y)\Big), \\ g(y) = -\varepsilon \log\Big(\int \exp(\frac{f(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dP(x)\Big). \end{cases}$$

Recíprocamente, si (f_0, g_0) verifica 4.11, entonces $d\pi(x,y) = r(x,y)dP(x)dQ(y)$ define una probabilidad en $\mathbb{R}^d \times \mathbb{R}^d$ con marginales P y Q (por la propiedad 4.12). □

Se denota por K al siguiente operador:

$$K(f,g) = \begin{pmatrix} -\varepsilon \log\Big(\int \exp(\frac{g(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dQ(y)\Big) \\ -\varepsilon \log\Big(\int \exp(\frac{f(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dP(x)\Big) \end{pmatrix}$$

Se ha comprobado que (f_0, g_0) es un punto fijo del operador K. Por lo tanto, se pueden calcular a partir de la iteración

$$\begin{cases} f_{n+1}(x) = -\varepsilon \log\Big(\int \exp(\frac{g_n(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dQ(y)\Big), \\ g_{n+1}(x) = -\varepsilon \log\Big(\int \exp(\frac{f_n(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon})dP(x)\Big). \end{cases}$$

Esta iteración se conoce como algoritmo de Sinkhorn. La versión discreta de este algoritmo supuso un salto en cuanto a la posibilidad de calcular de forma eficiente el transporte óptimo (entrópico) entre probabilidades. Esto se desarrolla ampliamente en [15]. El tratamiento del problema de transporte óptimo desde el punto de vista computacional queda fuera de los objetivos de este TFG.

Conclusiones

En el contexto del Aprendizaje "Justo", el primer paso para entrenar reglas de clasificación que no dependan de un atributo protegido binario $S \in \{0, 1\}$ consiste en estimar la parte común entre las dos distribuciones de probabilidad correspondientes a los subgrupos en los que se divide la población según dicho atributo. Formalmente, se ha comprobado que el nivel al que dos poblaciones están relacionadas se puede caracterizar a partir de los conjuntos de recortes y la distancia en variación total entre sus respectivas distribuciones. Por eso, este trabajo está motivado por un algoritmo propuesto por H. Inouzh, cuyo objetivo es estimar cotas de la distancia en variación total entre dos distribuciones de probabilidad $P = \mathcal{L}(X|S = 0)$ y $Q = \mathcal{L}(X|S = 1)$, buscando elementos comunes a sus conjuntos de recortes. Dicho algoritmo permite calcular el menor valor de α, en una partición del intervalo $[0, 1]$, tal que $d_{TV}(P, Q) \leq \alpha$. El desarrollo de la aplicación que motivó este trabajo, a la vez que el proyecto de beca de colaboración con el Departamento de Estadística e Investigación Operativa, está todavía por terminar y excede a los objetivos del TFG. Por lo tanto, aunque este trabajo tiene una motivación aplicada, se ha dedicado exclusivamente a estudiar las herramientas teóricas involucradas en el algoritmo.

En primer lugar, se ha propuesto utilizar la distancia de Wasserstein $\mathcal{W}_2$ como métrica para calcular los recortes óptimos entre P y Q, para cierto nivel $\alpha \in (0, 1)$, aprovechando su continuidad con respecto a aproximaciones empíricas y las buenas propiedades topológicas de los conjuntos de recortes respecto de esta métrica. Por eso, se ha estudiado el problema de transporte óptimo, demostrando la existencia de una probabilidad que minimiza el coste. Además, a partir de la fórmula de dualidad de Kantorovich, se ha llegado a caracterizar las aplicaciones de transporte óptimo: gradientes de funciones convexas.

No obstante, el alto coste computacional de calcular la distancia de Wasserstein empírica y su sensibilidad a la maldición de la dimensionalidad han motivado la introducción del coste de transporte entrópico $\mathcal{W}_{2,\varepsilon}$, que contamina la distancia de Wasserstein con la divergencia de Kullback-Leibler, para mejorar la eficiencia y la robustez del método, gracias al algoritmo de Sinkhorn.

Por último, se ha considerado usar un test de contraste de hipótesis basado en la Máxima Discrepancia en Media (MMD) para contrastar si los recortes óptimos P_α y Q_α son iguales, para cierto nivel $\alpha \in (0, 1)$. En caso afirmativo, se tendría que $d_{TV}(P, Q) \leq \alpha$. Con esta finalidad, se han estudiado los núcleos, que son resistentes a la maldición de la dimensionalidad, y sus ventajas en estadística: trabajar conociendo únicamente los productos internos de los datos en un espacio de Hilbert.

En conclusión, este trabajo me ha servido como introducción a un campo en el que algunas herramientas tales como los recortes, las métricas de máxima discre-

pancia media o el problema de transporte óptimo, en sus distintas versiones, son un elemento esencial. Lo que he aprendido con este TFG me ha puesto en condiciones adecuadas para poderme enfrentar a este tipo de problemas, en futuros proyectos.

Apéndice A

Teorema de Radon-Nikodym

A lo largo de varios capítulos, se trabaja con el concepto de medida absolutamente continua respecto de otra medida de referencia y se habla de la derivada de Radon-Nikodym. En concreto, este concepto es imprescindible en la definición de recorte de una probabilidad (1.1.3) y a lo largo de todo el capítulo 4, para definir la divergencia de Kullback-Leibler y el transporte entrópico. Se recuerdan, a continuación, las definiciones y teoremas necesarios en este trabajo, vistos en la asignatura de Análisis Real del grado. Se pueden encontrar las demostraciones en [16].

Definición A.0.1. *Sean λ y μ dos medidas positivas en el espacio medible $(\mathcal{X}, \mathcal{M})$. Se dice que λ es absolutamente continua con respecto de μ, y se escribe*

$$\lambda \ll \mu,$$

si para cualquier $E \in \mathcal{M}$ tal que $\mu(E) = 0$ se tiene que $\lambda(E) = 0$.

Si μ es una medida positiva, el teorema de Radon-Nikodym caracteriza cómo son todas las medidas absolutamente continuas respecto de μ. Se enuncia aquí una versión simplificada de dicho teorema (ya que se recogen en esta sección solo los resultados imprescindibles para el trabajo).

Teorema A.0.2 (Radon-Nikodym)**.** *Sean μ una medida positiva y λ una medida positiva σ-finita en $(\mathcal{X}, \mathcal{M})$ tal que $\lambda \ll \mu$. Entonces, existe una función*

$$h : \mathcal{X} \to [0, +\infty]$$

integrable respecto de μ tal que

$$\lambda(E) = \int_E h d\mu,$$

para todo conjunto $E \in \mathcal{M}$. Cualquier otra función que verifique estas propiedades coincide con h, $\mu - c.s.$

La función h definida en el teorema se denota por

$$\frac{d\lambda}{d\mu}.$$

Si la medida λ es finita (es el caso de las probabilidades), su derivada de Radon-Nikodym respecto de μ toma valores reales salvo en un conjunto de medida μ nula.

Por último, se enuncia la regla de la cadena, que se usará para dar definiciones equivalentes de la divergencia de Kullback-Leibler, en 4.1.2.

Proposición A.0.3 (Regla de la cadena)**.** *Sean λ, ν y μ tres medidas positivas σ-finitas en $(\mathcal{X}, \mathcal{M})$ tales que $\lambda \ll \nu$ y $\nu \ll \mu$. Se verifica:*

1. *Si $g : \mathcal{X} \to \mathbb{C}$ es $\mathcal{M}$-medible, entonces*

$$\int g \, d\lambda = \int g \, \frac{d\lambda}{d\nu} d\nu.$$

   En particular, si $g \in L_1(\lambda)$, entonces $g \, \frac{d\lambda}{d\nu} \in L_1(\nu)$.

2. *Se verifica que $\lambda \ll \mu$ y*

$$\frac{d\lambda}{d\mu} = \frac{d\lambda}{d\nu} \frac{d\nu}{d\mu}, \quad \mu - c.s.$$

Apéndice B

Regularidad de las medidas

Al trabajar con probabilidades en $\mathbb{R}^d$ con la σ-álgebra generada por los abiertos, denotada por β^d, se cuenta con unas propiedades de regularidad, útiles para muchos de los razonamientos de este trabajo. En este apéndice, se explica qué significa que una medida sea regular y bajo qué condiciones lo es. Los resultados que se exponen a continuación son una adaptación de [11].

Definición B.0.1. *Sea $\mathcal{X}$ un espacio topológico Hausdorff.*

- *La σ-álgebra de Borel es la generada por los abiertos de $\mathcal{X}$, se denota por β.*
- *Se denomina medida de Borel a toda medida definida sobre la σ-álgebra de Borel de $\mathcal{X}$.*
- *Si μ es una medida de Borel positiva, un conjunto de Borel $E \subset \mathcal{X}$ es regular exteriormente si*

$$\mu(E) = \text{ínf}\{\mu(V) : E \subseteq V, V \text{ abierto}\}$$

  y es regular interiormente si

$$\mu(E) = \sup\{\mu(K) : K \subseteq E, K \text{ compacto}\}.$$

- *Se dice que μ es una medida ajustada si $\mathcal{X}$ es regular interiormente.*

Si todo conjunto de Borel de $\mathcal{X}$ es a la vez regular exterior e interiormente, se dice que la medida μ es regular.

Lema B.0.2. *Sea $\mathcal{X}$ un espacio topológico Hausdorff, $\mathcal{S}$ una σ-álgebra en $\mathcal{X}$ y μ una medida finita y ajustada sobre $(\mathcal{X}, \mathcal{S})$. Se define:*

$$\mathcal{R} := \{A \in \mathcal{S} : A \text{ y } \mathcal{X} \setminus A \text{ son regulares interiormente para } \mu\}.$$

Entonces, $\mathcal{R}$ es una σ-álgebra. También es σ-álgebra la clase de conjuntos:

$$\mathcal{S} := \{A \in \mathcal{S} : A \text{ y } \mathcal{X} \setminus A \text{ son regulares exteriormente para } \mu\}.$$

Demostración. Por definición, $\mathcal{R}$ es cerrado para complementarios. Sea $\{A_n\}_{n=1}^{\infty} \subset \mathcal{R}$, y sea $A := \bigcup_{n\geq 1} A_n$. Dado $\varepsilon > 0$, para cada $n \in \mathbb{N}$, existen dos conjuntos compactos $K_n \subseteq A_n$ y $L_n \subseteq A_n^C$ tales que

$$\mu(A_n \setminus K_n) < \frac{\varepsilon}{3^n} \quad \text{y} \quad \mu(A_n^C \setminus L_n) < \frac{\varepsilon}{2^n}.$$

Por ser μ una medida finita, existe un $n_0 \in \mathbb{N}$ tal que

$$\mu(A) - \mu\Big(\bigcup_{n=1}^{n_0} A_n\Big) < \frac{\varepsilon}{2}.$$

Sea $K := \bigcup\limits_{n=1}^{n_0} K_n$. Entonces K es compacto, $K \subseteq A$, y

$$\mu\Big(\bigcup_{n=1}^{n_0} A_n\Big) - \mu(K) \leq \sum_{n=1}^{n_0} [\mu(A_n) - \mu(K_n)] \leq \frac{\varepsilon}{2}.$$

Por lo tanto,

$$\mu(A) - \mu(K) = \Big[\mu(A) - \mu\Big(\bigcup_{n=1}^{n_0} A_n\Big)\Big] - \Big[\mu\Big(\bigcup_{n=1}^{n_0} A_n\Big) - \mu(K)\Big] < \varepsilon.$$

o, equivalentemente, $\mu(K) \geq \mu(A) - \varepsilon$. Se tiene probado que A es regular exterior. Para probar que A^C también es regular exterior, se considera el compacto $L := \bigcap\limits_{n=1}^{\infty} L_n$. Entonces,

$$\mu(A^c \setminus L) = \mu\Big(\bigcup_{n=1}^{n_0} A_n^C\Big) - \ \mu(L) \leq \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon.$$

Además, $\mathcal{X} \in \mathcal{R}$ por ser μ una medida ajustada. Por lo tanto, $\mathcal{R}$ es una σ-álgebra. Se prueba, análogamente, que $\mathcal{S}$ es σ-álgebra. □

Teorema B.0.3. *En un espacio métrico $(\mathcal{X}, d)$, toda medida de Borel finita y ajustada μ es regular.*

Demostración. Sea U un abierto de $\mathcal{X}$, entonces, es regular exteriormente. Para probar que U es regular interiormente, se considera la sucesión creciente de conjuntos cerrados $\{C_n\}_{n=1}^{\infty}$ definidos de la siguiente forma:

$$C_n := \{x : d(x, U^C) \geq \frac{1}{n}\}, \quad n \in \mathbb{N}.$$

Como U^C es cerrado, $d(x, U^C) = 0$ si y solo si $x \notin U$. Entonces, para todo $n \in \mathbb{N}$, $C_n \subset U$ y $U = \bigcup\limits_{n=1}^{\infty} C_n$. Se tiene que:

$$\mu(U) = \mu\Big(\bigcup_{n=1}^{\infty} C_n\Big) = \lim_{n\to\infty} \mu(C_n).$$

De aquí, se concluye que

$$\mu(U) = \sup\{\mu(C) : C \subseteq U,\, C \text{ cerrado}\}.$$

Si μ es ajustada, entonces dado $\varepsilon > 0$, existe un conjunto compacto K tal que $\mu(\mathcal{X} \setminus K) < \frac{\varepsilon}{2}$. Se acaba de probar que existe $C \subset U$ cerrado en $\mathcal{X}$ tal que $\mu(U \setminus C) < \frac{\varepsilon}{2}$. Sea $L := K \cap C$. Entonces L es compacto, $L \subset U$ y

$$\mu(U \setminus L) \leq \mu(U \setminus C) + \mu(\mathcal{X} \setminus K) < \varepsilon.$$

Por lo tanto, U es regular interiormente.

Ahora, sea F cerrado en $\mathbb{R}^d$, entonces es regular interiormente (porque μ es ajustada). Para probar que F es regular exteriormente, se considera la sucesión decreciente de abiertos $\{V_n\}_{n=1}^{\infty}$ definidos como:

$$V_n = \big\{x \in \mathbb{R}^d : d(x, F) < \frac{1}{n}\big\}.$$

Entonces, para todo $n \in \mathbb{N}$, $F \subset V_n$ y $\bigcap_{n=1}^{\infty} V_n = F$. Se tiene que:

$$\mu(F) = \mu\Big(\bigcap_{n=1}^{\infty} V_n\Big) = \lim_{n\to\infty} \mu(V_n).$$

De aquí, se concluye que

$$\mu(F) = \text{ínf}\{\mu(V) : F \subseteq V,\, V \text{ abierto}\}.$$

Por lo tanto, si U es abierto de $\mathcal{X}$, entonces $U \in \mathcal{C}$ y $U \in \mathcal{S}$. Entonces, en $\mathcal{X}$ todos los conjuntos son regulares para cualquier medida finita.

□

Teorema B.0.4. *Si $(\mathcal{X}, d)$ es un espacio métrico, completo y separable (espacio Polaco), toda medida de Borel finita es regular.*

Demostración. Sea μ una medida finita en $\mathcal{X}$. Por el teorema B.0.3, para ver que es regular, basta probar que es ajustada. Existe un conjunto numerable $D = \{x_n : n \geq 1\}$ denso en $\mathcal{X}$. Fijado $\varepsilon > 0$, para cada $m \in \mathbb{N}$, existe un natural $n(m)$ tal que

$$\mu\Big(\mathcal{X} \setminus \bigcup_{n=1}^{n(m)} B(x_n, \frac{1}{m})\Big) < \frac{\varepsilon}{2^m}.$$

Se denota por $A_m := \bigcup_{n=1}^{n(m)} B(x_n, \frac{1}{m})$. Sea

$$K := \bigcap_{m\geq 1} \bigcup_{n=1}^{n(m)} B(x_n, \frac{1}{m}).$$

K es acotado y cerrado en un espacio métrico completo, entonces K es compacto, y se tiene que

$$\mu(K^C) = \mu\Big(\bigcup_{m=1}^{\infty} A_m^C\Big) \leq \sum_{m=1}^{\infty} \mu(A_m^C) \leq \sum_{m=1}^{\infty} \frac{\varepsilon}{2^m} = \varepsilon.$$

□

Lema B.0.5. *Sea $\mathcal{X}$ un espacio métrico, completo y separable (espacio Polaco) y β la σ-álgebra generada por los abiertos de $\mathcal{X}$. Se supone que p y q son probabilidades en $\mathcal{X}$. Si $\int f dp = \int f dq$ para toda función f continua y acotada en $\mathcal{X}$. Entonces, $p = q$.*

Demostración. Sea U un abierto en $\mathcal{X}$. Se considera la sucesión de funciones continuas y acotadas definidas de la siguiente forma:

$$f_n(x) = \text{mín}\{1, nd(x, U^c)\}, \quad \forall n \in \mathbb{N}.$$

Como U^c es cerrado, $d(x, U^c) = 0$ si y solo si $x \notin U$. Es fácil ver que $\{f_n\}_{n=1}^{\infty}$ es una sucesión creciente que converge puntualmente al indicador de U. Por el teorema de la convergencia monótona,

$$p(U) = \int \mathcal{X}_U dp = \lim_{n\to\infty} \int f_n dp = \lim_{n\to\infty} \int f_n dq = \int \mathcal{X}_U dq = q(U).$$

Si dos probabilidades coinciden en el conjunto de abiertos de $\mathcal{X}$, entonces son iguales. Para ver esto, se usa la regularidad de las probabilidades en un espacio polaco (ver apéndice B).

$$p(E) = \text{ínf}\{p(U) : E \subseteq U,\ U \text{ abierto}\} = \text{ínf}\{q(U) : E \subseteq U,\ U \text{ abierto}\} = q(E).$$

□

B.1. Teorema de representación de Riesz (funcionales acotados)

En la demostración de la dualidad de Kantorovich (teorema 3.2.1), se trabaja con el dual del espacio de las funciones continuas que se anulan en el infinito. Este espacio vectorial se identifica con las medidas de Borel regulares. El teorema de Riesz es el resultado que formaliza la idea anterior; se estudió en la asignatura Análisis Real del grado y se puede encontrar en [4].

Definición B.1.1. *Se dice que una función real f definida en $\mathcal{X}$ se anula en el infinito si, para cada $\varepsilon > 0$, existe un conjunto compacto $K \subset \mathcal{X}$ tal que $|f(x)| < \varepsilon$ para todo $x \in K$. Se denota por $\mathcal{C}_0(\mathcal{X})$ el espacio de las funciones continuas en $\mathcal{X}$ que se anulan en el infinito.*

Teorema B.1.2 (Teorema de representación de Riesz)**.** *Sea $\mathcal{X}$ un espacio Hausdorff localmente compacto. Todo funcional lineal y acotado φ en $\mathcal{C}_0(\mathcal{X})$ se representa por una única medida de Borel regular μ, en el sentido de que*

$$\varphi(f) = \int_X f\, d\mu, \quad \textit{para toda } f \in \mathcal{C}_0(X).$$

Apéndice C

Convergencia débil de probabilidades

Se resumen, en este apartado, los resultados sobre convergencia en distribución de probabilidades que se utilizan a lo largo del trabajo, todos ellos vistos en el grado (en la asignatura de Teoría de la Probabilidad). Estos resultados se pueden encontrar en [6]

Definición C.0.1. *Sea P una probabilidad en $(\mathbb{R}^d, \beta^d)$ con función de distribución F. Sea $\{P_n\}_{n=1}^{\infty}$ una sucesión de probabilidades en $(\mathbb{R}^d, \beta^d)$. Para cada $n \in \mathbb{N}$, se denota por F_n a la función de distribución de P_n. Se dice que la sucesión $\{P_n\}_{n=1}^{\infty}$ converge en distribución hacia P si*

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \forall x \in \mathcal{C}(F).$$

Se ha denotado por $\mathcal{C}(F)$ al conjunto de puntos de continuidad de F.

El teorema de Portmanteau da varias caracterizaciones de la convergencia en distribución. Se usará en la demostración de la existencia de un plan de transporte óptimo, bajo ciertas condiciones, en el teorema 3.1.1 y en la prueba de la unicidad de los recortes óptimos de dos probabilidades distintas, en la proposición 3.3.7, ambas del capítulo 3.

Teorema C.0.2 (Portmanteau)**.** *Sean $\{P_n\}_{n=1}^{\infty}$ y P probabilidades en $(\mathbb{R}^d, \beta^d)$ con funciones de distribución $\{F_n\}_{n=1}^{\infty}$ y F. Las condiciones siguientes son equivalentes:*

1. *$\{P_n\}_{n=1}^{\infty}$ converge en distribución a P.*
2. *$\lim_{n\to\infty} \int f dP_n = \int f dP$ para toda función f continua y acotada, es decir, para toda $f \in \mathcal{C}_b(\mathbb{R}^d)$.*
3. *Para todo A abierto de $\mathbb{R}^d$, $P(A) \leq \liminf_{n\to\infty} P_n(A)$.*
4. *Para todo C cerrado de $\mathbb{R}^d$, $P(C) \geq \limsup_{n\to\infty} P_n(C)$.*
5. *Para todo $H \in \beta$ tal que $P(\partial H) = 0$, $\lim_{n\to\infty} P_n(H) = P(H)$.*

También se recoge en este apéndice el teorema de Representación de Skorohod, que se usa en 3.3.3, para probar una caracterización de la convergencia de la distancia de Wasserstein.

Teorema C.0.3 (de Representación de Skorokhod)**.** *Sea $\{R_n\}_n$ una sucesión de probabilidades en $\mathbb{R}^d$ que converge en distribución hacia una probabilidad R. Entonces, existen X y $\{X_n\}_n$ vectores aleatorios definidos en un espacio probabilístico $(\Omega, \mathcal{M}, \mathbb{P})$ tales que*

$$X_n \sim R_n, \quad X \sim R \quad y \quad X_n \underset{c.s.}{\longrightarrow} X.$$

C.1. Sucesión de probabilidades ajustada

Es claro que para cualquier probabilidad en $(\mathbb{R}^d, \beta^d)$, para cada $\varepsilon > 0$, existe un compacto $K \subset \mathbb{R}^d$ tal que $P(K^C) < \varepsilon$. La siguiente definición se usa también en la demostración de 3.1.1, en el capítulo 3.

Definición C.1.1. *Sea $\{P_n\}_{n=1}^{\infty}$ una sucesión de probabilidades en $(\mathbb{R}^d, \beta^d)$. Se dice que es ajustada si para cada $\varepsilon > 0$, existe un compacto $K \subset \mathbb{R}^d$ tal que $P_n(K^C) < \varepsilon$, para todo $n \in \mathbb{N}$.*

Se enuncian, a continuación, un resultado que se deduce del teorema de Helly.

Teorema C.1.2. *Si la sucesión de probabilidades $\{P_n\}_{n=1}^{\infty}$ en $(\mathbb{R}^d, \beta^d)$ es ajustada, entonces de cualquier subsucesión $\{P_{n_k}\}_{k=1}^{\infty}$ se puede extraer otra subsucesión $\{P_{n_{k_m}}\}_{m=1}^{\infty}$ que converge en distribución hacia una probabilidad P.*

C.2. Función cuantil

Se trabaja, en varias ocasiones, con la función cuantil, en concreto en la sección 3.3.2 (donde se da una caracterización de la distancia de Wasserstein en $\mathbb{R}$ a partir de la función cuantil). Se recuerdan la definición y algunas propiedades.

Definición C.2.1. *Sea $F : \mathbb{R} \to \mathbb{R}$ una función de distribución, es decir, una función creciente, continua por la derecha y tal que $\lim_{x \to -\infty} F(x) = 0$ y $\lim_{x \to +\infty} F(x) = 1$. Entonces se define la función cuantil $F^{-1} : (0,1) \to \mathbb{R}$ como:*

$$F^{-1}(t) = \text{ínf}\{x \in \mathbb{R} : t \leq F(x)\}, \quad t \in (0,1).$$

Para todo $t \in (0,1)$ y para todo $x \in \mathbb{R}$, se verifica que

$$F^{-1}(t) \leq x \Leftrightarrow t \leq F(x). \tag{C.1}$$

También, se cumple que

$$F(F^{-1}(t)-) \leq t \leq F(F^{-1}(t)), \quad \forall t \in (0,1). \tag{C.2}$$

donde se ha denotado por $F^{-1}(t)-$ al límite por la izquierda de F^{-1} en t.

De estas propiedades, se deduce la siguiente proposición:

Proposición C.2.2. *La función cuantil $F^{-1} : (0,1) \to \mathbb{R}$ es creciente y continua por la izquierda. Además, es una variable aleatoria en el espacio probabilístico $((0,1), \beta_{(0,1)}, \mathcal{L})$, donde $\mathcal{L}$ es la medida de Lebesgue.*

Otra formulación equivalente de esta proposición es la siguiente:

Proposición C.2.3. *Sea U una variable aleatoria con distribución $\mathcal{U}(0,1)$, entonces $F^{-1}(U)$ tiene función de distribución F.*

Demostración. Sea $x \in \mathbb{R}$, por C.1, se tiene que

$$P\big(F^{-1}(U) \leq x\big) = P\big(U \leq F(x)\big) = F(x).$$

□

Proposición C.2.4. *Sea μ una probabilidad con densidad en $(\mathbb{R}, \beta)$ y X una variable aleatoria con distribución μ. Entonces, si F es la función de distribución de μ, se cumple que $F(X) \sim \mathcal{U}(0,1)$.*

Demostración. Sea $x \in \mathbb{R}$,

$$P\big(F(X) \geq x\big) = P\big(X \geq F^{-1}(x)\big) = 1 - F(F^{-1}(x)) = 1 - x.$$

La última igualdad es consecuencia de C.2 y de que F sea continua por tener μ densidad. □

Para terminar la sección de la función cuantil, se enuncia un resultado que se deduce de la demostración del teorema de Skorokhod en $\mathbb{R}$.

Teorema C.2.5. *Sean P y $\{P_n\}_{n=1}^{\infty}$ probabilidades en $(\mathbb{R}, \beta)$ tales que $\{P_n\}_{n=1}^{\infty}$ converge en distribución a P. Si F y $\{F_n\}_{n=1}^{\infty}$ son las respectivas funciones de distribución, entonces*

$$\lim_{n\to\infty} F_n^{-1}(t) = F^{-1}(t), \quad \textit{para casi todo } t \in (0,1).$$

C.3. Aproximación de funciones de distribución por suavizado

Por último, se explica cómo aproximar funciones de distribución mediante suavizado por convolución con un núcleo, procedimiento usado también en la sección 3.3.2.

Proposición C.3.1. *Sea P una probabilidad en $(\mathbb{R}, \beta)$ con función de distribución F. Entonces, existe una sucesión de probabilidades con densidad $\{P_n\}_{n=1}^{\infty}$ en $(\mathbb{R}, \beta)$ que converge en distribución hacia P.*

Demostración. Sea $\{\varphi_n\}_{n=1}^{\infty}$ una sucesión regularizante en $\mathbb{R}$, es decir, una sucesión de funciones que verifique las siguientes propiedades:

1. Para todo $n \in \mathbb{N}$, $\varphi_n \in C^{\infty}(\mathbb{R})$.

2. Para todo $n \in \mathbb{N}$, se tiene:

$$\varphi_n(x) \geq 0, \quad \text{para todo } x \in \mathbb{R}^d \quad \text{y} \quad \int_{\mathbb{R}^d} \varphi_n(x)\, dx = 1.$$

3. $\text{sop}(\varphi_n) \subset B\left(0, \frac{1}{n}\right)$.

Como F es localmente integrable en $\mathbb{R}$ por ser acotada, tiene sentido considerar la convolución con φ_n, y esta será derivable (por serlo φ_n). Además, $\lim_{n\to\infty} F * \varphi_n(x) = F(x)$ para todo $x \in \mathcal{C}(F)$. Se prueba:

Sea $x \in \mathcal{C}(F)$, para todo $\varepsilon > 0$, existe $n_0 \in \mathbb{N}$ tal que si $|x - z| \leq \frac{1}{n_0}$, entonces $|F(x) - F(z)| \leq \varepsilon$. Por lo tanto, para todo $n \geq n_0$,

$$\begin{aligned}
|(F * \varphi_n)(x) - F(x)| &= \left| \int_{\mathbb{R}} (F(z) - F(x))\varphi_n(x - z)\, dz \right| \\
&\leq \int_{\mathbb{R}} |F(z) - F(x)|\varphi_n(x - z)\, dz = \int_{B(x,\frac{1}{n})} |F(z) - F(x)|\varphi_n(x - z)\, dz \\
&\leq \varepsilon \int_{B(x,\frac{1}{n})} \varphi_n(x - z)\, dz = \varepsilon.
\end{aligned}$$

Solo falta probar que $F_n = F * \varphi_n$ es una función de distribución.

- F_n es creciente: Como F es creciente y φ_n mayor o igual que 0, esta propiedad se deduce de la monotonía de la integral, ya que

$$F_n(x) := \int F(x - z)\varphi_n(z)dz.$$

- $\lim_{x\to-\infty} F_n(x) = 0$:

  Para todo $z \in \mathbb{R}$, $\lim_{x\to-\infty} F(x - z)\varphi_n(z) = 0$. Además, se tiene la acotación de esta función por otra integrable:

$$|F(x - z)\varphi_n(z)| \leq \varphi_n(z), \quad \forall x \in \mathbb{R}.$$

  Se concluye, por el teorema de la convergencia dominada, que

$$\lim_{x\to-\infty} F * \varphi_n(x) = \lim_{x\to-\infty} \int F(x - z)\varphi_n(z)dz = 0.$$

- $\lim_{x\to+\infty} F_n(x) = 1$:

  Para todo $z \in \mathbb{R}$, $\lim_{x\to-\infty} F(x - z)\varphi_n(z) = \varphi_n(z)$ y se vuelve a tener la misma acotación que en el punto anterior, lo que permite aplicar el teorema de la convergencia dominada de nuevo.

$$\lim_{x\to+\infty} F * \varphi_n(x) = \lim_{x\to+\infty} \int F(x - z)\varphi_n(z)dz = \int \varphi_n(z)dz = 1.$$

Para cada $n \in \mathbb{N}$, sea P_n la única probabilidad en $\mathbb{R}$ con función de distribución F_n. Se acaba de probar que la sucesión $\{P_n\}_{n=1}^{\infty}$ converge en distribución hacia P $\square$

Definición C.3.2. *Sea μ una probabilidad en $(\mathbb{R}, \beta)$. Sea φ una función definida en $\mathbb{R}$ no negativa, con soporte compacto y tal que $\int \varphi = 1$. La convolución de μ con φ_n es la medida $\mu * \varphi$ definida por:*

$$(\mu * \varphi)(A) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathcal{X}_A(x+y)\varphi_n(y)\, dy \right) d\mu(x),$$

para cualquier conjunto medible $A \subseteq \mathbb{R}$.

Proposición C.3.3. *Sea μ una probabilidad en $(\mathbb{R}, \beta)$ con función de distribución F. Sea φ una función definida en $\mathbb{R}$ no negativa, con soporte compacto tal que $\int \varphi = 1$. Entonces, la función de distribución de la medida $\mu * \varphi$ es $F * \varphi$.*

Demostración. Si $t \in \mathbb{R}$,

$$\mu * \varphi((-\infty, t]) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathcal{X}_{(-\infty,t]}(x+y)\varphi(y)\, dy \right) d\mu(x).$$

Por Fubini (ambas funciones están acotadas), se puede cambiar el orden de integración.

$$\mu * \varphi((-\infty, t]) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathcal{X}_{(-\infty,t]}(x+y) d\mu(x) \right) \varphi(y)\, dy.$$

Desarrollando la primera integral,

$$\int_{\mathbb{R}} \mathcal{X}_{(-\infty,t]}(x+y) d\mu(x) = \mu\left(\{x : x+y \leq t\}\right) = \mu\left((-\infty, t-y]\right) = F(t-y).$$

Por lo tanto,

$$\mu * \varphi((-\infty, t]) = \int_{\mathbb{R}} F(t-y)\varphi(y)\, dy = F * \varphi(t).$$

$\square$

Si P es una probabilidad en $\mathbb{R}$, el último resultado de este apéndice garantiza que, escogiendo de forma adecuada la sucesión regularizante, se puede construir una sucesión de probabilidades con densidad que converjan débilmente hacia P y cuyos momentos de orden 2 también converjan.

Proposición C.3.4. *Sea P una probabilidad en $(\mathbb{R}, \beta)$ con momento de orden 2 finito. Sea $\rho \in C_c^{\infty}(\mathbb{R})$ una función no negativa, simétrica, con soporte compacto y tal que $\int \rho = 1$. Para cada $n \in \mathbb{N}$, se define*

$$\rho_n(x) := n\rho(nx), \quad \forall x \in \mathbb{R}.$$

Entonces la sucesión de medidas regularizadas por convolución $P_n := P * \rho_n$ verifica

$$\lim_{n\to\infty} \int_{\mathbb{R}} x^2\, dP_n(x) = \int_{\mathbb{R}} x^2\, dP(x).$$

Demostración. Por la definición de convolución, se tiene que:

$$\int x^2 \, dP_n(x) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} (x+y)^2 \rho_n(y) \, dy \right) dP(x).$$

Entonces,

$$\begin{aligned} &\int x^2 \, dP_n(x) \\ &= \int x^2 \, dP(x) + 2 \int x \cdot \left(\int y \rho_n(y) \, dy \right) dP(x) + \int \left(\int y^2 \rho_n(y) \, dy \right) dP(x). \end{aligned}$$

Como ρ es simétrica, se tiene $\int y\rho_n(y)dy = 0$ para todo $n \in \mathbb{N}$. Además, por un cambio de variable,

$$\int y^2 \rho_n(y) \, dy = \frac{1}{n^2} \int z^2 \rho(z) \, dz = \frac{C}{n^2},$$

donde $C := \int z^2 \rho(z) dz$. Por lo tanto,

$$\int x^2 \, dP_n(x) = \int x^2 \, dP(x) + \frac{C}{n^2}.$$

Por último, cuando n tiende a ∞,

$$\lim_{n \to \infty} \int x^2 \, dP_n(x) = \int x^2 \, dP(x).$$

□

Apéndice D

Convexidad

D.1. Teorema de Rademacher

En el problema del transporte óptimo, estudiado en el capítulo 3, se trabaja con funciones convexas. Por eso, se necesitan los siguientes resultados, sus demostraciones se pueden encontrar en [12], concretamente en la sección 3.1.

Definición D.1.1. *Una función $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ es inferiormente semicontinua si, fijado $x_0 \in \mathbb{R}^d$, para toda sucesión $\{x_n\}_{n=1}^{\infty}$ tal que $\lim_{n\to\infty} x_n = x_0$ se cumple que*

$$\lim_{n\to\infty} \inf f(x_n) \geq f(x_0).$$

Definición D.1.2. *$f : U \subset \mathbb{R}^d \to \mathbb{R}^m$ es Lipschitz si existe una constante $L > 0$ tal que*

$$\|f(x) - f(y)\| \leq L\|x - y\|, \quad \text{para todo } x, y \in U.$$

Proposición D.1.3. *Si la función $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ es convexa, entonces es inferiormente semicontinua en los puntos del interior de su dominio. Es decir, si $x \in Int(Dom\varphi)$, entonces, para toda sucesión $\{x_n\}_{n=1}^{\infty}$ que converge a x, se tiene que*

$$\liminf_{n\to\infty} \varphi(x_n) \geq \varphi(x).$$

Además, φ es localmente Lipschitz en los puntos del interior de su dominio, es decir, para cada punto $x \in Int(Dom\varphi)$, existe un entorno suyo $U \subset \mathbb{R}^d$ tal que φ es Lipschitz en U.

Teorema D.1.4 (Teorema de Rademacher)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}^m$ una función localmente Lipschitz. Entonces, f es diferenciable para casi todo punto.*

Corolario D.1.5. *Una función $\varphi : \mathbb{R}^n \to \mathbb{R}$ convexa es diferenciable salvo en un conjunto de medida nula.*

D.2. Segunda forma geométrica del teorema de Hahn-Banach

En un espacio vectorial normado, para separar convexos por hiperplanos, se razona a partir del teorema de Hahn-Banach, visto en la asignatura de Introducción a los Espacios de Funciones del grado. Este resultado se puede encontrar en [7].

Teorema D.2.1. *Sean A, B convexos no vacíos disjuntos del espacio vectorial normado E, tales que A es cerrado y B es compacto. Existe $\phi \in E'$, y $a, b \in \mathbb{R}$ tales que si $x \in A$ e $y \in B$, se tiene que*

$$\phi(x) \leq a < b \leq \phi(y).$$

D.3. Dualidad de Fencher-Rockafellar

Para demostrar la dualidad de Kantorovich (teorema 3.2.1), imprescindible para el estudio del problema de transporte óptimo asociado al coste cuadrático, se usa un resultado de dualidad de funcionales convexos. Se demuestra a partir de la segunda forma geométrica del teorema de Hahn Banach; se puede encontrar en [7], en concreto, el teorema I.11.

Definición D.3.1. *Sea E un espacio vectorial normado y E' su dual topológico. Sea Θ una función convexa en E con valores en $\mathbb{R} \cup \{+\infty\}$. Se define su transformada de Legendre-Fenchel $\Theta^* : E' \to \mathbb{R} \cup \{+\infty\}$ como:*

$$\Theta^*(T) = \sup_{u \in E} \{Tu - \Theta(u)\}, \quad \forall T \in E'.$$

Teorema D.3.2 (Dualidad de Fencher-Rockafellar)**.** *Sea E un espacio vectorial normado y E' su dual topológico. Sean Θ y Ξ dos funciones convexas en E con valores en $\mathbb{R} \cup \{+\infty\}$. Se supone que existe $v_0 \in E$ tal que Θ es continuo en v_0 y, además,*

$$\Theta(v_0) < +\infty, \quad \Xi(v_0) < +\infty.$$

Entonces,

$$\inf_{u \in E} \{\Theta(u) + \Xi(u)\} = \sup_{\pi \in E'} \{-\Theta^*(-T) - \Xi(T)\}.$$

Apéndice E

Convergencia débil de funciones

En el capítulo 3, concretamente en la demostración del teorema 3.2.13, se necesita trabajar con sucesiones de funciones en $L^1(\mu)$ que convergen débilmente, es decir, en el dual. En este apéndice se da la definición y algunas propiedades, que se pueden encontrar en el capítulo 6 de [13].

Definición E.0.1. *Sea μ una probabilidad en $(\mathbb{R}^d, \beta^d)$. Se dice que una sucesión $\{f_n\}_{n=1}^{\infty} \subset L^1(\mu)$ converge débilmente a $f \in L^1(\mu)$ si converge en el dual de $L^1(\mu)$. En otras palabras, si para toda función $\phi \in L^\infty(\mathbb{R}^d)$ se cumple que*

$$\lim_{n\to\infty} \int f_n(x)\,\phi(x)\,d\mu(x) = \int f(x)\,\phi(x)\,d\mu(x).$$

Esta convergencia es más débil que la convergencia en norma. En la demostración de algunos teoremas del trabajo, se usa que la convergencia débil preserva el orden; esta propiedad se prueba a partir del siguiente lema.

Lema E.0.2. *Sea μ una probabilidad en $(\mathbb{R}^d, \beta^d)$. Sea $f \in L^1(\mu)$. Si*

$$\int f(x)\,\phi(x)\,d\mu(x) \geq 0, \quad \forall \phi \in L^\infty(\mathbb{R}^d),\ \phi \geq 0,$$

entonces $f(x) \geq 0$, $\mu-$casi seguro.

Teorema E.0.3. *Sea μ una probabilidad en $(\mathbb{R}^d, \beta^d)$. Sea $\{f_n\}_{n=1}^{\infty} \subset L^1(\mu)$ una sucesión que converge débilmente a $f \in L^1(\mu)$. Sea $g \in L^1(\mathbb{R}^d)$ tal que, para todo $n \in \mathbb{N}$, $f_n(x) \geq g(x)$ para casi todo $x \in \mathbb{R}^d$, con respecto a la probabilidad μ. Entonces $f(x) \geq g(x)$ para casi todo $x \in \mathbb{R}^d$.*

E.1. Integrabilidad uniforme

Si se tiene una sucesión de funciones en $L^1(\mu)$, para garantizar que existe alguna subsucesión convergente débilmente, no es suficiente con que la sucesión esté acotada con la norma de $L^1(\mu)$, se necesitan más condiciones. Por esta razón, se introduce el concepto de integrabilidad uniforme.

Definición E.1.1. *Sea $(\mathcal{X}, \mathcal{M}, \mu)$ un espacio medible. La familia de funciones $\mathcal{F} \subset L^1(\mu)$ es uniformemente integrable si:*

- $\sup_{f\in\mathcal{F}} \int_{\mathbb{R}^d} |f(x)|\, d\mu(x) < \infty$.
- *Para todo $\varepsilon > 0$, existe $\delta > 0$ tal que, para todo conjunto $E \subset \beta^d$ con $\mu(E) < \delta$, se cumple:*

$$\sup_{f\in\mathcal{F}} \int_E |f(x)|\, d\mu(x) < \varepsilon.$$

Teorema E.1.2 (Dunford–Pettis)**.** *Sea μ una medida en $\mathbb{R}^d$ y $\{f_n\}_{n=1}^{\infty} \subset L^1(\mu)$ una sucesión de funciones uniformemente integrable. Entonces, existe una subsucesión $\{f_{n_k}\}_{k=1}^{\infty}$ y una función $f \in L^1(\mu)$ tal que $\{f_{n_k}\}_{k=1}^{\infty}$ converge débilmente a f.*

La prueba de este teorema se puede ver en [8], teorema 4.30. Por último, se enuncia un resultado que se deduce del teorema de Vitali. Establece condiciones para garantizar cuándo hay convergencia en norma p, con $p \geq 1$, y se usa para probar una caracterización de la convergencia con la distancia de Wasserstein, en 3.3.3. Se puede encontrar en [19], teorema 5.5.

Teorema E.1.3 (Vitali)**.** *Sea $\{X_n\}_n$ una sucesión de vectores aleatorios en $\mathbb{R}^d$ que converge en probabilidad hacia X. Entonces, para $p \geq 1$,*

$$\lim_{n\to\infty} E\|X_n\|^p = E\|X\|^p < \infty$$

si, y solo si,

$$\lim_{n\to\infty} \|X_n - X\|_p = 0.$$

Apéndice F

Desintegración de medidas

Para demostrar el lema de pegado, en 3.3.1, se necesita recurrir a la desintegración de medidas. Los resultados que se enuncian a continuación se pueden encontrar en [11].

Proposición F.0.1. *Sean $(\mathcal{X}_1, \mathcal{M}_1)$ y $(\mathcal{X}_2, \mathcal{M}_2)$ dos espacios medibles. Sea P_1 una probabilidad en $(\mathcal{X}_1, \mathcal{M}_1)$. Si $K : \mathcal{X}_1 \times \mathcal{M}_2 \to \mathbb{R}$ es una función tal que*

- *Para cada $x \in \mathcal{X}_1$, $K(x, \cdot)$ es una probabilidad en $(\mathcal{X}_2, \mathcal{M}_2)$.*
- *$x \mapsto K(x, B)$ es medible $\forall B \in \mathcal{M}_2$.*

Entonces,

$$\pi(E) = \int_{\mathcal{X}_1} K(x, E_x)\, dP_1(x), \quad \forall E \in \mathcal{M}_1 \otimes \mathcal{M}_2.$$

define una probabilidad en $(\mathcal{X}_1 \times \mathcal{X}_2, \mathcal{M}_1 \otimes \mathcal{M}_2)$.

Teorema F.0.2. *Sean $(\mathcal{X}_1, \mathcal{M}_1)$ y $(\mathcal{X}_2, \mathcal{M}_2)$ dos espacios medibles. Además, se supone que $\mathcal{X}_1$ y $\mathcal{X}_2$ son espacios métricos, completos y separables. Sea π una probabilidad en $(\mathcal{X}_1 \times \mathcal{X}_2, \mathcal{M}_1 \otimes \mathcal{M}_2)$ cuya marginal sobre $\mathcal{X}_1$ es P_1. Entonces existe una función $K : \mathcal{X}_1 \times \mathcal{M}_2 \to \mathbb{R}$ de forma que*

- *Para cada $x \in \mathcal{X}_1$, $K(x, \cdot)$ es una probabilidad en $(\mathcal{X}_2, \mathcal{M}_2)$.*
- *$x \mapsto K(x, B)$ es medible $\forall B \in \mathcal{M}_2$.*

que verifica

$$\pi(E) = \int_{\mathcal{X}_1} K(x, E_x)\, dP_1(x), \quad \forall E \in \mathcal{M}_1 \otimes \mathcal{M}_2.$$

Bibliografía

[1] P. C. Álvarez-Esteban et al. "Uniqueness and approximate computation of optimal incomplete transportation plans". En: *Ann. Inst. Henri Poincaré Probab. Stat.* 2 (2011), págs. 358-375.

[2] Pedro C. Álvarez-Esteban et al. "Similarity of samples and trimming". En: *Bernoulli* 18.2 (2012), págs. 606-634.

[3] Pedro César Álvarez-Esteban et al. "Trimmed comparison of distributions". En: *J. Amer. Statist. Assoc.* 103.482 (2008), págs. 697-704.

[4] Robert B. Ash. *Real analysis and probability*. Academic Press, 1972.

[5] Peter J. Bickel y David A. Freedman. "Some Asymptotic Theory for the Bootstrap". En: *The Annals of Statistics* 9.6 (1981), págs. 1196-1217.

[6] P. Billingsley. *Convergence of probability measures*. John Wiley, Sons Inc., 1999.

[7] H. Brézis. *Analyse fonctionnelle: théorie et applications*. Masson, 1983.

[8] H. Brézis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2011.

[9] Nello Cristianini y John. Shawe-Taylor. *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, 2004.

[10] I. Csiszar. "I-Divergence Geometry of Probability Distributions and Minimization Problems". En: *The Annals of Probability* 3 (1975), págs. 146-158.

[11] R. M. Dudley. *Real analysis and probability*. Second edition. Cambridge University Press, 2002.

[12] Lawrence C. Evans y Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. Revised. Chapman y Hall/CRC, 2015.

[13] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. 2nd. Wiley, 1999.

[14] Arthur Gretton et al. "A Kernel Two-Sample Test". En: *Journal of Machine Learning Research* 13.25 (2012), págs. 723-773.

[15] Gabriel Peyré y Marco Cuturi. "Computational Optimal Transport: With Applications to Data Science". En: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), págs. 355-607.

[16] Walter Rudin y Lorenzo Abellanas Rapun. *Análisis real y complejo*. 3ª ed. MacGraw-Hill, 1987.

[17] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980.

[18] John. Shawe-Taylor y Nello. Cristianini. *Kernel methods for pattern analysis*. 1st ed. Cambridge University Press, 2004.

[19] Galen R. Shorack. *Probability for Statisticians*. 2nd. Springer, 2000.

[20] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008.

[21] C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

[22] Cédric Villani. *Topics in optimal transportation*. American Mathematical Society, Providence, RI, 2003.