Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE MÁSTER

MÁSTER EN MATEMÁTICAS

LA LEY DE CONVERGENCIA DE TIPOS Y EL PROBLEMA DEL TRANSPORTE ÓPTIMO

AUTORA:
Lucía Trapote Reglero

TUTOR:
Carlos Matrán Bea

Julio 2024

A Carlos Matrán, por aceptar tutorizarme este trabajo y ofrecerme la oportunidad de trabajar con él. Por su incansable curiosidad y por saber transmitir su pasión por las matemáticas. Por la dedicación, la paciencia y el esfuerzo para que el trabajo haya salido lo mejor posible en las mejores condiciones, sin él esta memoria no hubiera sido posible.

A mi familia, por regalarme una educación y enseñarme el valor del esfuerzo. Por apoyar mis decisiones y confiar en mí incondicionalmente.

A mis compañeros del máster por el tiempo compartido y por hacer más llevadero el curso. A mis amigos, por seguir eligiendo compartir nuevas etapas conmigo, por estar siempre ahí y dar luz a mi camino.

Índice general

Resumen

La Ley de Convergencia de Tipos es un resultado clave en la Teoría de la Probabilidad. Determina que las distribuciones límite que podemos obtener al modificar una sucesión de vectores aleatorios, a través de cambios en localizaciones y parámetros de forma, son necesariamente del mismo tipo. Como ejemplo notable, garantiza que en el Teorema Central del Límite solo puede obtenerse una distribución Gaussiana, y que la velocidad $\sqrt{n}$ es esencialmente la única posible. Por otra parte, el Problema del Transporte Óptimo garantiza la existencia de funciones óptimas para minimizar el coste cuadrático del transporte de una distribución de probabilidades a otra. El trabajo consistirá en el estudio de ambos problemas y en el análisis de las conexiones y perspectivas que la denominada "descomposición polar" de Brenier, en el segundo, podría aportar al primero.

Palabras clave

Ley de Convergencia de Tipos, Problema del Transporte Óptimo, distancia de Wasserstein, misma estructura de dependencia, descomposición polar de Brenier...

Abstract

The Law of Convergence of Types is a key result in Probability Theory. It determines that the limit distributions that we can obtain by modifying a sequence of random vectors, through changes in locations and shape parameters, are necessarily of the same type. As a notable example, it guarantees that in the Central Limit Theorem only a Gaussian distribution can be obtained, and that the velocity $\sqrt{n}$ is essentially the only one possible. On the other hand, the Optimal Transport Problem guarantees the existence of optimal functions to minimize the quadratic cost of transport from one probability distribution to another. The work will consist in the study of both problems and in the analysis of the connections and perspectives that the so-called "polar decomposition" of Brenier, in the second one, could bring to the first one.

Keywords

Law of Convergence of Types, Optimal Transport Problem, Wasserstein distance, same dependence structure, Brenier polar decomposition...

Introducción

Recordamos de la asignatura de Teoría de la Probabilidad y Estadística Matemática que la Ley de Convergencia de Tipos versa sobre las posibles distribuciones que pueden esperarse al realizar transformaciones afines a una sucesión de variables aleatorias que, inicialmente, converge en distribución a una variable aleatoria no degenerada. Uno de los objetivos principales de este trabajo es presentar una generalización de este resultado a espacios de dimensión mayor que uno.

En el primer capítulo del trabajo, tras realizar un breve recordatorio de la Ley de Convergencia de Tipos, se presentan, siguiendo el trabajo de Billingsley [3] como referencia principal, resultados que recogen condiciones necesarias y suficientes sobre las aplicaciones afines y los vectores límite de las hipótesis para asegurar resultados en la línea de la Ley de Convergencia de Tipos para el caso multivariado.

Las demostraciones que presenta Billingsley en su artículo están basadas, principalmente, en teoría de grupos y en argumentos de compacidad y convergencia. Por el contrario, aunque existen distintas formas de demostrar la Ley de Convergencia de Tipos en $\mathbb{R}$, en el presente trabajo se aborda vía el Teorema de Representación de Skorohod que permite pasar de convergencias en ley entre variables aleatorias a convergencias casi seguro entre sus funciones cuantiles asociadas. De esta manera, se utiliza la función cuantil como representante de la probabilidad y se aprovechan las propiedades de las que goza la misma contando, además, con las ventajas que supone la convergencia casi seguro. Si se quiere enmarcar el caso multivariado en este mismo esquema, el papel que las funciones cuantiles cumplen en el caso unidimensional en este caso lo desempeñan los denominados "transportes óptimos". Con el objetivo de realizar las demostraciones del caso multivarado siguiendo este

esquema se requiere desarrollar un estudio de la teoría del transporte óptimo, que se recoge en el segundo capítulo del trabajo.

Introducida la teoría del transporte óptimo, y observada su estrecha relación con la teoría de la factorización polar, se puede tratar de presentar nuevas pruebas de los resultados clave expuestos en el primer capítulo utilizando el Teorema de Representación de Skorohod y recurriendo además a los recursos que nos proporciona el transporte óptimo. Esto es lo que se recoge en el tercer capítulo del trabajo que se ha dividido en dos partes.

En la primera parte se trata el caso de las aplicaciones lineales, en esencia, lo que se realiza es revisitar los resultados relativos a la generalización de la Ley de Convergencia de Tipos para el caso multivariado y presentar pruebas aprovechando la convergencia casi seguro que se obtiene al utilizar la representación de Skorohod y las ventajas que supone hacer uso de la descomposición polar de matrices al poder trabajar con matrices ortogonales (que forman un grupo compacto y, por lo tanto gozan de buenas propiedades de convergencia) y matrices simétricas y definidas positivas (que representan aplicaciones de transporte óptimo). Por otro lado, en la segunda parte se trata el caso de las estructuras de dependencia, la idea principal es adaptar los resultados relativos a la Ley de Convergencia de Tipos para el nuevo contexto en el que se trabaja y ver qué conclusiones se pueden obtener, nuevamente teniendo como herramientas la representación de Skorohod y propiedades bien conocidas de las aplicaciones de transporte óptimo.

Capítulo 1

Ley de convergencia de tipos en espacios de dimensión k

La estadística matemática y la teoría de la probabilidad tratan con frecuencia el problema de determinar la distribución límite de una sucesión de variables aleatorias. Un resultado clásico en este ámbito es el Teorema Central del Límite, el cual trabajando con una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ independientes e igualmente distribuidas con media μ y varianza σ^2 asegura que

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

La Ley de los Grandes Números garantiza la convergencia casi seguro de la sucesión de medias muestrales $\{\overline{X}_n\}_{n=1}^{\infty}$ hacia la media μ, por su parte, lo que afirma esencialmente el Teorema Central del Límite es que para un índice n suficientemente avanzado, la diferencia entre $\overline{X}_n$ y μ, cuando se multiplica por el factor $\sqrt{n}$, sigue aproximadamente una distribución normal de media 0 y varianza σ^2. Esto es, para n suficientemente grande la distribución de $\sqrt{n}(\overline{X}_n - \mu)$ está tan próxima como queramos a una normal de media 0 y varianza σ^2. En otras palabras, la distribución normal sustituye a las cotas determinísticas asociadas a los errores en las aproximaciones matemáticas, introduciendo una medida sobre el comportamiento de esos errores.

La Ley de Convergencia de Tipos, en el contexto de una sucesión de variables aleatorias $\{Y_n\}_{n=1}^{\infty}$ convergente hacia una variable aleatoria no degenerada Y, versa sobre las posibles distribuciones que pueden esperarse al realizar

transformaciones lineales a la sucesión anterior, esto es considerar una nueva sucesión $\{a_nY_n + b_n\}_{n=1}^{\infty}$, con la única condición de que los coeficientes a_n sean estrictamente positivos. La respuesta que da este resultado resulta muy interesante, pues asegura que la convergencia de la sucesión transformada es o bien una constante (es decir, una variable aleatoria degenerada), o bien una variable aleatoria no degenerada del "mismo tipo" que Y.

Una consecuencia inmediata de Ley de Convergencia de Tipos, teniendo en cuenta el Teorema Central del Límite, es el hecho de que si existen sucesiones $\{a_n\}_{n=1}^{\infty}$ y $\{b_n\}_{n=1}^{\infty}$ tales que $a_n\overline{X}_n + b_n \xrightarrow{d} Z$ entonces, necesariamente Z es una variable normal (es decir, del mismo "tipo" que una distribución normal de media 0 y varianza σ^2) o es degenerada (esto es, constante casi seguro).

Así, el principal objetivo de este primer capítulo del trabajo es recordar el clásico resultado de la ley de convergencia de tipos e introducir una generalización del mismo a espacios de dimensión k, siendo k un número natural mayor que uno.

1.1. Ley de convergencia de tipos

Antes de enunciar y demostrar la Ley de Convergencia de Tipos vamos a recordar algunos conceptos básicos e introducir la notación que utilizaremos a lo largo de todo el trabajo.

Dada una variable aleatoria X y una probabilidad P, para indicar que X sigue la ley de probabilidad P, utilizaremos indistintamente la notación

$$X \overset{d}{\sim} P \qquad \text{ó} \qquad \mathscr{L}(X) = P.$$

Recordamos además que, dadas dos variables aleatorias X e Y, se dice que están **igualmente distribuidas** si tienen la misma ley de probabilidad, esto es, si $\mathscr{L}(X) = \mathscr{L}(Y)$. Este fenómeno se suele representar como $X \overset{d}{=} Y$, notación que utilizaremos en el presente trabajo.

Con vistas a enunciar el resultado principal de este apartado, dadas dos **probabilidades** P y Q sobre la recta real $\mathbb{R}$, se dice que son **del mismo tipo** si existen variables aleatorias X e Y y constantes $a > 0$ y b tales que $\mathscr{L}(X) = P$, $\mathscr{L}(Y) = Q$ y $\mathscr{L}(Y) = \mathscr{L}(aX + b)$. En este contexto, se tiene que $Y \overset{d}{=} aX + b$ y se dice que X e Y son del **mismo tipo**.

Existen otras caracterizaciones útiles para este fenómeno, una de ellas en términos de la función cuantil, que es la que emplearemos para realizar la prueba del resultado de interés y la desarrollamos a continuación, tras realizar un breve recordatorio sobre la función cuantil.

Definición 1.1. Dada una función de distribución $F : \mathbb{R} \to [0,1]$ se define su función cuantil asociada, que denotamos por F^{-1}, como

$$F^{-1}(t) = \text{mín}\{x \in \mathbb{R} \,:\, t \leq F(x)\}, \quad \text{para todo } t \in (0,1)\,.$$

La propiedad fundamental característica de la función cuantil es la siguiente:

$$F^{-1}(t) \leq x \iff t \leq F(x) \quad \text{para todo } t \in (0,1)\,,$$

que presenta como consecuencia inmediata el hecho de que $t \leq F(F^{-1}(t))$, una relación que podemos completar comparando a través de la función $F(x-) := \lim_{y \to x-} F(y)$, escribiendo

$$F(F^{-1}(t)-) \leq t \leq F(F^{-1}(t)).$$

La función cuantil es una función $F^{-1} : (0,1) \to \mathbb{R}$ que es creciente y continua por la izquierda, pero el recíproco también se verifica, como se recoge en el siguiente resultado.

Proposición 1.2. *Toda función definida en el intervalo* $(0,1)$*, creciente y continua por la izquierda es la función cuantil de una, y solo una, función de distribución.*

La función cuantil juega además un papel fundamental en el estudio de la convergencia en ley, a través de la equivalencia existente entre la convergencia débil de funciones de distribución y la convergencia en casi todo punto de las correspondientes inversas, dadas por las funciones cuantil. Este resultado se conoce como teorema elemental de representación de Skorohod y lo presentamos a continuación:

Teorema 1.3. *Sean* F *y* $\{F_n\}_{n=1}^{\infty}$ *funciones de distribución en la recta real y, como de costumbre,* F^{-1}*,* $\{F_n^{-1}\}_{n=1}^{\infty}$ *sus correspondientes funciones cuantil. Entonces, las relaciones siguientes son equivalentes:*

(*a*) $F_n(x) \to F(x)$ *para todo* $x \in C(F)$*, donde* $C(F)$ *representa el conjunto de puntos de continuidad de* F*.*

(*b*) $F_n^{-1}(t) \to F^{-1}(t)$ *para casi todo* $t \in (0,1)$.

El calificativo elemental se utiliza para destacar que el resultado de Skorohod es válido para convergencias en espacios muy generales. En cualquier caso estos teoremas son actualmente una de las herramientas más poderosas para estudiar la convergencia de probabilidades.

Volviendo a la caracterización de las probabilidades del mismo tipo, sean X e Y variables aleatorias, P y Q probabilidades y $a > 0$ y b constantes. Supongamos que tenemos $aX + b \stackrel{d}{=} Y$ con $\mathscr{L}(X) = P$, $\mathscr{L}(Y) = Q$ y funciones de distribución asociadas F y G respectivamente. Ahora bien, dado que estamos trabajando con $a > 0$ tenemos que $aF^{-1} + b$ es una función creciente, continua por la izquierda y definida en $(0,1)$ y, por lo tanto, se trata de una función cuantil. Es conocido que $F^{-1} \stackrel{d}{=} X$ y $G^{-1} \stackrel{d}{=} Y$, de donde se sigue que:

$$aF^{-1} + b \stackrel{d}{=} aX + b \stackrel{d}{=} Y \stackrel{d}{=} G^{-1}$$

luego, en términos de la función cuantil tenemos que G^{-1} está igualmente distribuida que $aF^{-1} + b$ y, como ambas son funciones crecientes y continuas por la izquierda, tienen que ser iguales.

La Ley de Convergencia de Tipos puede ser demostrada por medio de una cadena de lemas como presenta Billingsley en [4] (páginas 204 - 206) o haciendo uso de la función característica como realiza Loeve en [14] (páginas 216 - 218). En este trabajo, siguiendo el artículo de Cuesta et al [8], presentamos una demostración, a nuestro parecer más sencilla, sin más que tener en cuenta la caracterización presentada sobre variables del mismo tipo en términos de la función cuantil y el Teorema de Skorohod.

Teorema 1.4 (Ley de Convergencia de Tipos)**.** *Sean* $\{a_n\}_{n=0}^{\infty}$ *y* $\{b_n\}_{n=0}^{\infty}$ *dos sucesiones de números reales con* $a_n > 0$*, y sea* $\{X_n\}_{n=1}^{\infty}$ *una sucesión de variables aleatorias tal que* $X_n \xrightarrow{L} X$ *y* $a_n X_n + b_n \xrightarrow{L} Y$*, donde* X *e* Y *son variables aleatorias con leyes no degeneradas. Entonces existen* $a > 0$ *y* b *tales que* $a_n \xrightarrow{n\to\infty} a$*,* $b_n \xrightarrow{n\to\infty} b$*, y además* $Y \stackrel{d}{=} aX + b$*.*

Demostración. Denotamos por F_n, G_n, F y G a las funciones de distribución asociadas a las variables X_n, $a_n X_n + b_n$, X e Y respectivamente y representamos las respectivas funciones cuantiles por F_n^{-1}, G_n^{-1}, F^{-1} y G^{-1}, como es usual.

Para variables aleatorias reales el Teorema de Skorohod nos asegura que

$$X_n \xrightarrow{L} X \quad \Longleftrightarrow \quad F_n^{-1} \xrightarrow{c.s.} F^{-1}.$$

Esto es:

$$\begin{array}{lll} F_n \xrightarrow{d} F & \Longleftrightarrow & F_n^{-1} \xrightarrow{c.s.} F^{-1} \\ G_n \xrightarrow{d} G & \Longleftrightarrow & G_n^{-1} \xrightarrow{c.s.} G^{-1} \quad \Longleftrightarrow \quad a_n F_n^{-1} + b_n \xrightarrow{c.s.} G^{-1} \end{array}$$

En resumen, lo que tenemos es $F_n^{-1} \to F^{-1}$, $a_n F_n^{-1} + b_n \to G^{-1}$ con F^{-1} y G^{-1} no constantes y queremos probar que las sucesiones $\{a_n\}_{n=1}^{\infty}$ y $\{b_n\}_{n=1}^{\infty}$ tienen límites respectivos $a > 0$ y b; y además queremos comprobar que $G^{-1} = aF^{-1} + b$ casi seguro.

Sean x_1, x_2 tales que $F^{-1}(x_1) \neq F^{-1}(x_2)$. Entonces

$$F_n^{-1}(x_1) \to F^{-1}(x_1) \quad \text{y} \quad F_n^{-1}(x_2) \to F^{-1}(x_2)$$

y, en consecuencia,

$$F_n^{-1}(x_1) - F_n^{-1}(x_2) \to F^{-1}(x_1) - F^{-1}(x_2).$$

Además, también se verifican

$$a_n F_n^{-1}(x_1) + b_n \to G^{-1}(x_1) \quad \text{y} \quad a_n F_n^{-1}(x_2) + b_n \to G^{-1}(x_2),$$

de donde se deduce que

$$a_n \left(F_n^{-1}(x_1) - F_n^{-1}(x_2)\right) \to G^{-1}(x_1) - G^{-1}(x_2)$$

y, en consecuencia,

$$a_n \to \frac{G^{-1}(x_1) - G^{-1}(x_2)}{F^{-1}(x_1) - F^{-1}(x_2)} =: a.$$

Ahora bien, si consideramos y_1, y_2 tales que $G^{-1}(y_1) \neq G^{-1}(y_2)$ tenemos que

$$a_n \left(F_n^{-1}(y_1) - F_n^{-1}(y_2)\right) \to G^{-1}(y_1) - G^{-1}(y_2) \neq 0$$

dado que $a_n \to a$ y $F_n^{-1}(y_1) - F_n^{-1}(y_2) \to F^{-1}(y_1) - F^{-1}(y_2)$, necesariamente $a \neq 0$. Como además $a_n > 0$, se deduce que $a > 0$.

Por último, tenemos que $a_n F^{-1}(x_1) + b_n \to G^{-1}(x_1)$ luego $b_n \to G^{-1}(x_1) - aF^{-1}(x_1)$.

Ahora bien, como tenemos que $F_n^{-1} \to F^{-1}$ y las sucesiones $\{a_n\}_{n=1}^{\infty}$ y $\{b_n\}_{n=1}^{\infty}$ tienen límites respectivos $a > 0$ y b, entonces también se verifica $a_n F_n^{-1} + b_n \to aF^{-1} + b$. Teniendo en cuenta que el teorema de Skorohod nos aseguraba que $a_n F_n^{-1} + b_n \xrightarrow{c.s.} G^{-1}$, es evidente que $G^{-1} = aF^{-1} + b$. □

De este resultado se deduce realizando cambios triviales en la demostración previa el siguiente corolario.

Corolario 1.5. *Sean $\{a_n\}_{n=0}^{\infty}$ y $\{b_n\}_{n=0}^{\infty}$ dos sucesiones de números reales con $a_n > 0$, y sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de variables aleatorias tal que $X_n \xrightarrow{L} X$ y $a_n X_n + b_n \xrightarrow{L} X$, donde X es una variable aleatoria no degenerada. Entonces se tiene que $a_n \xrightarrow{n\to\infty} 1$ y $b_n \xrightarrow{n\to\infty} 0$.*

Por otro lado, cabe destacar que si relajamos las hipótesis de la Ley de Convergencia de Tipos permitiendo que la sucesión $\{a_n\}_{n=1}^{\infty}$ sea de números no nulos, esto es, permitiendo que la sucesión pueda tener signos negativos y positivos; el resultado sigue siendo válido en cierto sentido. Trabajando con aplicaciones lineales por simplicidad, si escribimos estos coeficientes como $a_n = sign(a_n)|a_n|$, donde $sign(a_n)$ denota el signo de a_n tenemos que, en virtud de la Ley de Convergencia de Tipos $|a_n| \to a > 0$ y entonces:

- Si $sign(a_n) \to 1$ se tiene que $aX \stackrel{d}{=} Y$ y como $a > 0$ las variables X e Y son del mismo tipo.
- Si $sign(a_n) \to -1$ se tiene que $-aX \stackrel{d}{=} Y$ y como $-a < 0$ entonces las variables X e Y son del mismo tipo si, y sólo si, X es una variable simétrica, esto es, $X \stackrel{d}{=} -X$. En caso de que X no sea una variable aleatoria simétrica las variables X e Y no son del mismo tipo, pero verifican una condición similar.
- Si la sucesión $\{sign(a_n)\}_{n=1}^{\infty}$ no converge, entonces necesariamente X es una variable simétrica y, en consecuencia, $aX \stackrel{d}{=} -aX \stackrel{d}{=} Y$ luego las variables X e Y son del mismo tipo.

Esto nos proporciona una idea de cómo se va a abordar el caso multidimensional, en el que las transformaciones afines α_n van a descomponerse como

producto de transformaciones que convergen a la identidad y transformaciones que dejan fija la distribución del vector límite.

1.2. Preliminares

Antes de seguir avanzado introducimos la notación que utilizaremos en lo que resta de capítulo y algunas observaciones pertinentes. Seguiremos en esencia la exposición del trabajo de Billingsley (1966) [3].

Notación. Denotaremos:

- por r, s, ... a números reales,
- por a, b, ..., x, y, ... a vectores (columna) de números reales de dimensión k,
- por A, B, ... a matrices de dimensión $k \times k$ con coeficientes reales,
- por α, β, ... a transformaciones afines de $\mathbb{R}^k$, teniendo en cuenta que δ denotará siempre a la transformación identidad,
- por X, Y, ... a vectores (columna) aleatorios de dimensión k.

Con la notación expuesta, recordamos que una transformación afín α es de la forma $\alpha\, x = Ax + a$. Se dice que la transformación afín α es no singular si, y sólo si, la matriz A que la define es no singular. En particular, una transformación afín no singular se caracteriza por ser una aplicación biyectiva de $\mathbb{R}^k$ en $\mathbb{R}^k$. Además, las transformaciones afines forman un semigrupo bajo la composición, que representaremos por $\mathscr{S}$, y el conjunto de transformaciones afines no singulares forman un grupo en $\mathscr{S}$, que denotaremos por $\mathscr{G}$.

Por otro lado, al escribir $\alpha_n \to \alpha$, nos estaremos refieriendo siempre a que existe convergencia $\alpha_n\, x \to \alpha\, x$ para cada vector $x \in \mathbb{R}^k$, en el sentido de la norma euclídea. De esta forma, dotamos a $\mathscr{S}$ de una topología bajo la cual la composición es continua.

Definición 1.6. Se dice que un **vector aleatorio** X es **degenerado** si existe una variedad afín H de dimensión $k-1$ tal que X pertenece a H con probabilidad igual a uno.

Estudiamos dos lemas previos para, posteriormente, poder abordar los resultados principales del capítulo.

Notación. Denotamos por $\|Aa\|$ al máximo de los valores absolutos de las entradas del vector Aa y además, como es usual, representamos al producto interno de dos vectores x e y de $\mathbb{R}^k$ por $< x, y >$, esto es, $< x, y >= \sum_{i=1}^{k} x_i y_i$.

Definición 1.7. Se dice que la sucesión $\{X_n\}_{n=1}^{\infty}$ está **acotada en probabilidad** si para todo $\varepsilon > 0$ existe un número real r_ε tal que $\|X_n\| \leq r_\varepsilon$ se cumple con probabilidad mayor que $1 - \varepsilon$ para todo n.

Nótese que en el ámbito de la convergencia de medidas de probabilidad, esta condición es idéntica a la de que la sucesión de distribuciones asociadas $\{\mathscr{L}(X_n)\}_{n=1}^{\infty}$ sea ajustada.

Observación 1. Si $k = 1$ y $\{X_n\}_{n=1}^{\infty}$ está acotada en probabilidad y $s_n \to \infty$, entonces $X_n/s_n \xrightarrow{d} 0$.

Lema 1.8. *Sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de vectores aleatorios verificando $X_n \xrightarrow{d} X$, donde X es un vector aleatorio no degenerado. Supongamos que $\{< x_n, X_n > + r_n\}_{n=1}^{\infty}$ está acotada en probabilidad, entonces* $\sup_n \|x_n\| < \infty$ *y* $\sup_n |r_n| < \infty$.

Demostración. Consideramos $s_n = \|x_n\| + |r_n|$ y suponemos que $\{s_n\}_{n=1}^{\infty}$ no está acotada superiormente. Elegimos una subsucesión $\{s_{n_k}\}_{k=1}^{\infty}$ que tienda hacia infinito y de ella extraemos una subsucesión $\{s_{n_{k_l}}\}_{l=1}^{\infty}$ tal que

$$\frac{x_{n_{k_l}}}{s_{n_{k_l}}} \longrightarrow x_0 \in \mathbb{R}^k \qquad \text{y} \qquad \frac{r_{n_{k_l}}}{s_{n_{k_l}}} \longrightarrow r_0 \in \mathbb{R}.$$

Ahora bien:

$$\left\| \frac{x_{n_{k_l}}}{s_{n_{k_l}}} \right\| \longrightarrow \|x_0\| \quad \Longrightarrow \quad \frac{\|x_{n_{k_l}}\|}{\|x_{n_{k_l}}\| + |r_{n_{k_l}}|} \longrightarrow \|x_0\|$$

$$\left| \frac{r_{n_{k_l}}}{s_{n_{k_l}}} \right| \longrightarrow |r_0| \quad \Longrightarrow \quad \frac{|r_{n_{k_l}}|}{\|x_{n_{k_l}}\| + |r_{n_{k_l}}|} \longrightarrow |r_0|$$

de donde se sigue que

$$1 = \frac{\|x_{n_{k_l}}\|}{\|x_{n_{k_l}}\| + |r_{n_{k_l}}|} + \frac{|r_{n_{k_l}}|}{\|x_{n_{k_l}}\| + |r_{n_{k_l}}|} \longrightarrow \|x_0\| + |r_0|$$

y, por lo tanto, $\|x_0\| + |r_0| = 1$. En consecuencia, no puede verificarse simultáneamente $x_0 = \mathbf{0}$ y $r_0 = 0$. Luego:

$$< \frac{x_{n_{k_l}}}{s_{n_{k_l}}}, X_{n_{k_l}} > + \frac{r_{n_{k_l}}}{s_{n_{k_l}}} \xrightarrow{d} < x_0, X > + r_0. \tag{1.1}$$

Además, como $\{(x_n, X_n) + r_n\}_{n=1}^{\infty}$ está acotada en probabilidad y $s_{n_{k_l}} \to \infty$ entonces

$$< \frac{x_{n_{k_l}}}{s_{n_{k_l}}}, X_{n_{k_l}} > + \frac{r_{n_{k_l}}}{s_{n_{k_l}}} \xrightarrow{d} 0 \tag{1.2}$$

Combinando (1.1) y (1.2) se deduce que $(x_0, X) + r_0 \stackrel{d}{=} 0$, luego X pertenece con probabilidad 1 al conjunto $H = \{x : (x_0, x) + r_0 = 0\}$. Teniendo en cuenta que x_0 y r_0 no pueden anularse simultáneamente y que H es un conjunto no vacío, entonces H es un hiperplano de dimensión $k-1$, de donde se deduciría que X es un vector aleatorio degenerado, lo cual entra en contradicción con las hipótesis del resultado.

Luego necesariamente $\{s_n\}_{n=1}^{\infty}$ es una sucesión acotada y, en consecuencia, se tiene que

$$\sup_n \|x_n\| < \infty \quad \text{y} \quad \sup_n |r_n| < \infty.$$

□

Notación. Si α es una transformación afín dada por $\alpha x = Ax + a$, escribiremos $\|\alpha\| = \text{máx}\{\|A\|, \|a\|\}$. Tomando $d(\alpha, \beta) = \|\alpha - \beta\|$, donde $\alpha - \beta$ es la diferencia punto a punto de las transformaciones α y β, dotamos a $\mathscr{S}$ de una métrica equivalente a la topología de la convergencia puntual.

Lema 1.9. *Sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de vectores aleatorios con $X_n \xrightarrow{d} X$, donde X es un vector aleatorio no degenerado y sea $\{\alpha_n\}_{n=1}^{\infty}$ una sucesión de transformaciones afines. Supongamos que $\{\alpha_n X_n\}_{n=1}^{\infty}$ está acotada en probabilidad, entonces* $\sup_n \|\alpha_n\| < \infty$.

Demostración. Supongamos $\alpha_n X = A_n X + a_n$ y denotamos por A_n^t a la matriz traspuesta de A_n.

Como $\{\alpha_n X_n\}_{n=1}^{\infty}$ está acotada en probabilidad, dado $\varepsilon > 0$ existe un número real $r_\varepsilon > 0$ tal que $\|\alpha_n X_n\| \leq r_\varepsilon$ se verifica con probabilidad mayor que $1 - \varepsilon$.

Ahora bien, dado $x \in \mathbb{R}^k$ tenemos que $\{< x, \alpha_n X_n >\}_{n=1}^{\infty}$ está acotada en probabilidad pues, haciendo uso de la desigualdad de Cauchy–Schwarz se tiene que

$$| < x, \alpha_n X_n > | \leq \|x\| \|\alpha_n X_n\| \leq \|x\| r_\varepsilon$$

con probabilidad mayor que $1 - \varepsilon$.

Por otro lado, dado que se tiene la igualdad

$$< A_n^t x, X_n > + < x, a_n > = < x, \alpha_n X_n >,$$

del lema previo se sigue que

$$\sup_n \|A_n^t x\| < \infty \quad \text{y} \quad \sup_n | < x, a_n > | < \infty.$$

Ahora bien:

- Veamos que $\sup_n\{\|A_n\|\} < \infty$:

  Dado que estamos trabajando en $\mathbb{R}^k$, que es un espacio de dimensión finita, sabemos que todas las normas matriciales son equivalentes. Por lo tanto nos limitamos a trabajar con la norma del máximo y probaremos que $\sup_n\{\|A_n\|_{\text{máx}}\} < \infty$, donde la norma del máximo se define como $\|A\|_{\text{máx}} = \text{máx}_{i,j} |a_{ij}|$.

  Por hipótesis, sabemos que $\sup_n \|A_n^t x\|_\infty < \infty$ para $x \in \mathbb{R}^k$ arbitrario luego, en particular, $\sup_n \|A_n^t e_i\|_\infty < \infty$ para $i = 1, \ldots, k$. Así, para cada $i \in \{1, \ldots, k\}$ existe una constante $M_i > 0$ de manera que $\sup_n \|A_n^t e_i\|_\infty < M_i$ tomando $M = \text{máx}\{M_1, \ldots, M_k\}$ tenemos que $\sup_{1 \leq i \leq k} \sup_n \|A_n^t e_i\|_\infty < M$ o, equivalentemente,

  $$\sup_n \|A_n\|_{\text{máx}} = \sup_n \|A_n^t\|_{\text{máx}} < M < \infty$$

  como queríamos probar.

- Vemos que $\sup_n \|a_n\| < \infty$:

  Como $\sup_n | < x, a_n > | < \infty$ para $x \in \mathbb{R}^k$ arbitrario, en particular,

  $$\sup_n \left\|a_n^{(i)}\right\| = \sup_n | < e_i, a_n > | < \infty.$$

Así, para cada $i \in \{1, \ldots, k\}$ existe una constante $M_i > 0$ de manera que $\sup_n |a_n^{(i)}| < M_i$ tomando $M = \max\{M_1, \ldots, M_k\}$ tenemos que $\sup_{1 \leq i \leq k} \sup_n |a_n^{(i)}| < M$ o, equivalentemente, $\sup_n \|a_n\| < M < \infty$ como queríamos probar.

Ahora bien, por definición $\|\alpha_n\| = \max\{\|A_n\|, \|a_n\|\}$ y como hemos probado que $\sup_n \|A_n\| < \infty$ y $\sup_n \|a_n\| < \infty$, es claro que $\sup_n \|\alpha_n\| < \infty$.

□

1.3. Resultados principales

Empezamos probando que el conjunto de transformaciones afines que conservan la distribución de un vector aleatorio no degenerado es un subgrupo compacto del grupo de transformaciones afines no singulares.

Teorema 1.10. *Si X es un vector aleatorio no degenerado, entonces el conjunto $\mathscr{G}_X$ de todas las transformaciones afines que verifican $\alpha X \stackrel{d}{=} X$ forman un subgrupo compacto del grupo $\mathscr{G}$.*

Demostración. Antes de probar que $\mathscr{G}_X$ es un subgrupo compacto, notamos que toda transformación $\alpha \in \mathscr{G}_X$ debe ser no singular.

Supongamos, razonando por reducción al absurdo, que $\alpha_0 \in \mathscr{G}_X$ con α_0 transformación afín singular, dada por $\alpha_0 X = A_0 X + a_0$. Que α_0 sea singular implica que A_0 es una matriz singular y, por tanto existe un vector $x_0 \in \mathbb{R}^k \setminus \{0\}$ tal que $A_0^t x_0 = 0$.

Ahora bien, como tenemos la igualdad

$$< x_0, a_0 >=< A_0^t x_0, X > + < x_0, a_0 >=< x_0, \alpha_0 X > \stackrel{d}{=} < x_0, X >$$

deducimos que X es un vector aleatorio degenerado, en contradicción con la hipótesis del teorema. En consecuencia, si $\alpha \in \mathscr{G}_X$ tenemos que α es no singular.

Pasamos a probar que $\mathscr{G}_X$ es subgrupo de $\mathscr{G}$.

- Es claro que $\mathscr{G}_X$ es cerrado para la composición, pues si $\alpha, \beta \in \mathscr{G}_X$ tenemos que $\alpha X \stackrel{d}{=} X$ y $\beta X \stackrel{d}{=} X$. Por lo tanto $\beta(\alpha X) \stackrel{d}{=} \beta X \stackrel{d}{=} X$, lo que prueba que $\beta \circ \alpha \in \mathscr{G}_X$.

- Si $\alpha \in \mathscr{G}_X$ entonces

  $$\alpha X \stackrel{d}{=} X \implies AX+a \stackrel{d}{=} X \implies X \stackrel{d}{=} A^{-1}X+A^{-1}a \implies X \stackrel{d}{=} \alpha^{-1}X$$

  lo que prueba que $\alpha^{-1} \in \mathscr{G}_X$.

- Es evidente que la transformación afín definida por la matriz identidad de tamaño k y el vector nulo pertenece a $\mathscr{G}_X$ y se trata del elemento neutro del grupo $\mathscr{G}$.

Ya hemos probado que $\mathscr{G}_X$ es un subgrupo de $\mathscr{G}$. Ahora nos queda demostrar que se trata de un subgrupo compacto.

Si $\alpha_n \in \mathscr{G}_X$ y $\alpha_n \to \alpha$ entonces $X \stackrel{d}{=} \alpha_n X \xrightarrow{d} \alpha X$ de donde se sigue que $\alpha X \stackrel{d}{=} X$ y, por tanto, $\alpha \in \mathscr{G}_X$. Lo que prueba que $\mathscr{G}_X$ es cerrado.

Nos faltaría ver que se trata de un conjunto compacto. Si $\{\alpha_n\}_{n=1}^{\infty} \subset \mathscr{G}_X$ entonces $\{\alpha_n\}_{n=1}^{\infty}$ es acotada en probabilidad pues $\alpha_n X \stackrel{d}{=} X$ para todo n. Esto, junto con el hecho de que $X \xrightarrow{d} X$, con X no degenerado, nos permite aplicar el Lema 1.9 del que se deduce que $\sup_n \|\alpha_n\| < \infty$. Por lo tanto, $\sup\{\alpha : \alpha \in \mathscr{G}_X\} < \infty$. Con lo que concluimos la prueba del resultado. □

A continuación, dada una sucesión de vectores aleatorios que converge en distribución a un vector aleatorio no degenerado, se establece una condición necesaria y suficiente sobre una sucesión de transformaciones afines para que la sucesión de los vectores aleatorios transformada por las transformaciones afines correspondientes siga convergiendo en distribución al mismo vector aleatorio no degenerado.

Teorema 1.11. *Sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de vectores aleatorios tales que $X_n \xrightarrow{d} X$, donde X es un vector aleatorio no degenerado. Entonces, para que se verifique $\alpha_n X_n \xrightarrow{d} X$ es necesario y suficiente que las transformaciones afines α_n tengan la forma*

$$\alpha_n = \delta_n \gamma_n, \tag{1.3}$$

donde

$$\delta_n \to \delta \quad \text{y} \quad \gamma_n \in \mathscr{G}_X. \tag{1.4}$$

Demostración. Primero probamos que se verifica la condición suficiente.

Como $\mathscr{G}_X$ es un grupo compacto cualquier subsucesión $\{\gamma_{n_k}\}_{k=1}^{\infty}$ de la sucesión $\{\gamma_n\}_{n=1}^{\infty}$ admite una subsucesión $\{\gamma_{n_{k_l}}\}_{l=1}^{\infty}$ convergente hacia un elemento $\gamma \in \mathscr{G}_X$. Por lo tanto,

$$\gamma_{n_{k_l}} X_{n_{k_l}} \xrightarrow{d} \gamma X \stackrel{d}{=} X.$$

Además, si cualquier subsucesión $\{\gamma_{n_k}\}_{k=1}^{\infty}$ admite una subsucesión $\{\gamma_{n_{k_l}}\}_{l=1}^{\infty}$ que converge en distribución a X, entonces $\{\gamma_n X_n\}_{n=1}^{\infty}$ converge en distribución a X. En consecuencia,

$$\alpha_n X_n = \delta_n \gamma_n X_n \xrightarrow{d} \delta X \stackrel{d}{=} X,$$

como se quería probar.

Pasamos ahora a probar la condición necesaria del enunciado.

Ya sabemos que si $\alpha_n X_n \xrightarrow{d} X$ entonces la sucesión $\{\alpha_n X_n\}_{n=1}^{\infty}$ es acotada en probabilidad y, del Lema 1.9, junto con la hipótesis de que X es un vector aleatorio no degenerado y $X_n \xrightarrow{d} X$, se sigue que $\sup_n \|\alpha_n\| < \infty$. Luego la sucesión $\{\alpha_n\}_{n=1}^{\infty}$ tiene adherencia compacta. Por lo tanto, si una subsucesión $\{\alpha_{n_k}\}_{k=1}^{\infty}$ converge hacia α tenemos $\alpha_{n_k} X_{n_k} \xrightarrow{d} X$ y $\alpha_{n_k} X_{n_k} \xrightarrow{d} \alpha X$, de lo que se sigue que $\alpha \in \mathscr{G}_X$. Con esto, hemos probado que el conjunto F de todos los puntos límites de la sucesión $\{\alpha_n\}_{n=1}^{\infty}$ es un subconjunto de $\mathscr{G}_X$.

Si denotamos r_n a la distancia de α_n al conjunto F, entonces $r_n \xrightarrow{n\to\infty} 0$. Elegimos $\gamma_n \in F$ con $\|\alpha_n - \gamma_n\| < r_n + 1/n$ y consideramos $\delta_n = \alpha_n \gamma_n^{-1}$.

Ahora bien, $\alpha_n = \delta_n \gamma_n$ y $\gamma_n \in \mathscr{G}_X$ y sólo queda probar que $\delta_n \to \delta$ o, equivalentemente $\delta_n - \delta = (\alpha_n - \gamma_n)\,\gamma_n^{-1} \to 0$. Para ello, observamos que $\|\alpha_n - \gamma_n\| < r_n + 1/n \xrightarrow{n\to\infty} 0$, lo que prueba que $\alpha_n - \gamma_n \xrightarrow{n\to\infty} 0$. Además teniendo en cuenta que $\sup_n \|\gamma_n^{-1}\| \leq \sup\{\|\gamma\| \,:\, \gamma \in \mathscr{G}_X\} < \infty$, se concluye que $\delta_n - \delta \to 0$ como queríamos. □

La caracterización obtenida para las transformaciones afines α_n, en función de transformaciones que convergen hacia la identidad compuestas con transformaciones no singulares, nos asegura que para n suficientemente grande éstas son siempre no singulares.

Sabemos qué significa que dos variables aleatorias sean del mismo tipo, si queremos llevar este concepto a dimensión mayor que uno debemos pensar en una matriz definida positiva (que sería el equivalente a la constante mayor que

cero) y un vector de localización (que sería el equivalente a la constante que no presentaba restricciones), es decir, tenemos que pensar en una transformación afín con matriz definida positiva. De esta manera, diremos que dos **vectores aleatorios** X e Y son **del mismo tipo**, si existe un transformación afín definida positiva α tal que $\mathscr{L}(\alpha X) = \mathscr{L}(Y)$.

Una vez conocido el concepto de vectores aleatorios del mismo tipo, podemos introducir un resultado preliminar que nos servirá para, posteriormente, presentar una generalización de la Ley de Convergencia de Tipos a espacios de dimensión mayor que uno.

Teorema 1.12. *Sea $\{\alpha_n\}_{n=1}^{\infty}$ una sucesión de transformaciones afines y sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de vectores aleatorios. Si $X_n \xrightarrow{d} X$ y $\alpha_n X_n \xrightarrow{d} Y$, donde X e Y son vectores aleatorios no degenerados, entonces $\alpha X \stackrel{d}{=} Y$ para algún α en $\mathscr{G}$.*

Demostración. Como $\alpha_n X_n \xrightarrow{d} Y$ sabemos que la sucesión $\{\alpha_n X_n\}_{n=1}^{\infty}$ es acotada en probabilidad. Además, como por hipótesis $X_n \xrightarrow{d} X$ siendo X un vector aleatorio no degenerado, del Lema 1.9 se tiene que $\sup_n \|\alpha_n\| < \infty$. Consideramos una subsucesión $\{\alpha_{n_k}\}_{k=1}^{\infty}$ de $\{\alpha_n\}_{n=1}^{\infty}$ tal que $\alpha_{n_k} \to \alpha$. Ahora bien, como

$$\alpha_{n_k} X_{n_k} \xrightarrow{d} \alpha X \quad \text{y} \quad \alpha_{n_k} X_{n_k} \xrightarrow{d} Y$$

se deduce que $\alpha X \stackrel{d}{=} Y$. Dado que X e Y son vectores aleatorios no degenerados por hipótesis, tenemos que necesariamente α debe de ser una transformación no singular y, por tanto, $\alpha \in \mathscr{G}$. □

Estamos en condiciones de presentar y demostrar una generalización de la Ley de Convergencia de Tipos a espacios de dimensión mayor que uno. La prueba de este resultado se limita a combinar lo obtenido en los dos teoremas previos.

Teorema 1.13. *Sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de vectores aleatorios y sean $\{\alpha_n\}_{n=1}^{\infty}$ y $\{\beta_n\}_{n=1}^{\infty}$ sucesiones de transformaciones afines, siendo las transformaciones α_n no singulares para todo $n \in \mathbb{N}$. Supongamos que $\beta_n X_n \xrightarrow{d} X$, donde X es un vector aleatorio no degenerado. Entonces β_n es no singular para n suficientemente grande, y para que se cumpla $\alpha_n X_n \xrightarrow{d} Y$, donde Y es un vector aleatorio no degenerado, es necesario y suficiente que $Y \stackrel{d}{=} \alpha X$*

para algún $\alpha \in \mathscr{G}$ y que para n suficientemente grande la transformación afín α_n sea de la forma

$$\alpha_n = \alpha\delta_n\gamma_n\beta_n,$$

donde

$$\delta_n \to \delta \quad \text{y} \quad \gamma_n \in \mathscr{G}_X.$$

Demostración. Que las transformaciones afines β_n sean no singulares para n suficientemente grande se deduce del hecho de que X es un vector aleatorio no degenerado.

Probemos en primer lugar la condición suficiente, esto es, si $Y \stackrel{d}{=} \alpha X$ para algún $\alpha \in \mathscr{G}$ y para n suficientemente grande (digamos $n \geq n_0$) las transformaciones afines α_n son de la forma $\alpha_n = \alpha\delta_n\gamma_n\beta_n$ con $\delta_n \to \delta$ y $\gamma_n \in \mathscr{G}_X$, entonces $\alpha_n X_n \xrightarrow{d} Y$.

Como $\mathscr{G}_X$ es compacto, cualquier subsucesión $\{\gamma_{n_k}\}_{k=1}^{\infty}$ de la sucesión $\{\gamma_n\}_{n=n_0}^{\infty}$ (que está contenida en $\mathscr{G}_X$) admite una subsucesión $\{\gamma_{n_{k_l}}\}_{l=1}^{\infty}$ convergente hacia un elemento $\gamma \in \mathscr{G}_X$, por lo tanto, $\gamma_{n_{k_l}} \to \gamma$.

Ahora bien, como $\beta_n X_n \xrightarrow{d} X$ entonces

$$\gamma_{n_{k_l}}\beta_{n_{k_l}}X_{n_{k_l}} \xrightarrow{d} \gamma X \stackrel{d}{=} X,$$

sin embargo, el Teorema de compacidad de Helly nos asegura que esta convergencia se verifica para todo $n \in \mathbb{N}$. Luego tenemos $\gamma_n\beta_n X_n \xrightarrow{d} X$ y, en consecuencia, $\alpha_n X_n = \alpha\delta_n\gamma_n\beta_n X_n \xrightarrow{d} \alpha\delta X \stackrel{d}{=} \alpha X$.

Por hipótesis se tiene $\alpha_n X_n \xrightarrow{d} Y$ y hemos probado que $\alpha_n X_n \xrightarrow{d} \alpha X$, en consecuencia, $Y \stackrel{d}{=} \alpha X$ como queríamos probar.

Pasamos a demostrar la condición necesaria, esto es, si $\alpha_n X_n \xrightarrow{d} Y$ entonces $Y \stackrel{d}{=} \alpha X$ para algún $\alpha \in \mathscr{G}$ y para n suficientemente grande las transformaciones afines α_n son de la forma $\alpha_n = \alpha\delta_n\gamma_n\beta_n$ donde $\delta_n \to \delta$ y $\gamma_n \in \mathscr{G}_X$.

Por hipótesis tenemos que $\beta_n X_n \xrightarrow{d} X$ y que las transformaciones afines β_n son no singulares para n suficientemente grande. Vamos a escribir $Y_n = \beta_n X_n$ y entonces $X_n = \beta_n^{-1} Y_n$ para n suficientemente grande, pues β_n es no singular.

Además, por hipótesis también sabemos que $\alpha_n X_n \xrightarrow{d} Y$ y si sustituimos en esta expresión X_n por $\beta_n^{-1} Y_n$ se sigue que $\alpha_n\beta_n^{-1} Y_n \xrightarrow{d} Y$.

Ahora bien, tenemos las convergencias $Y_n = \beta_n X_n \xrightarrow{d} X$ y $\alpha_n \beta_n^{-1} Y_n \xrightarrow{d} Y$ con X e Y vectores aleatorios no degenerados. Así, estamos en condiciones de aplicar el Teorema 1.12 del que se deduce que $\alpha X \stackrel{d}{=} Y$ para algún $\alpha \in \mathscr{G}$.

Con esto tenemos

$$\alpha_n \beta_n^{-1} Y_n \xrightarrow{d} \alpha X \implies \alpha^{-1} \alpha_n \beta_n^{-1} Y_n \xrightarrow{d} X,$$

teniendo en cuenta que $X_n = \beta_n^{-1} Y_n$, se sigue que

$$Y_n \xrightarrow{d} X \quad \text{y} \quad \alpha^{-1} \alpha_n \beta_n^{-1} Y_n \xrightarrow{d} X$$

donde X es un vector aleatorio no degenerado y, en virtud del Teorema 1.11 se sigue que $\alpha^{-1} \alpha_n X_n \beta_n^{-1} = \delta_n \gamma_n$ donde $\delta_n \to \delta$ y $\gamma_n \in \mathscr{G}_X$, para n suficientemente grande (basta tomar n como el número natural a partir del cual las transformaciones afines β_n son no singulares).

Despejando en la igualdad anterior, se deduce que, para n suficientemente grande, las transformaciones afines α_n son de la forma $\alpha_n = \alpha \delta_n \gamma_n \beta_n$, donde $\alpha \in \mathscr{G}$, $\delta_n \to \delta$ y $\gamma_n \in \mathscr{G}_X$, como se quería probar. □

Por último, se presenta una caracterización de los subgrupos compactos del grupo de transformaciones afines no singulares en términos de un grupo cerrado de matrices ortogonales, una matriz simétrica y definida positiva y un punto de $\mathbb{R}^k$.

Teorema 1.14. *Todo subgrupo compacto $\mathscr{G}_0$ de $\mathscr{G}$ tiene la siguiente forma: Existe un subgrupo cerrado $\mathcal{O}_0$ de $\mathcal{O}$, un punto $x_0 \in \mathbb{R}^k$, y una matriz V simétrica y definida positiva, tal que $\mathscr{G}_0$ consiste exactamente en las trasformaciones afines*

$$\alpha x = V^{-1} W V (x - x_0) + x_0$$

con $W \in \mathcal{O}_0$.

Demostración. Dada una transformación afin α definida por $\alpha x = Ax + a$, denotamos por $J(\alpha)$ al determinante de la matriz A, esto es, $J(\alpha) = \det(A)$. Como $\alpha^n x = A^n x + A^{n-1} a + \cdots + Aa + a$ y $\alpha^{-1} x = A^{-1} x - A^{-1} a$, tenemos que

$$J(\alpha^n) = J(\alpha)^n \quad \text{y} \quad J\left(\alpha^{-1}\right) = J(\alpha)^{-1}. \tag{1.5}$$

Luego la función J es continua en la topología de $\mathscr{S}$ y, al tratarse $\mathscr{G}_0$ de un subgrupo cerrado de $\mathscr{G}$ se tiene que J está acotada en $\mathscr{G}_0$. Entonces, existe una constante $M > 0$ de modo que $|J(\alpha)| < M$ para todo $\alpha \in \mathscr{G}_0$.

Sabiendo que J es una función continua, vamos a probar que sólamente puede ser $J(\alpha) = \pm 1$.

- Supongamos $\beta \in \mathscr{G}_0$ con $J(\beta) > 1$. Como $\mathscr{G}_0$ es un subgrupo, se tiene que $\beta^n \in \mathscr{G}_0$ para todo $n \in \mathbb{N}$. Entonces $J(\beta^n) = J(\beta)^n \xrightarrow{n\to\infty} \infty$ pues $J(\beta) > 1$. Luego, de la definición de límite, se tiene que existe $n_0 \in \mathbb{N}$ tal que $|J(\beta^n)| \geq M$ para todo $n \geq n_0$, lo cual entra en contradicción con la acotación de la función J en $\mathscr{G}_0$.
- El caso en el que $\beta \in \mathscr{G}_0$ con $J(\beta) < -1$ es totalmente análogo al anterior, teniendo en cuenta que en este caso el límite de $J\left(\beta^{(2n-1)}\right)$ será $-\infty$ y el límite de $J\left(\beta^{(2n)}\right)$ será ∞.
- Si $\beta \in \mathscr{G}_0$ con $J(\beta) \in (-1, 1) \setminus \{0\}$ entonces nuevamente se tiene que $\beta^n \in \mathscr{G}_0$ y $J(\beta^n) \xrightarrow{n\to\infty} 0$. Ahora bien, como $\mathscr{G}_0$ es compacto y $\{\beta^n\}_{n=1}^{\infty} \subset \mathscr{G}_0$ existe una subsucesión $\{\beta^{n_k}\}_{k=1}^{\infty}$ convergente hacia un elemento $\gamma \in \mathscr{G}_0$. Dado que J es continua tenemos que $J(\gamma) = 0$, lo cual es absurdo pues $\gamma \in \mathscr{G}$ y, por tanto γ debe ser una transformación afín no singular.

Luego necesariamente $J(\alpha) = \pm 1$ para todo $\alpha \in \mathscr{G}_0$.

Sea L un conjunto abierto y acotado de $\mathbb{R}^k$ y consideramos

$$M = \bigcup \{\alpha L \,:\, \alpha \in \mathscr{G}_0\}.$$

Entonces M es acotado y abierto, y además verifica que $\alpha M = M$ para todo $\alpha \in \mathscr{G}_0$. Sea x_0 el vector de componentes $|M|^{-1} \int_M x_i dx$, donde $|M|$ denota la medida de Lebesgue del conjunto M. Como el jacobiano $|J(\alpha)| = 1$ un cambio de variable demuestra que $\alpha x_0 = x_0$, para todo $\alpha \in \mathscr{G}_0$. Luego x_0 es un punto fijo de todas las transformaciones afines de $\mathscr{G}_0$. Así, el subgrupo compacto $\mathscr{G}_0$ consiste en las transformaciones afines de la forma

$$\alpha x = A(x - x_0) + x_0$$

donde $A \in \mathscr{G}_1$ y $\mathscr{G}_1$ es un grupo compacto de transformaciones lineales.

Veamos que efectivamente el conjunto $\mathscr{G}_1$ definido como $\mathscr{G}_1 = \{A \in \mathbb{R}^{k\times k} : \alpha x = A(x - x_0) + x_0, \alpha \in \mathscr{G}_0\}$ es un subgrupo compacto de transformaciones lineales.

- Veamos en primer lugar que $\mathscr{G}_1$ es cerrado para la composición, esto es, queremos probar que si $A, B \in \mathscr{G}_1$ entonces $AB \in \mathscr{G}_1$.

  Si $A, B \in \mathscr{G}_1$ existen transformaciones afines $\alpha, \beta \in \mathscr{G}$ tales que $\alpha x = A(x - x_0) + x_0$ y $\beta x = B(x - x_0) + x_0$. Como $\mathscr{G}_0$ es un subgrupo de $\mathscr{G}$ sabemos que $\alpha \circ \beta \in \mathscr{G}_0$, además observamos que

  $$\begin{aligned}(\alpha \circ \beta)x &= \alpha(B(x - x_0) + x_0) = \\ &= A\left[(B(x - x_0) + x_0) - x_0\right] + x_0 = AB(x - x_0) + x_0\end{aligned}$$

  lo que prueba que la transformación lineal dada por la matriz AB pertenece a $\mathscr{G}_1$.

- Veamos que el elemento neutro del grupo de transformaciones lineales, que es claramente la transformación lineal definida por la matriz identidad, pertenece a $\mathscr{G}_1$. Esto es equivalente a mostrar que la transformación afín $ix = I(x - x_0) + x_0$ pertenece a $\mathscr{G}_0$. Ahora bien, la transformación anterior es la identidad y, al tratarse $\mathscr{G}_0$ de un subgrupo de $\mathscr{G}$, sabemos que el elemento neutro estará contenido en $\mathscr{G}_0$.

- A continuación probamos que $\mathscr{G}_1$ es cerrado para elementos inversos, esto es, si $A \in \mathscr{G}_1$ entonces $A^{-1} \in \mathscr{G}_1$.

  Si $A \in \mathscr{G}_1$ entonces la transformación $\alpha x = A(x - x_0) + x_0 = Ax + (I - A)x_0$ está en $\mathscr{G}_0$ y, nuevamente por ser $\mathscr{G}_0$ subgrupo sabemos que $\alpha^{-1} \in \mathscr{G}_0$. Ya sabemos que $\alpha^{-1}x = A^{-1}x - A^{-1}(I - A)x_0 = A^{-1}(x - x_0) + x_0$ luego por definición de $\mathscr{G}_1$, tenemos que $A^{-1} \in \mathscr{G}_1$.

- La asociatividad de $\mathscr{G}_1$ es evidente de la asociatividad del procuto de matrices.

- Para probar que $\mathscr{G}_1$ es compacto vamos a ver que es acotado y cerrado.

  La acotación es clara, pues para cada transformación lineal $A \in \mathscr{G}_1$ existe una transformación afín $\alpha \in \mathscr{G}_0$ con matriz asociada A, entonces $\|A\| \leq \|\alpha\|$, así $\sup\{\|A\| : A \in \mathscr{G}_1\} \leq \sup\{\|\alpha\| : \alpha \in \mathscr{G}_0\} < \infty$ lo que asegura la acotación de $\mathscr{G}_1$.

Para ver que se trata de un conjunto cerrado razonamos tomando $A \in \overline{\mathscr{G}_1} \setminus \mathscr{G}_1$, entonces existe una sucesión de transformaciones lineales $\{A_n\}_{n=1}^{\infty} \subset \mathscr{G}_1$ tal que $A_n x \xrightarrow{n\to\infty} Ax$ o, equivalentemente

$$A_n(x - x_0) + x_0 \xrightarrow{n\to\infty} A(x - x_0) + x_0.$$

Definiendo $\alpha_n x = A_n(x - x_0) + x_0$ para todo número natural n, se tiene que la sucesión de transformaciones afiens $\{\alpha_n\}_{n=1}^{\infty}$ está contenida en $\mathscr{G}_0$ que se trata de un subgrupo compacto, entonces considerando la transformación afín $\alpha x = A(x - x_0) + x_0$ se tiene que $\alpha \in \mathscr{G}_0$ y, en consecuencia, $A \in \mathscr{G}_1$.

Sea L un conjunto acotado y abierto de $\mathbb{R}^k$, como antes, el conjunto

$$N = \bigcup \{A^t L \ : \ A \in \mathscr{G}_1\}$$

es acotado y abierto, y además $A^t N = N$ para toda transformación lineal $A \in \mathscr{G}_1$. Sea U la matriz de entradas $u_{ij} = \int_N z_i z_j dz$. Entonces U es simétrica y

$$< x, Uy >= \int_N < x, z >< y, z > dz.$$

Luego $< x, Ux > \geq 0$ y además, si $< x, Ux >= 0$ entonces $< x, z >= 0$ para todo $z \in N$ lo cual implica que $x = 0$ por ser N un conjunto abierto. En consecuencia, la matriz U es definida positiva. Y como para toda transformación lineal $A \in \mathscr{G}_1$, se tiene que $\det(A) = \pm 1$, un cambio de variable prueba que

$$< Ax, UAy >=< x, Uy > . \tag{1.6}$$

Por otro lado, como U es simétrica y definida positiva entonces existe una matriz V simétrica y definida positiva tal que $V^2 = U$ (la matriz V se denomina raíz cuadrada de U). Observamos que $V^{-1}UV^{-1} = I$ y vamos a probar que la matriz VAV^{-1} es ortogonal.

Recordamos que en un espacio euclídeo, en nuestro caso estamos trabajando en $\left(\mathbb{R}^k, < \cdot, \cdot >\right)$, las matrices ortogonales se definen de la siguiente manera:

$$\text{“}O \in \mathcal{O} \iff < x, y >=< Ox, Oy >, \ \ \forall x, y \in \mathbb{R}^k\text{.”}$$

Por lo tanto, vamos a probar que $< x, y >=< VAV^{-1}x, VAV^{-1}y >$ para todo par de vectores $x, y \in \mathbb{R}^k$, haciendo uso de la igualdad (1.6) y de la

clásica propiedad $< Mx, y >=< x, M^t y >$, teniendo en cuenta que $V = V^t$ y $V^2 = U$:

$$\begin{aligned} < VAV^{-1}x, VAV^{-1}y > &=< AV^{-1}x, VVAV^{-1}y >= \\ &=< A\left(V^{-1}x\right), UA\left(V^{-1}y\right) >=< V^{-1}x, UV^{-1}y >= \\ &=< V^{-1}x, V^t y >=< x, (V^{-1})^t V^t y >=< x, y > \end{aligned}$$

como se quería probar.

Ahora bien, para probar que la familia $\mathcal{O}_0 = \{VAV^{-1} : A \in \mathscr{G}_1\}$ es un subgrupo compacto de $\mathcal{O}$ basta deducirlo del hecho de que $\mathscr{G}_1$ es un subgrupo compacto:

- Si $U_1, U_2 \in \mathcal{O}_0$ entonces existen matrices $A_1, A_2 \in \mathscr{G}_1$ de manera que $U_i = VA_iV^{-1}$. Así, $U_1U_2 = V(A_1A_2)V^{-1} \in \mathcal{O}_0$ pues $A_1A_2 \in \mathscr{G}_1$.
- Nuevamente la asociatividad de $\mathcal{O}_0$ se debe a la asociatividad del producto de matrices.
- Claramente la matriz identidad pertenece a $\mathcal{O}_0$, puesto que $I \in \mathscr{G}_1$ entonces $\mathcal{O}_0 \ni VIV^{-1} = I$.
- Dada $U \in \mathcal{O}_0$ existe $A \in \mathscr{G}_1$ de manera que $U = VAV^{-1}$. Observamos que $U^{-1} = VA^{-1}V^{-1}$ y como $A^{-1} \in \mathscr{G}_1$, entonces $U^{-1} \in \mathcal{O}_0$.
- La compacidad es clara pues la aplicación $\Phi : \mathscr{G} \to \mathcal{O}$ definida por $\Phi(A) = VAV^1$ es continua y verifica $\Phi(\mathscr{G}_1) = \mathcal{O}_0$. Como $\mathscr{G}_0$ es compacto en $\mathscr{G}$ entonces su imagen por Φ es compacta en $\mathcal{O}$.

En consecuencia, la familia $\mathcal{O}_0 = \{VAV^{-1} : A \in \mathscr{G}_1\}$ es un subgrupo compacto del grupo de las matrices ortogonales.

Así, cada transformación afín $\alpha \in \mathscr{G}_0$ la podemos escribir como

$$\alpha x = V^{-1}(VAV^{-1})V(x - x_0) + x_0$$

donde x_0 es un vector de $\mathbb{R}^k$, V es una matriz simétrica y definida positiva y la matriz VAV^{-1} es ortogonal.

□

Este último resultado viene a decir que con una elección adecuada del origen de coordenadas y de la correspondiente base, el grupo compacto $\mathscr{G}_0$ se puede identificar con un subgrupo compacto de las matrices ortogonales.

Para el caso unidimensional, es claro que el grupo de matrices ortogonales está formado únicamente por la identidad y por su opuesta, esto es, la transformación identidad y la reflexión respecto del punto cero. Por tanto, los únicos subgrupos son el propio grupo $\mathcal{O}$ y el grupo formado únicamente por la identidad. Así, como ya se ha comentado, el grupo de transformaciones que conservan la distribución de una variable no degenerada consiste o bien en sólo la identidad, o bien en la identidad junto con la reflexión respecto de algún punto.

Este resultado invita a profundizar en la caracterización de los subgrupos compactos del grupo de transformaciones afines no singulares, pero en este trabajo no vamos a seguir trabajando sobre este tipo de resultados.

Capítulo 2

El problema del transporte óptimo

2.1. Problema del transporte óptimo

Supongamos que tenemos una montaña de arena y un agujero que necesitamos rellenar con esa arena. Obviamente la montaña de arena y el agujero deben tener el mismo volumen que vamos a normalizar para que sea uno, y así poder hablar en un contexto de probabilidades.

En nuestra modelización del problema, tanto la masa de arena como el agujero van a modelizarse mediante probabilidades μ y ν definidas respectivamente en espacios medibles Ω_1 y Ω_2. De esta manera, para conjuntos medibles cualesquiera $A \subset \Omega_1$ y $B \subset \Omega_2$, $\mu(A)$ representará la medida de la cantidad de arena localizada en A y $\nu(B)$ representará la medida de la cantidad de arena que debe ser utilizada para rellenar B.

Razonablemente, transportar arena de un lugar a otro va a requerir de un esfuerzo, que va a ser modelado mediante una función medible que vamos a llamar función coste definida en $\Omega_1 \times \Omega_2$. De manera informal, $c(x, y)$ representará el coste de transportar una unidad de masa desde la localización x hasta la localización y. Es natural suponer que la función c, además de ser medible, se trata de una función positiva; a priori, no se debe excluir la posibilidad de que la función c tome valores infinitos, luego c es una función medible tal que $c : \Omega_1 \times \Omega_2 \to \mathbb{R} \cup \{\infty\}$.

La cuestión central que trata de resolver el problema del transporte óptimo es: ¿cómo se puede realizar el transporte descrito minimizando el coste del mismo?

Antes de adentrarnos en el estudio de la pregunta expuesta, debemos clarificar que entendemos por transportar o por **función de transferencia**. Debemos modelar las funciones de transferencia por medidas de probabilidad π en el espacio producto $\Omega_1 \times \Omega_2$. De manera informal, $d\pi(x, y)$ mide la cantidad de masa transferida de la localización x a la localización y. A priori no vamos a excluir la posibilidad de que la masa localizada en un punto x pueda ser repartida en diversas localizaciones (esto es, que tenga diversos destinos de llegada y). Para que una función de transferencia $\pi \in P(\Omega_1 \times \Omega_2)$ sea admisible, informalmente necesitaríamos que toda la masa tomada del punto x coincida con $d\mu(x)$ y, análogamente, toda la masa transportada al punto y coincida con $d\nu(y)$. Formalmente, establecemos este requisito a través de las relaciones:

$$\int_{\Omega_2} d\pi(x, y) = d\mu(x), \quad \int_{\Omega_1} d\pi(x, y) = d\nu(y),$$

es decir, que $\pi \in P(\Omega_1 \times \Omega_2)$, $\mu \in P(\Omega_1)$ y $\nu \in P(\Omega_2)$ verifiquen

$$\pi(A \times \Omega_2) = \mu(A), \quad \pi(\Omega_1 \times B) = \nu(B), \tag{2.1}$$

para todos los conjuntos medibles A de Ω_1 y B de Ω_2. Estas probabilidades $\pi \in P(\Omega_1 \times \Omega_2)$ que verifican la condición inmediatamente anterior se dice que tienen marginales μ y ν, y cualquier probabilidad (conjunta), con estas marginales, se considerará como una **función de transferencia admisible**. Denotaremos al conjunto de estas probabilidades por $\Pi(\mu, \nu)$, esto es, considerando $(\Omega_1, \mu, \sigma_1)$ y $(\Omega_2, \nu, \sigma_2)$ espacios probabilíticos

$$\Pi(\mu, \nu) = \left\{ \pi \in P(\Omega_1 \times \Omega_2) \ : \begin{array}{l} \pi(A \times \Omega_2) = \mu(A), \ \forall A \in \sigma_1 \\ \\ \pi(\Omega_1 \times B) = \nu(B), \ \forall B \in \sigma_2 \end{array} \right\}.$$

Además, considerando $(\varphi, \psi) \in L^1(\Omega_1, \mu) \times L^1(\Omega_2, \nu)$ y teniendo en cuenta la linealidad de la esperanza, la condición (2.1) es equivalente a

$$\int_{\Omega_1 \times \Omega_2} (\varphi(x) + \psi(y)) d\pi(x, y) = \int_{\Omega_1} \varphi(x) d\mu(x) + \int_{\Omega_2} \psi(y) d\nu(y). \tag{2.2}$$

Observación 2. Dadas $\mu \in P(\Omega_1)$ y $\nu \in P(\Omega_2)$ el conjunto $\Pi(\mu, \nu)$ es siempre no vacío. Veámoslo.

Demostración. Consideramos el espacio producto $\Omega_1 \times \Omega_2$ y denotamos por σ a la correspodiente σ-álgebra producto.

Para cada $D \in \sigma$ y para cada $x \in \Omega_1$, $y \in \Omega_2$ se definen las secciones

$$D_x = \{\tilde{y} \in \Omega_2 \,:\, (x, \tilde{y}) \in D\} \subset \Omega_2 \quad \Longrightarrow \quad D_x \in \sigma_2,$$

$$D_y = \{\tilde{x} \in \Omega_1 \,:\, (\tilde{x}, y) \in D\} \subset \Omega_1 \quad \Longrightarrow \quad D_y \in \sigma_1.$$

Definiendo las funciones $\phi : \Omega_1 \to [0, \infty]$ y $\psi : \Omega_2 \to [0, \infty]$ como

$$\phi(x) = \nu(D_x) \quad \Longrightarrow \quad \phi \text{ medible en } \Omega_1,$$

$$\psi(y) = \mu(D_y) \quad \Longrightarrow \quad \psi \text{ medible en } \Omega_2.$$

Siguiendo el libro de Ash [1] (Corolario 2.6.3), consideramos la probabilidad producto como

$$(\mu \times \nu)(D) = \int_{\Omega_1} \phi(x) d\mu(x) = \int_{\Omega_2} \psi(y) d\nu(y),$$

sabiendo que $\mu \times \nu$ es la única probabilidad en $\Omega_1 \times \Omega_2$ verificando

$$(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$$

para todos $A \in \sigma_1$ y $B \in \sigma_2$. Así, es claro que $\mu \times \nu \in \Pi(\mu, \nu)$. □

Con los conceptos expuestos, siguiendo el clásico texto de Villani [20], ya podemos introducir la formulación del problema del transporte óptimo debida a Kantorovich.

Definición 2.1 (**Problema de Kantorovich**)**.** En el contexto introducido y con la notación expuesta el problema de Kantorovich del transporte óptimo es el siguiente problema de minimización:

$$\underset{\pi \in \Pi(\mu,\nu)}{\text{Minimizar}} \quad I(\pi) := \int_{\Omega_1 \times \Omega_2} c(x, y) d\pi(x, y).$$

Este problema de minimización fue estudiado en los años 40 por el matemático de la antigua Unión Soviética Kantorovich que fue galardonado con un premio Nobel por sus aportaciones en Economía. La relación entre el problema del transporte óptimo con la Economía es más que evidente, basta con

pensar en vez de una montaña de arena y un agujero que rellenar, en una densidad de producción y una densidad de consumidores que satisfacer, en un contexto discreto.

Para una función de transferencia $\pi \in \Pi(\mu, \nu)$, la cantidad no negativa pero posiblemente infinita $I(\pi)$ se denomina **coste total del transporte asociado a $\boldsymbol{\pi}$**. En la misma línea, el **coste óptimo de trasporte entre $\boldsymbol{\mu}$ y $\boldsymbol{\nu}$** es el valor

$$T_c(\mu, \nu) = \text{ínf}\{I(\pi) \ : \ \pi \in \Pi(\mu, \nu)\}.$$

Las probabilidades $\pi \in \Pi(\mu, \nu)$ que verifican $I(\pi) = T_c(\mu, \nu)$, en caso de que existan, se denominan **planes de transferencia óptimos** (optimal transference plans, en inglés).

Esta formulación debida a Kantorovich se puede interpretar en términos más probabilísticos teniendo en cuenta lo siguiente. Dado un espacio (Ω, σ, P) recordamos que cada variable aleatoria X en Ω define una medida de probabilidad π_X de la siguiente manera $\pi_X(A) = P(X^{-1}(A))$ para cada $A \in \sigma$, y además, escribimos $\mathscr{L}(X) = \pi_X$. En el contexto que estamos trabajando nos interesan variables aleatorias U y V en un espacio probabilistico común (Ω, σ, P) que toman valores en Ω_1 y Ω_2 respectivamente y que, además, verifican $\mathscr{L}(U) = \mu$ y $\mathscr{L}(V) = \nu$.

Definición 2.2 (**Problema de Kantorovich - Versión probabilista**)**.** Dados dos espacios de probabilidad $(\Omega_1, \sigma_1, \mu)$ y $(\Omega_2, \sigma_2, \nu)$, el problema de Kantorovich del transporte óptimo en su versión probabilista trata de encontar el mínimo valor de

$$I(U, V) := E(c(U, V))$$

en el conjunto de variables aleatorias U, V sobre un espacio probabilístico común (Ω, σ, P) de manera que U toma valores en (Ω_1, σ_1), V toma valores en (Ω_2, σ_2) y, además verifican $\mathscr{L}(U) = \mu$ y $\mathscr{L}(V) = \nu$. Equivalentemente, esta versión probabilista plantea el siguiente problema de minimización:

$$\begin{aligned} \text{Minimizar} \quad & I(U, V) := E(c(U, V)) \\ \text{sujeto a} \quad & U \text{ variable aleatoria de } (\Omega, \sigma, P) \text{ en } (\Omega_1, \sigma_1) \text{ con } \mathscr{L}(U) = \mu \\ & V \text{ variable aleatoria de } (\Omega, \sigma, P) \text{ en } (\Omega_2, \sigma_2) \text{ con } \mathscr{L}(V) = \nu. \end{aligned}$$

Al final, con las restricciones lo que estamos haciendo es quedarnos con los vectores aleatorios que toman valores en $\Omega_1 \times \Omega_2$ cuyas funciones de distribución conjuntas vienen dadas por una probabilidad del conjunto $\Pi(\mu, \nu)$.

Con lo cual es más que evidente que ambos problemas de minimización son equivalentes.

Ahora bien, el problema de Kantorovich no es más que una versión más débil del problema del transporte óptimo original considerado por Monge. La diferencia entre ambas formulaciones reside en el hecho de que el problema de Monge no permite repartir masa, esto es, cada localización x tiene un único destino y. En términos de variables aleatorias, esta condición extra significa que, con la notación anterior, buscamos una variable aleatoria V que sea función de la variable aleatoria U. Esto mismo en términos de funciones de transferencia, se traduce en buscar una probabilidad $\pi \in \Pi(\mu, \nu)$ de la forma

$$d\pi(x, y) = d\pi_T(x, y) = d\mu(x)\delta\left[y = T(x)\right], \tag{2.3}$$

donde $T : \Omega_1 \to \Omega_2$ es una función medible y, como es usual, δ representa la función Kronecker (recordamos que toma el valor 1 si efectivamente $y = T(x)$ y 0 en otro caso).

La probabilidad que aparece en el lado derecho de la ecuación (2.3), que también suele ser denotada como $(Id \times T)\#\mu$, es una medida de probabilidad en $\Omega_1 \times \Omega_2$ que satisface la siguiente propiedad: "para cualquier función medible en $\Omega_1 \times \Omega_2$ y positiva $\mathcal{C}$, se verifica

$$\int_{\Omega_1 \times \Omega_2} \mathcal{C}(x, y) d\pi_T(x, y) = \int_{\Omega_1} \mathcal{C}(x, T(x)) d\mu(x)."$$

En particular, el coste total asociado al transporte es

$$I(\pi_T) = \int_{\Omega_1} c(x, T(x)) d\mu(x).$$

Ahora bien, ¿qué condición debe cumplir π_T en (2.3) para que $\pi_T \in \Pi(\mu, \nu)$? Teniendo en cuenta la expresión inmediatamente anterior, la condición (2.2) se traduce en

$$\int_{\Omega_1} \left[\varphi(x) + \psi \circ T(x)\right] d\mu(x) = \int_{\Omega_1} \varphi(x) d\mu(x) + \int_{\Omega_2} \psi(y) d\nu(y),$$

lo que, cancelando los sumandos en φ, se convierte en la condición

$$\int_{\Omega_1} (\psi \circ T) d\mu = \int_{\Omega_2} \psi \, d\nu, \tag{2.4}$$

que debe verificarse para toda función $\psi \in L^1(d\nu)$. Esto es, para cada función $\psi \in L^1(d\nu)$, la función medible $\psi \circ T$ debe pertenecer a $L^1(d\mu)$ y los valores de ambas integrales en (2.4) deben coincidir. Esto mismo en términos de conjuntos medibles puede reescribirse como:

$$\forall\, B \in \sigma_2, \quad \nu(B) = \mu\left(T^{-1}(B)\right). \tag{2.5}$$

Cuando se verifiquen las condiciones (2.4) y (2.5), escribiremos

$$\nu = T\#\mu$$

y diremos que ν es una **medida engendrada** por T (push-forward en inglés) o **medida imagen** de μ por T, o también diremos que T transporta μ a ν.

En estas condiciones podemos introducir la formulación de Monge al problema del trasporte óptimo.

Definición 2.3 (**Problema de Monge**)**.** Dados dos espacios de probabilidad $(\Omega_1, \sigma_1, \mu)$ y $(\Omega_2, \sigma_2, \nu)$, el problema de Monge del transporte óptimo propone encontrar el mínimo valor de

$$I(T) := \int_{\Omega_1} c(x, T(x)) d\mu(x)$$

en el conjunto de aplicaciones medibles $T : \Omega_1 \to \Omega_2$ que verifican $T\#\mu = \nu$.

En lo que sigue nos referiremos por el nombre de "problema de Monge-Kantorovich" a cualquiera de los problemas de minimación de Monge o Kantorovich que han sido comentados. Así, una vez que tenemos el problema expuesto una pregunta natural que surge es la existencia de minimizantes.

2.2. La distancia de Wasserstein

Sean (Ω, σ, P) un espacio de probabilidad y B un espacio de Banach con norma $\|\cdot\|$ y sigma álgebra asociada β_B. Vamos a denotar por $\Pi(B)$ al conjunto de todas las probabilidades definidas en (B, β_B) y además, dado $0 < p < \infty$, representaremos por $\Pi_p(B)$ al subconjunto de aquellas probabilidades P en $\Pi(B)$ que verifican $E\|t\|^p dP < \infty$.

Definición 2.4. Sea $c : B \times B \to \mathbb{R}$ una función continua y no negativa. Si μ, ν son dos probabilidades en (B, β_B) y denotamos $\Pi(\mu, \nu)$ al conjunto de

probabilidades en $(B \times B, \beta_B \otimes \beta_B)$ con marginales μ y ν entonces

$$K_c(\mu,\nu) = \text{ínf}\left\{\int_{B\times B} c(x,y)P(dx,dy) \ : \ P \in \Pi(\mu,\nu)\right\}$$

define el **funcional de Kantorovich**, dada la función c.

Siguiendo el artículo de Bickel et al [2], en este trabajo nos vamos a centrar en el caso especial en el cual el espacio de Banach B es $\mathbb{R}^d$ para $d \geq 1$, y la función c es elegida como $c = c_p$, donde c_p viene definida por $c_p(x,y) = \|x-y\|^p$ con $1 \leq p < \infty$. En este contexto podemos introducir el concepto de distancia de Wasserstein.

Definición 2.5. Sea $1 \leq p < \infty$ y consideramos $\mu, \nu \in \Pi_p(\mathbb{R}^d)$. Se define la ***p*-distancia de Wasserstein** entre μ y ν en $\Pi_p(\mu,\nu)$, como la distancia en $\Pi_p(\mu,\nu)$ dada por $d_p(\mu,\nu) = \big(K_{c_p}(\mu,\nu)\big)^{1/p}$, esto es,

$$\begin{aligned} d_p(\mu,\nu) &= \left(\text{ínf}\int_{\mathbb{R}^d\times\mathbb{R}^d} c_p(x,y)d\pi(x,y) \ : \ \pi \in \Pi(\mu,\nu)\right)^{1/p} \\ &= \left(\text{ínf}\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\pi(x,y) \ : \ \pi \in \Pi(\mu,\nu)\right)^{1/p}. \end{aligned}$$

Razonando de manera totalmente análoga al caso de la versión probabilista del problema de Kantorovich, es claro que también podemos escribir

$$d_p(\mu,\nu) = \text{ínf}\{(E\|U-V\|^p)\}^{1/p} \tag{2.6}$$

donde el ínfimo se toma sobre los pares de variables aleatorias U y V con valores en $\mathbb{R}^d$, definidas en cierto espacio de probabilidad común y tales que $\mathscr{L}(U) = \mu$ y $\mathscr{L}(V) = \nu$.

Definición 2.6. Sean U y V vectores aleatorios definidos en un espacio de probabilidad común que verifican $\mathscr{L}(U) = \mu$ y $\mathscr{L}(V) = \nu$. En estas condiciones el par (U,V) se dice que es un **emparejamiento** de μ y ν.

Observación 3. Abusando de la notación también diremos que la ley de probabilidad conjunta del par (U,V), $\pi \in \Pi(\mu,\nu)$, es un emparejamiento de μ y ν.

Una vez conocido el concepto de emparejamiento, nos interesamos en aquellos que se denominan óptimos.

Definición 2.7. En el contexto expuesto, se dice que (X, Y) es un **emparejamiento óptimo** para la p-distancia de Wasserstein entre μ y ν si $\mathscr{L}(X) = \mu$, $\mathscr{L}(Y) = \nu$ y se verifica que $d_p(\mu, \nu) = (E(\|X - Y\|^p))^{1/p}$.

Si pensamos en el problema de Monge, la p-distancia de Wasserstein entre dos probabilidades μ y ν tomaría la forma:

$$d_p(\mu, \nu) = \left(\text{ínf} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - T(x)\|^p d\mu(x) \ : \ T\#\mu = \nu \right\} \right)^{1/p}.$$

Luego, pasando a trabajar con variables aleatorias, un emparejamiento (óptimo) para este problema debe verificar que una de las variables sea función de la otra, que es lo que denominamos emparejamiento (óptimo) determinista.

Definición 2.8. Se dice que un emparejamiento (X, Y) es **determinista** si existe una aplicación medible $T : \Omega_1 \to \Omega_2$ de manera que $T(X) = Y$.

Observación 4. Dadas dos variables aleatorias X e Y definidas sobre un mismo espacio de probabilidad, cuando escribamos $d_p(X, Y)$ nos estaremos refiriendo a la p-distancia de Wasserstein entre las leyes de probabilidad que siguen, esto es, si $\mathscr{L}(X) = \mu$ y $\mathscr{L}(Y) = \nu$ entonces $d_p(X, Y) = d_p(\mu, \nu)$.

Antes de probar que efectivamente la distancia de Wasserstein es una métrica (esto es, una distancia) vamos a demostrar que se alcanza el mínimo en (2.6).

Proposición 2.9. *En el contexto anterior, el ínfimo en* (2.6) *es alcanzado.*

Demostración. Ya sabemos que el ínfimo se toma en un conjunto no vacío. También sabemos que la p-distancia de Wasserstein la podemos definir de forma analítica como

$$d_p(\mu, \nu) = \inf_{\pi \in \Pi_p(\mu, \nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) \right\}^{1/p}.$$

Dados $x, y \in \mathbb{R}^d$, es conocida la desigualdad $\|x + y\|^p \leq k_p(\|x\|^p + \|y\|^p)$ donde k_p toma el valor 1 si $p \leq 1$ o el valor 2^{p-1} si $p \geq 1$, luego

$$\begin{aligned} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) &\leq k_p \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x\|^p d\pi(x, y) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y\|^p d\pi(x, y) \right) \\ &= k_p \left(\int_{\mathbb{R}^d} \|x\|^p d\mu(x) + \int_{\mathbb{R}^d} \|y\|^p d\nu(y) \right) < \infty, \end{aligned}$$

puesto que μ y ν pertenecen a $\Pi_p(\mathbb{R}^d)$ y por tanto tienen momentos de orden p finitos.

Esto nos asegura que el ínfimo es finito y, denotando Π_1 y Π_2 a las proyecciones sobre la primera y segunda componentes respectivamente, sabemos que existe una sucesión $\{\pi_n\}_{n=1}^{\infty}$ verificando $\pi_n \circ \Pi_1^{-1} = \mu$ y $\pi_n \circ \Pi_2^{-1} = \nu$ y además

$$\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi_n(x, y) \right)^{1/p} \downarrow d_p(\mu, \nu).$$

Denotando por U_n y V_n a las variables aleatorias resultantes de considerar Π_1 y Π_2 sobre el espacio $(\mathbb{R}^d \times \mathbb{R}^d, \pi_n)$ tenemos que

$$\mathscr{L}(U_n) = \mu, \quad \mathscr{L}(V_n) = \nu \quad \text{y} \quad (E\|U_n - V_n\|^p)^{1/p} \downarrow d_p(\mu, \nu).$$

Ahora bien, como:

$$\begin{aligned} \mathscr{L}(U_n) = \mu, \quad \forall n \in \mathbb{N} \quad &\Longrightarrow \quad \{U_n\}_{n=1}^{\infty} \text{ es ajustada,} \\ \mathscr{L}(V_n) = \nu, \quad \forall n \in \mathbb{N} \quad &\Longrightarrow \quad \{V_n\}_{n=1}^{\infty} \text{ es ajustada,} \end{aligned}$$

en consecuencia, la sucesión $\{(U_n, V_n)\}_{n=1}^{\infty}$ es ajustada y, por tanto, la sucesión $\{\mathscr{L}(U_n, V_n)\}_{n=1}^{\infty} = \{\pi_n\}_{n=1}^{\infty}$ es ajustada. Del Teorema de compacidad de Helly se deduce la existencia de una subsucesión $\{\pi_{n_k}\}_{k=1}^{\infty}$ y una probabilidad π en $\mathbb{R}^d \times \mathbb{R}^d$ tales que $\pi_{n_k} \xrightarrow{d} \pi$. Denotando por U y V a las variables obtenidas al considerar las proyecciones Π_1 y Π_2 sobre el espacio $(\mathbb{R}^d \times \mathbb{R}^d, \pi)$, de la continuidad de las funciones coordenadas se sigue que:

$$\begin{aligned} \mu = \pi_n \circ \Pi_1^{-1} \xrightarrow{d} \pi \circ \Pi_1^{-1} \quad &\Longrightarrow \quad \mathscr{L}(U) = \pi \circ \Pi_1^{-1} = \mu, \\ \nu = \pi_n \circ \Pi_2^{-1} \xrightarrow{d} \pi \circ \Pi_2^{-1} \quad &\Longrightarrow \quad \mathscr{L}(V) = \pi \circ \Pi_2^{-1} = \nu. \end{aligned}$$

Nuevamente, por la continuidad de la función $\|\cdot\|^p : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ dada por $(x, y) \longmapsto \|x - y\|^p$ se tiene que

$$\|U_{n_k} - V_{n_k}\|^p \xrightarrow{d} \|U - V\|^p.$$

Ahora bien, del Lema de Fatou se deduce que

$$\begin{aligned} E\|U - V\|^p &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \|U - V\|^p d\pi = \int_{\mathbb{R}^d \times \mathbb{R}^d} \liminf \|U_{n_k} - V_{n_k}\|^p d\pi \\ &\leq \liminf \int_{\mathbb{R}^d \times \mathbb{R}^d} \|U_{n_k} - V_{n_k}\|^p d\pi = (d_p(\mu, \nu))^p, \end{aligned}$$

luego elevando a $1/p$ ambos lados de la desigualdad se obtiene que

$$(E\|U - V\|^p)^{1/p} \leq d_p(\mu, \nu).$$

Por último, de la definición de p-distancia de Wasserstein es evidente la desigualdad contraria, esto es

$$d_p(\mu, \nu) \leq (E\|U - V\|^p)^{1/p},$$

en consecuencia, se tiene que efectivamente

$$d_p(\mu, \nu) = (E\|U - V\|^p)^{1/p}.$$

□

Para probar que la distancia de Wasserstein es una métrica vamos a requerir un resultado auxiliar, denominado Lema del Pegado que, siguiendo el desarrollo del libro de Santambrogio [18], puede demostrarse recurriendo a la desintegración de medidas. Realizamos una breve introducción a este concepto y exponemos el resultado de interés.

Definición 2.10. Consideramos un espacio de medida (Ω, μ) y una aplicación $f : \Omega \to \Lambda$ tomando valores en un espacio topológico Λ. Decimos que una familia $(\mu_y)_{y\in\Lambda}$ es una **desintegración de $\boldsymbol{\mu}$ con respecto de $\boldsymbol{f}$** si cada μ_y es una medida de probabilidad concentrada en $f^{-1}(\{y\})$ y para cada función test $\phi \in \mathscr{C}(\Omega)$ la aplicación $y \mapsto \int_\Omega \phi d\mu_y$ es medible Borel y verifica

$$\int_\Omega \phi d\mu = \int_\Lambda d\nu(y) \int_\Omega \phi d\mu_y,$$

donde $\nu = f\#\mu$.

Observación 5. En el caso particular en el que Ω es un espacio producto, esto es $\Omega = \Omega_1 \times \Omega_2$, y cuando f es la proyección sobre Ω_1, identificaremos las medidas μ_y, que formalmente están definidas en $\Omega_1 \times \Omega_2$ concentradas en $\{y\} \times \Omega_2$, con medidas en Ω_2 con lo que tenemos

$$\int_{\Omega_1\times\Omega_2} \phi(y, z)d\mu(y, z) = \int_{\Omega_1} d\nu(y) \int_{\Omega_2} \phi(y, z)d\mu_y(z).$$

Observación 6. La desintegración de una medida μ se corresponde exactamente con la ley de probabilidad condicionada. En probabilidad, se habla a menudo de la ley condicionada de una variable X conocida $Y = y$. Esto significa que la probabilidad $\mathbb{P}$ en el espacio de probabilidad Ω es desintegrada respecto de la aplicación $Y : \Omega \to E$ en probabilidades $\mathbb{P}_y$ y consideramos la ley de X bajo $\mathbb{P}_y$.

La existencia y unicidad de la desintegración de medidas depende de las hipótesis sobre los espacios que se trabajan, pero para $\Omega = \mathbb{R}^d$ es cierto.

Lema 2.11 (del pegado)**.** *Sean (Ω_1, μ), (Ω_2, ρ) y (Ω_3, ν) espacios de medida. Dadas medidas $\alpha \in \Pi(\mu, \rho)$ y $\beta \in \Pi(\rho, \nu)$, existe $\sigma \in P(\Omega_1 \times \Omega_2 \times \Omega_3)$ tal que $\Pi_{12}\#\sigma = \alpha$ y $\Pi_{23}\#\sigma = \beta$, donde Π_{12} y Π_{23} denotan las proyecciones sobre las dos primeras y las dos segundas variables, respectivamente.*

Demostración. En primer lugar, consideramos la desintegración de α respecto de la proyección Π_2. De esta manera conseguimos una familia de medidas $\alpha_y \in \Pi(\Omega_1)$ (identificándolas como medidas sobre Ω_1, en vez de sobre $\Omega_1 \times \{y\} \subset \Omega_1 \times \Omega_2$). Estas medidas estaban definidas por:

$$\int_{\Omega_1 \times \Omega_2} \phi(x, y) d\alpha(x, y) = \int_{\Omega_2} d\rho(y) \int_{\Omega_1} \phi(x, y) d\alpha_y(x)$$

para toda función medible ϕ de dos variables.

De la misma manera, tenemos una familia $\beta_y \in \Pi(\Omega_3)$ de manera que para cada ϕ tenemos:

$$\int_{\Omega_2 \times \Omega_3} \phi(y, z) d\beta(y, z) = \int_{\Omega_2} d\rho(y) \int_{\Omega_3} \phi(y, z) d\beta_y(z).$$

Para cada $y \in \Omega_2$, consideramos $\alpha_y \times \beta_y$, que es una medida sobre $\Omega_1 \times \Omega_3$. Definimos $\sigma \in P(\Omega_1 \times \Omega_2 \times \Omega_3)$ como

$$\int_{\Omega_1 \times \Omega_2 \times \Omega_3} \phi(x, y, z) d\sigma(x, y, z) := \int_{\Omega_2} d\rho(y) \int_{\Omega_1 \times \Omega_3} \phi(x, y, z) d(\alpha_y \times \beta_y)(x, z).$$

Para cada ϕ que depende sólo de x e y, se tiene:

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2 \times \Omega_3} \phi(x, y) d\sigma &= \int_{\Omega_2} d\rho(y) \int_{\Omega_1 \times \Omega_3} \phi(x, y) d(\alpha_y \times \beta_y)(x, z) \\ &= \int_{\Omega_2} d\rho(y) \int_{\Omega_1} \phi(x, y) d\alpha_y(x) = \int_{\Omega_1 \times \Omega_2} \phi d\alpha(x, y); \end{aligned}$$

lo que prueba que, en efecto, $(\Pi_{12})\#\sigma = \alpha$. Análogamente, considerando funciones test ϕ que sólo dependan de las variables y y z, se prueba que $(\Pi_{23})\#\sigma = \beta$. □

Proposición 2.12. *La p-distancia de Wasserstein definida en el espacio de probabilidades $\Pi_p(\mathbb{R}^d)$ es una métrica.*

Demostración.

(i) Sean $\mu, \nu \in \Pi_p(\mathbb{R}^d)$, veamos que si $d_p(\mu, \nu) = 0$ entonces $\mu = \nu$.

Si $d_p(\mu, \nu) = 0$, entonces existe una probablidad $\gamma \in \Pi(\mu, \nu)$ de manera que

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

Esto significa que γ está concentrada en el conjunto $\{x = y\}$, luego para cualquier función test $\phi \in \mathscr{C}(\mathbb{R}^d)$ se tiene que

$$\int \phi d\mu = \int \phi(x) d\gamma = \int \phi(y) d\gamma = \int \phi d\nu$$

lo cual implica que $\mu = \nu$.

(ii) Sean $\mu, \nu \in \Pi_p(\mathbb{R}^d)$, veamos que $d_p(\mu, \nu) \geq 0$.

Es directo de la definición de p-distancia de Wasserstein, pues se define como la potencia de un valor esperado que es siempre una cantidad positiva o igual a cero, luego efectivamente $d_p(\mu, \nu) \geq 0$.

(iii) Sean $\mu, \nu \in \Pi_p(\mathbb{R}^d)$, es claro que $d_p(\mu, \nu) = d_p(\nu, \mu)$ de la simetría de la definición.

(iv) Sean $\mu, \rho, \nu \in \Pi_p(\mathbb{R}^d)$, $\alpha \in \Pi(\mu, \rho)$ y $\beta \in \Pi(\rho, \nu)$, de manera que α y β son óptimos en las respectivas p-distancias de Wasserstein.

Del Lema del pegado existe una medida $\sigma \in P(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ tal que $\Pi_{12}\#\sigma = \alpha$ y $\Pi_{23}\#\sigma = \beta$, donde Π_{12} y Π_{23} denotan las proyecciones sobre las dos primeras y las dos segundas variables, respectivamente.

Consideramos $\gamma := \Pi_{13}\#\sigma$ y observamos que:

$$\Pi_1\#\gamma = \Pi_1\#\sigma = \Pi_1\#\alpha = \mu$$

y análogamente

$$\Pi_2 \# \gamma = \Pi_2 \# \sigma = \Pi_2 \# \beta = \nu;$$

lo que significa que $\gamma \in \Pi(\mu, \nu)$. De esta manera, haciendo uso de la desigualdad de Minkowski, tenemos que

$$\begin{aligned} d_p(\mu,\nu) &\leq \left(\int \|x-z\|^p d\gamma\right)^{1/p} = \left(\int \|x-z\|^p d\sigma\right)^{1/p} \\ &= \|x-z\|_{L^p(\sigma)} \leq \|x-y\|_{L^p(\sigma)} + \|y-z\|_{L^p(\sigma)} \\ &= \left(\int \|x-y\|^p d\sigma\right)^{1/p} + \left(\int \|y-z\|^p d\sigma\right)^{1/p} \\ &= \left(\int \|x-y\|^p d\alpha\right)^{1/p} + \left(\int \|y-z\|^p d\beta\right)^{1/p} \\ &= d_p(\mu,\rho) + d_p(\rho,\nu). \end{aligned}$$

□

2.2.1. Caso de la recta real y la función cuantil

Al trabajar sobre la recta real la p-distancia de Wasserstein entre dos probabilidades se puede escribir de manera sencilla en términos de las funciones cuantiles asociadas a cada probabilidad.

Este resultado puede probarse de varias maneras, destacamos en particular las dadas en Major [15] y en Cambanis et al [7]. En el primero se realiza en un primer lugar la prueba para el caso discreto, suponiendo una propiedad que determina completamente la distribución conjunta de dos variables y que verifica la función cuantil, para luego tratar el caso general haciendo uso de dos sucesiones de variables discretas que convergen casi seguro a las variables que son óptimas y acabar utilizando el Teorema de Skorohod para poder trabajar con funciones cuantiles. Por otro lado, en el segundo se realiza una demostración alternativa, que es la que presentamos en este trabajo, basada en el estudio de desigualdades para la expresión $E(k(X,Y))$ en el caso de ser k una función "casi-monótona".

La desigualdades para el caso de variables aleatorias son inmediatas. Para poder presentar el estudio en cuestión necesitamos introducir el concepto de variable aleatoria estocásticamente más pequeña.

Definición 2.13. Sean X y X' dos variables aleatorias, se dice que X es **estocásticamente más pequeña** que X', y escribimos $X \subset X'$, si para todo $x \in \mathbb{R}$ se verifica

$$P(X < x) \geq P(X' < x).$$

Sean X y X' variables aleatorias tales que X es estocásticamente más pequeña que X'. Si consideramos una función $k : \mathbb{R} \to \mathbb{R}$ monótona creciente, esto es $k(x) \leq k(y)$ para todo $x \leq y$, sabemos que $k(X)$ y $k(X')$ vuelven a ser variables aleatorias. Además es evidente que $k(X)$ es estocásticamente más pequeña que $k(X')$ y, en consecuencia, $E(k(X)) \leq E(k(X'))$.

Nuestro objetivo ahora es conseguir un resultado análogo para el caso de dos dimensiones. El problema al trabajar con vectores aleatorios bidimensionales es que los conceptos de ser estocásticamente más pequeño y de función casi-monóntona se complican y es más laborioso realizar un estudio de las desigualdades para la expresión $E(k(X, Y))$.

Definición 2.14. Sean (X, Y) y (X', Y') dos pares de vectores aleatorios, se dice que (X, Y) es **estocásticamente más pequeño** que (X', Y'), y escribimos $(X, Y) \subset (X', Y')$ si para todos $x, y \in \mathbb{R}$ se verifica

$$P(X < x, Y < y) \geq P(X' < x, Y' < y).$$

Con el objetivo expuesto, consideramos variables aleatorias X e Y en un espacio de probabilidad (Ω, σ, P), siendo $F(x)$ y $G(y)$ sus respectivas funciones de distribución y $H(x, y)$ la función de distribución conjunta del par (X, Y).

Definición 2.15. Una función $k(x, y)$ se denomina **casi-monótona** si para todo $x \leq x'$, $y \leq y'$ se verifica

$$\Delta_{(x,x')}^{(y,y')} k = k(x, y) + k(x', y') - k(x, y') - k(x', y) \geq 0.$$

De la misma forma, se dice que la función $k(x, y)$ es **casi-antimonótona** si para todo $x \leq x'$, $y \leq y'$ se verifica $\Delta_{(x,x')}^{(y,y')} k \leq 0$.

Observación 7. Si k es casi-monótona y continua por la derecha entonces determina una única medida μ (σ-finita y no negativa) en los subconjuntos de Borel β^2 en el plano $\mathbb{R}^2$, de manera que para todos $x \leq x'$ e $y \leq y'$ verifica

$$\mu\left((x, x'] \times (y, y']\right) = \Delta_{(x,x')}^{(y,y')} k.$$

Veamos que un cambio en el orden de integración en una integral doble apropiada nos da la expresión deseada para $E(k(X,Y))$.

CASO PARTICULAR

Sea $k(x,y)$ simétrica, continua por la derecha y casi-monótona. Definimos la función $f(x,y,\omega)$ como sigue:

$$f(x,y,\omega) = \begin{cases} & X(\omega) < x,\ y \le Y(\omega) \\ 1 \quad \text{si} & \text{ó} \\ & Y(\omega) < x,\ y \le X(\omega) \\ \\ 0 & \text{en otro caso} \end{cases}$$

Así f es una función medible en $(\mathbb{R}^2 \times \Omega, \beta^2 \otimes \sigma, \mu \times P)$ y como es no negativa, podemos aplicar el Teorema de Fubini que afirma

$$E\int_{\mathbb{R}^2} f d\mu = \int_{\mathbb{R}^2} Ef d\mu.$$

Ahora bien, la integral que aparece en la parte izquierda de la igualdad anterior es

$$\int_{\mathbb{R}^2} f d\mu = k(X,X) + k(Y,Y) - 2k(X,Y);$$

y por otro lado si calculamos el valor esperado que se encuentra en la parte derecha se tiene que

$$\begin{aligned} Ef &= P(X < x \wedge y,\ Y \ge x \vee y) + P(X \ge x \vee y,\ Y < x \wedge y) = \\ &= F(x \wedge y) + G(x \wedge y) - H(x \wedge y, x \vee y) - H(x \vee y, x \wedge y) =: A(x,y). \end{aligned}$$

Y teniendo en cuenta que podemos aplicar el teorema de Fubini como comentamos anteriormente:

$$E\left(k(X,X) + k(Y,Y) - 2k(X,Y)\right) = \int_{\mathbb{R}^2} A(x,y) d\mu(x,y).$$

Luego si $E(k(X,X)), E(k(Y,Y)) < \infty$ tenemos la expresión

$$2E(k(X,Y)) = E(k(X,X)) + E(k(Y,Y)) - \int_{\mathbb{R}^2} A d\mu$$

para determinar $E(k(X,Y))$ que no es la propia definición. Además, si las distribuciones de F y G están fijas (que va a ser nuestro caso porque, en

el contexto de la distancia de Wasserstein trabajamos con probabilidades marginales prefijadas) $E(k(X,Y))$ depende de H solo a través de A y como μ es no negativa, incrementando H resulta que incrementamos $E(k(X,Y))$. Esto es, $E(k(X,Y))$ es monótona creciente respecto de H.

CASO GENERAL

Sea $k(x,y)$ casi monótona y continua por la derecha. En este caso vamos a definir la función $f(x,y,\omega)$ como sigue:

$$f\,(x,y,\omega) = \begin{cases} 1 \quad \text{si} & \begin{array}{c} x_0 < x \leq X(\omega),\ y_0 < y \leq Y(\omega) \\ \text{ó} \\ X(\omega) < x \leq x_0,\ Y(\omega) < y \leq y_0 \end{array} \\ -1 \quad \text{si} & \begin{array}{c} X(\omega) < x < x_0,\ y_0 < y \leq Y(\omega) \\ \text{ó} \\ x_0 < x < X(\omega),\ Y(\omega) < y \leq y_0 \end{array} \\ 0 & \text{en otro caso} \end{cases}$$

donde x_0 e y_0 son puntos fijos apropiados. De nuevo, la función f es $\beta^2 \otimes \sigma$-medible. Si denotamos por f^+ y f^-, respectivamente, a las partes positiva y negativa de f el Teorema de Fubini nos asegura que

$$E\int_{\mathbb{R}^2} f^+ d\mu = \int_{\mathbb{R}^2} Ef^+ d\mu \quad \text{y} \quad E\int_{\mathbb{R}^2} f^- d\mu = \int_{\mathbb{R}^2} Ef^- d\mu.$$

Considerando la función $k_0(x,y) = k(x,y) - k(x,y_0) - k(x_0,y) + k(x_0,y_0)$ obtenemos

$$E(k_0^+(X,Y)) = \int_{\mathbb{R}^2} B^+ d\mu \quad \text{y} \quad E(k_0^-(X,Y)) = \int_{\mathbb{R}^2} B^- d\mu$$

donde k_0^+ y k_0^- son, respectivamente las partes positiva y negativa de la

función k_0 y las funciones B^+ y B^- vienen definidas como sigue :

$$B^+(x,y) = \begin{cases} 1 + H(x,y) - F(x) - G(y) & \text{si} \quad x_0 < x,\ y_0 < y \\ H(x,y) & \text{si} \quad x \le x_0,\ y \le y_0 \\ 0 & \text{en otro caso,} \end{cases}$$

$$B^-(x,y) = \begin{cases} F(x) - H(x,y) & \text{si} \quad x_0 \le x,\ y_0 < y \\ G(y) - H(x,y) & \text{si} \quad x_0 < x,\ y \le y_0 \\ 0 & \text{en otro caso.} \end{cases}$$

Observación 8. Recordamos que si Z^+ y Z^- son la parte positiva y negativa de una variable aleatoria Z, usando la terminología estándar, se dice que EZ existe (aunque sea infinita) si al menos una de las esperanzas EZ^+ o EZ^- es finita, y entonces definimos $EZ = EZ^+ - EZ^-$.

Así, si $Ek_0(X,Y)$ existe, tenemos que $Ek_0(X,Y) = \int_{\mathbb{R}^2} B d\mu$ donde la función $B(x,y)$ es $B = B^+ - B^-$. De la definición de k_0 se tiene que $Ek_0(X,Y)$ existe si $Ek(X,Y)$ existe y $Ek(X,y_0), Ek(x_0,Y) < \infty$ en cuyo caso se tiene que

$$Ek(X,Y) = Ek(X,y_0) + Ek(x_0,Y) - k(x_0,y_0) + \int_{\mathbb{R}^2} B d\mu.$$

Esta expresión obtenida para $Ek(X,Y)$, si F y G son fijas (como es en nuestro caso), depende únicamente de H solo a través de B y, además es monótona creciente respecto de H.

Recogemos en el siguiente resultado lo que se deduce del desarrollo realizado (tanto en el caso particular de que la función $k(x,y)$ sea simétrica, como en el caso más general), sin más que tener en cuenta la definción de vector aleatorio estocásticamente más pequeño y su relación con la función de distribución conjunta del mismo.

Teorema 2.16. *Sean $X \overset{d}{=} X'$ e $Y \overset{d}{=} Y'$ tales que $(X,Y) \subset (X',Y')$. Si $k(x,y)$ es una función casi-monótona y continua por la derecha entonces*

$$Ek(X,Y) \ge Ek(X',Y')$$

cuando los valores de la expresión anterior existen (aunque puedan ser infinitos) y cualquiera de las condiciones siguientes se verifica:

(i) $k(x,y)$ *es simétrica y* $Ek(X,X), Ek(Y,Y) < \infty$.

(ii) $Ek(X,y_0), Ek(x_0,Y) < \infty$ *para algún* x_0, y_0.

El resultado anterior es muy interesante pues, bajo ciertas condiciones, relaciona los valores esperados $Ek(X,Y)$ y $Ek(X',Y')$ dependiendo únicamente de la distribución conjunta que presentan los distintos vectores aleatorios, ya que las distribuciones marginales son fijas. Podemos caracterizar este resultado para el conjunto de distribuciones conjuntas con distribuciones marginales F y G respectivamente y obtendremos el resultado de real interés.

Por lo tanto, vamos a denotar por $\mathcal{H}(F,G)$ al conjunto de todas las funciones de distribución conjuntas $H(x,y)$ con distribuciones marginales $F(x)$ y $G(y)$, y por $E_H k(X,Y)$ al valor esperado cuando H es la función de distribución conjunta del vector (X,Y). A continuación presentamos algunas propiedades del conjunto $\mathcal{H}(F,G)$ conocidas:

- El conjunto $\mathcal{H}(F,G)$ tiene una cota superior e inferior (las denominadas cotas de Fréchet).
- $H(x,y) \in \mathcal{H}(F,G)$ si, y sólo si $H_-(x,y) \leq H(x,y) \leq H_+(x,y)$ para todos x, y donde H_- y H_+ son:

  $$H_-(x,y) = \text{máx}\{F(x) + G(y) - 1, 0\} \;\leftarrow\; \text{cota inferior de } \mathcal{H}(F,G),$$
  $$H_+(x,y) = \text{mín}\{F(x), G(y)\} \;\leftarrow\; \text{cota superior de } \mathcal{H}(F,G).$$

- $\mathcal{H}(F,G)$ es una familia convexa de funciones de distribución.

Sea $\mathcal{H}$ una familia convexa de funciones de distribución en 2 variables. Si $H, H' \in \mathcal{H}$ son tales que $E_H k(X,Y)$ y $E_{H'} k(X,Y)$ existen y son finitas, entonces cada número en el intervalo

$$[\text{mín}\{E_H k(X,Y), E_{H'} k(X,Y)\}, \text{máx}\{E_H k(X,Y), E_{H'} k(X,Y)\}]$$

es igual a $E_{H''} k(X,Y)$ para algún $H'' \in \mathcal{H}$. Si para cada $\alpha \in [0,1]$ consideramos $H_\alpha(x,y) = \alpha H(x,y) + (1-\alpha)H'(x,y)$, entonces $H_\alpha \in \mathcal{H}$ y se tiene que

$$EH_\alpha k(X,Y) = \alpha E_H k(X,Y) + (1-\alpha) E_{H'} k(X,Y).$$

Esto no es válido si $-\infty < E_H k(X,Y) < E_{H'} k(X,Y) = +\infty$ como muestra el ejemplo $\mathcal{H} = \{H_\alpha : \alpha \in [0,1]\}$. El conjunto de los valores $E_H k(X,Y)$ cuando H recorre una familia de funciones de distribución no es necesariamente convexo (puede ser de la forma $I \cup \{\infty\}, \{-\infty\} \cup I, \{-\infty\} \cup I \cup \{\infty\}$).

En el próximo teorema probaremos que bajo ciertas hipótesis que incluyen la casi-monotonía de k se puede asegurar que cuando H recorre $\mathcal{H}(F,G)$ el conjunto de valores $E_H k(X,Y)$ es cerrado, convexo y su supremo e ínfimo son alcanzados. Antes introducimos un lema previo que ayudará a la prueba del resultado principal.

Lema 2.17. *Sean X e Y variables aleatorias con funciones de distribución respevtivas $F(x)$ y $G(y)$, y sea $H(x,y)$ la función de distribución conjunta de (X,Y). Sea $k(x,y)$ una función medible Borel, localmente acotada en el plano.*
Si los valores esperados $E_{H_+}k(X,Y)$ y $E_{H_-}k(X,Y)$ existen y al menos uno de ellos es finito, entonces el intervalo cerrado (posiblemente no acotado) con puntos finales $E_{H_+}k(X,Y)$ y $E_{H_-}h(X,Y)$ pertenece al conjunto de valores de $E_H k(X,Y)$ cuando H recorre $\mathcal{H}(F,G)$.

Demostración. Basta con probar que si

$$-\infty < E_{H_+}k(X,Y) < E_{H_-}k(X,Y) = +\infty$$

existe una sucesión $H_n \in \mathcal{H}(F,G)$ tal que $E_{H_n}k(X,Y) \to \infty$, puesto que el resto de casos son totalmente análogos.

De la definición de H_+ y H_- tenemos que

- bajo H_+: $(X,Y) \stackrel{d}{=} (F^{-1}(U), G^{-1}(U))$,
- bajo H_-: $(X,Y) \stackrel{d}{=} (F^{-1}(U), G^{-1}(1-U))$,

donde U sigue una distribución uniforme en el intervalo $[0,1]$ y, F^{-1} y G^{-1} denotan a las respectivas funciones cuantiles de F y G. Además, por hipótesis,

$$E_{H_-}k(X,Y) = Ek(F^{-1}(U), G^{-1}(1-U)) = \int_0^1 k(F^{-1}(u), G^{-1}(1-u))du = +\infty.$$

Luego para cada $0 \leq \alpha \leq 1/2$ definimos $g_\alpha(u)$ en $[0,1]$ como

$$g_\alpha(u) = \begin{cases} G^{-1}(1-u) & \text{si} \quad \alpha < u < 1-\alpha \\ \\ G^{-1}(u) & \text{si} \quad \begin{array}{l} 0 \leq u \leq \alpha \\ \text{ó} \\ 1-\alpha \leq u \leq 1 \end{array} \end{cases}$$

y sea H_α la función de distribución del par $(F^{-1}(U), g_\alpha(U))$. Para cada x tenemos

$$\begin{aligned} P(g_\alpha(U) < x) &= \ell\{(\alpha, 1-\alpha) \cap (1-G(x), 1]\} + \\ &\quad + \ell\{(0,\alpha] \cup [1-\alpha, 1] \cap [0, G(x))\} = G(x). \end{aligned}$$

Esto prueba que las distribuciones marginales de H_α son respectivamente F y G, luego $H_\alpha \in \mathcal{H}(F,G)$. Además, también tenemos que

$$\begin{aligned} E_{H_\alpha} k(X,Y) = Ek(F^{-1}(U), g_\alpha(U)) &= \int_\alpha^{1-\alpha} k(F^{-1}(u), G^{-1}(1-u))du + \\ &+ \int_0^\alpha k(F^{-1}(u), G^{-1}(u))du \; + \; \int_{1-\alpha}^1 k(F^{-1}(u), G^{-1}(u))du. \end{aligned}$$

Como para $u \in [\alpha, 1-\alpha]$, $F^{-1}(u)$ y $G^{-1}(1-u)$ están acotadas y como k es localmente acotada se tiene que

$$\int_\alpha^{1-\alpha} k(F^{-1}(u), G^{-1}(1-u))du < \infty$$

y

$$\begin{aligned} \lim_{\alpha \downarrow 0} \int_\alpha^{1-\alpha} k(F^{-1}(u), G^{-1}(1-u))du &= \int_0^1 k(F^{-1}(u), G^{-1}(1-u))du \\ &= E_{H_-} k(X,Y) = +\infty. \end{aligned}$$

Por otra parte, como $\int_0^1 |k(F^{-1}(u), G^{-1}(u))|du = E_{H_+}|k(X,Y)| < \infty$ tenemos

$$\lim_{\alpha \downarrow 0} \left(\int_0^\alpha k(F^{-1}(u), G^{-1}(u))du \; + \; \int_{1-\alpha}^1 k(F^{-1}(u), G^{-1}(u))du \right) = 0.$$

Por lo tanto, $\lim_{\alpha\downarrow 0} E_{H_\alpha} k(X,Y) = +\infty$, como queríamos probar. □

Teorema 2.18. *Sean X e Y variables aleatorias con funciones de distribución respectivas $F(x)$ y $G(y)$, y función de distribución conjunta $H(x,y)$. Sea $k(x,y)$ una función casi-monótona y continua por la derecha.*
Si los valores esperados $E_{H_+} k(X,Y)$, $E_{H_-} k(X,Y)$ existen (aunque puedan ser infinitos), entonces el conjunto de los valores de $E_H k(X,Y)$ cuando H recorre $\mathcal{H}(F,G)$ pertenecen al intervalo cerrado $\left[E_{H_-} k(X,Y), E_{H_+} k(X,Y)\right]$ (posiblemente no acotado) cuando alguna de las siguiente condiciones se verifica.

(a) $k(x,y)$ *es simétrica y (i) se verifica (en este caso,* $-\infty \leq E_{H_-}k(X,Y) \leq E_{H_+}k(X,Y) < +\infty$*).*

(b) Para algún x_0 *e* y_0*, (ii) se verifica y al menos unos de los valores* $E_{H_-}k(X,Y)$ *o* $E_{H_+}k(X,Y)$ *es finito.*

Demostración. (a) Por las expresiones obtenidas para el valor esperado $E_H k(X,Y)$, que sabíamos que eran monótonas crecientes respecto de la función de distibución conjunta $H(x,y)$ considerada, se deduce que para todo $H \in \mathcal{H}(F,G)$ el valor esperado $E_H k(X,Y)$ existe y satisface

$$-\infty \leq E_{H_-}k(X,Y) \leq E_H k(X,Y) \leq E_{H_+}k(X,Y) < +\infty.$$

- Si $E_{H_+}k(X,Y) = -\infty$ entonces $E_H k(X,Y) = -\infty$ para todo $H \in \mathcal{H}(F,G)$.
- Si $-\infty < E_{H_+}k(X,Y) < \infty$ el resultado es trivial y se deduce del lema previo cuando $E_{H_-}k(X,Y) = -\infty$.

(b) Supongamos $E_{H_+}k(X,Y)$ es finito entonces hay dos posibilidades:

- Si $E_{H_-}k(X,Y) = E_{H_+}k(X,Y)$ entonces $E_H k(X,Y) = E_{H_+}k(X,Y)$ para todo $H \in \mathcal{H}(F,G)$.
- Si $-\infty < E_{H_+}k(X,Y) < \infty$ el resultado es trivial y se deduce del lema previo cuando $E_{H_-}k(X,Y) = -\infty$.

En el caso de suponer $E_{H_-}k(X,Y)$ finito hay dos casos totalmente análogos:

- Si $E_{H_+}k(X,Y) = E_{H_-}k(X,Y)$ entonces $E_H k(X,Y) = E_{H_-}k(X,Y)$ para todo $H \in \mathcal{H}(F,G)$.
- Si $E_{H_+}k(X,Y) < \infty$ el resultado es trivial y se deduce del lema previo cuando $E_{H_+}k(X,Y) = +\infty$.

□

Del teorema previo se deduce, para el caso en el que k es una función casi-monótona, que el ínfimo y el supremo de $E_H k(X,Y)$ para $H \in \mathcal{H}(F,G)$ se alcanzan en H_- y H_+ respectivamente y además, en particular, se tiene que

$$E_{H_-}k(X,Y) = \int_0^1 k(F^{-1}(u), G^{-1}(u))du$$

y

$$E_{H_+}k(X,Y) = \int_0^1 k(F^{-1}(u), G^{-1}(1-u))du.$$

Análogamente, si k es una función casi-antimonótona, el ínfimo y el supremo de $E_H k(X,Y)$ para $H \in \mathcal{H}(F,G)$ se alcanzan en H_+ y H_- respectivamente y, además en particular se tiene que

$$E_{H_-}k(X,Y) = \int_0^1 k(F^{-1}(u), G^{-1}(1-u))du$$

y

$$E_{H_+}k(X,Y) = \int_0^1 k(F^{-1}(u), G^{-1}(u))du.$$

Tomando $k(x,y) = \|x-y\|^p$ que se trata de una función simétrica, casi-antimonótona y continua estamos en condiciones de utilizar el resultado previo que nos asegura que

$$E_{H_+}\|X-Y\|^p = \int_0^1 \|F^{-1}(u)-G^{-1}(u)\|^p du = \text{ínf}\left\{ E\|X-Y\|^p \ : \ \begin{array}{c} \mathscr{L}(X)=F \\ \mathscr{L}(Y)=G \end{array} \right\}$$

$$\Longrightarrow \ d_p(X,Y) = \left(\int_0^1 \|F^{-1}(u) - G^{-1}(u)\|^p du \right)^{1/p}. \tag{2.7}$$

En este contexto, si trabajando sobre la recta real consideramos $p=1$ se deduce inmediatamente el siguiente resultado que nos permite expresar la distancia de Wasserstein en términos de las funciones de distribución, teniendo únicamente en cuenta que el área encerrada por la función de distribución en toda la recta real es el mismo que el encerrado por la función cuantil en el intervalo $[0,1]$.

Corolario 2.19. *Sea $B = \mathbb{R}$ y consideramos $p=1$ y μ, ν probabilidades en $\mathbb{R}$ que vienen dadas por funciones de distribución F y G respectivamente. Entonces considerando $\|x\| = |x|$ se tiene*

$$d_1(\mu,\nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt = \int_{-\infty}^{\infty} |F(x) - G(x)|dx.$$

Demostración. La primera igualdad es consecuencia de (2.7) aplicada al caso particular de la recta real y $p=1$.

Ahora bien, para la segunda igualdad basta observar que el área encerrada por las funciones de distribución en $\mathbb{R}$ es el mismo que el encerrado por las funciones cuantiles en el intervalo $[0, 1]$.

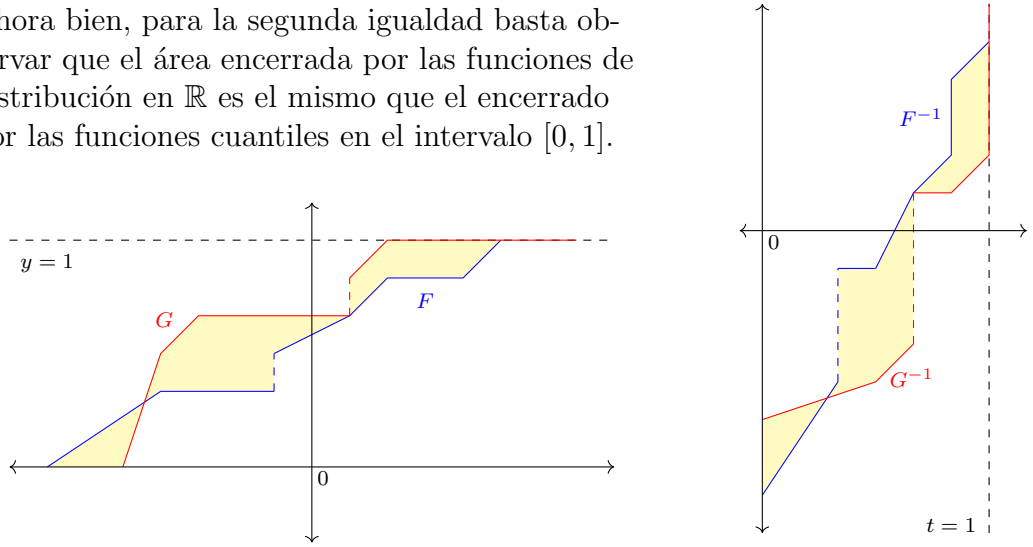


Figura 2.1: Funciones de distribución (izquierda) y cuantiles (derecha).

□

2.2.2. Algunas propiedades básicas

Los resultados que se recogen en esta subsección se pueden consultar en el apéndice del trabajo de Bickel et al [2], clásica referencia que incluye una recopilación muy variada de propiedades de la p-distancia de Wasserstein.

Proposición 2.20. *Sean $\alpha_n, \alpha \in \Pi_p(\mathbb{R}^d)$. Entonces, las siguientes afirmaciones son equivalentes:*

(1) $d_p(\alpha_n, \alpha) \xrightarrow{n\to\infty} 0$.

(2) $\alpha_n \to \alpha$ *débilmente y* $\int \|x\|^p d\alpha_n(x) \to \int \|x\|^p d\alpha(x)$.

(3) $\alpha_n \to \alpha$ *débilmente y la función* $\|x\|^p$ *es uniformemente integrable respecto de* $\{\alpha_n\}_{n=1}^{\infty}$.

Demostración.

(1) $\Longrightarrow$ (2) Supongamos que $d_p(\alpha_n, \alpha) \xrightarrow{n\to\infty} 0$. Consideramos variables aleatorias X_n y X con leyes de probabilidad $\mathscr{L}(X_n) = \alpha_n$ y $\mathscr{L}(X) = \alpha$, y verificando $d_p(\alpha_n, \alpha) = E\left(\|X_n - X\|^p\right)^{1/p}$. Entonces haciendo uso de la segunda

desigualdad trinagular observamos que

$$\left(\int \|x\|^p d\alpha_n(x)\right)^{1/p} - \left(\int \|x\|^p d\alpha(x)\right)^{1/p} = (E\|X_n\|^p)^{1/p} - (E\|X\|^p)^{1/p}$$
$$\leq (E\|X_n - X\|^p)^{1/p} \xrightarrow{n\to\infty} 0.$$

De lo anterior, sin más que despejar se deduce que

$$\left(\int \|x\|^p d\alpha_n(x)\right)^{1/p} \xrightarrow{n\to\infty} \left(\int \|x\|^p d\alpha(x)\right)^{1/p},$$

y de la continuidad de la función $\phi(x) = x^p$ se sigue que

$$\int \|x\|^p d\alpha_n(x) \xrightarrow{n\to\infty} \int \|x\|^p d\alpha(x),$$

o equivalentemente,

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

A continuación, vamos a probar que la sucesión $\{X_n\}_{n=1}^{\infty}$ es ajustada. Sea $\varepsilon > 0$. Como $E\|X_n\|^p \xrightarrow{n\to\infty} E\|X\|^p$ existe un número natural n_0 tal que $E\|X_n\|^p \leq 2E\|X\|^p$ para todo $n \geq n_0$. Así, aplicando la desigualdad de Markov tenemos que

$$P(\|X_n\| > a) \leq \frac{E\|X_n\|^p}{a^p} \leq \frac{2E\|X\|^p}{a^p} \xrightarrow{a\to\infty} 0.$$

Y, por tanto, para a suficientemente grande podemos conseguir $P(\|X_n\| > a) < \varepsilon$ y, en consecuencia, $\{X_n\}_{n=1}^{\infty}$ es una sucesión ajustada.

Ahora bien, observamos que si f es una función Lipschitziana, esto es, si existe una constante K tal que $\|f(x) - f(y)\| \leq K\|x - y\|$ entonces

$$\left\|\int f(x)d\alpha_n(x) - \int f(x)d\alpha(x)\right\| = \|E(f(X_n)) - E(f(X))\|$$
$$= \|E(f(X_n) - f(X))\| \leq E\,\|f(X_n) - f(X)\| \leq KE\|X_n - X\|$$
$$\overset{(*)}{\leq} K\,(E\|X_n - X\|^p)^{1/p} = Kd_p(\alpha_n, \alpha) \xrightarrow{n\to\infty} 0,$$

donde en $(*)$ se ha utilizado la desigualdad de Jensen, lo que implica que

$$\int f(x)d\alpha_n(x) \xrightarrow{n\to\infty} \int f(x)d\alpha(x).$$

En consecuencia, como las funciones Lipschitzianas son una clase separante, tenemos asegurada la convergencia débil de la sucesión $\{\alpha_n\}_{n=1}^{\infty}$ hacia α sin más que recordar el siguiente resultado:

"Sea ε una clase separante y $\{F_n\}_{n=1}^{\infty}$ una sucesión ajustada. Si F es una función de distribución tal que $\int f dF_n \xrightarrow{n\to\infty} \int f dF$ para toda función $f \in \varepsilon$, entonces $F_n \xrightarrow{d} F$ ".

(2) $\Longrightarrow$ (3) Del teorema de representación de Skorohod existen un espacio probabilístico (Ω, σ, P) y variables aleatorias U_n y U de manera que

$$\mathscr{L}(U_n) = \alpha_n, \quad \mathscr{L}(U) = \alpha \quad \text{y} \quad U_n \xrightarrow{c.s.} U.$$

Ahora bien, sabemos que la convergencia casi seguro implica la convergencia en probabilidad, luego $U_n \xrightarrow{p.} U$. Además también sabemos que el hecho de que la función $\|x\|^p$ sea uniformemente integrable respecto de $\{\alpha_n\}_{n=1}^{\infty}$ es equivalente a que la sucesión $\{\|U_n\|^p\}_{n=1}^{\infty}$ sea uniformemente integrable. Por último, como $U_n = \alpha_n$ y $\mathscr{L}(U) = \alpha$ sabemos que

$$E\|U_n\|^p = \int \|x\|^p d\alpha_n(x) \quad \text{y} \quad E\|U\|^p = \int \|x\|^p d\alpha(x).$$

Ahora bien, por hipótesis tenemos que

$$\int \|x\|^p d\alpha_n(x) \to \int \|x\|^p d\alpha(x), \quad \text{o equivalentemente,} \quad E\|U_n\|^p \to E\|U\|^p.$$

Trabajando en espacios L^p, recordamos que son equivalentes:

(I) $U_n \xrightarrow{L_p(\Omega)} U$.

(II) $U_n \xrightarrow{p.} U$ y $E\|U_n\|^p \to E\|U\|^p$.

(III) $U_n \xrightarrow{p.} U$ y $\{\|U_n\|^p\}_{n=1}^{\infty}$ es uniformemente integrable.

En estas condiciones tenemos que $U_n \xrightarrow{p.} U$ y $E\|U_n\|^p \to E\|U\|^p$ que es exactamente la afirmación (II). Como (II) es equivalente a (III), tenemos asegurado que $U_n \xrightarrow{p.} U$ y $\{\|U_n\|^p\}_{n=1}^{\infty}$ es uniformemente integrable y ya hemos comentado que esto último es equivalente a que la función $\|x\|^p$ sea uniformemente integrable respecto de $\{\alpha_n\}_{n=1}^{\infty}$ que es lo que queríamos probar.

(3) $\Longrightarrow$ (2) Suponemos que $\alpha_n \to \alpha$ débilmente y la función $\|x\|^p$ es uniformemente integrable respecto de $\{\alpha_n\}_{n=1}^{\infty}$. Recurriendo nuevamente al teorema de representación de Skorohod tenemos que $U_n \xrightarrow{p.} U$ y $\{\|U_n\|^p\}_{n=1}^{\infty}$ uniformemente integrable lo cual es equivalente a $U_n \xrightarrow{p.} U$ y $E\|U_n\|^p \to E\|U\|^p$ (es la implicación (III) $\Longrightarrow$ (II)).

Luego $\alpha_n \to \alpha$ débilimente y $E\|U_n\|^p \to E\|U\|^p$ lo cual es equivalente a $\int \|x\|^p d\alpha_n(x) \to \int \|x\|^p d\alpha(x)$, con lo que se obtiene (2).

(2) $\Longrightarrow$ (1)

Suponemos que $\alpha_n \to \alpha$ débilmente y $\int \|x\|^p d\alpha_n(x) \to \int \|x\|^p d\alpha(x)$. Razonando como antes tenemos $U_n \xrightarrow{c.s.} U$ y $E\|U_n\|^p \to E\|U\|^p$, luego la equivalencia (I) $\Longleftrightarrow$ (II) nos asegura que $U_n \xrightarrow{L_p} U$, esto es, $(E\|U_n - U\|^p)^{1/p} \xrightarrow{n\to\infty} 0$. En consecuencia, como

$$d_p(\alpha_n, \alpha) \leq (E\|U_n - U\|^p)^{1/p} \xrightarrow{n\to\infty} 0,$$

se deduce que $d_p(\alpha_n, \alpha) \xrightarrow{n\to\infty} 0$. □

Proposición 2.21. *Sea $1 \leq p < \infty$, entonces se verifican las siguientes propiedades para la p-distancia de Wasserstein:*

(1) $d_p(aU, aV) = |a| d_p(U, V)$ para todo escalar $a \in \mathbb{R}$.

(2) $d_p(LU, LV) \leq \|L\| d_p(U, V)$ para toda aplicación lineal $L : \mathbb{R} \to \mathbb{R}$.

Demostración.

(1) El caso $a = 0$ es evidente, por lo tanto, tratamos el caso $a \neq 0$. Supongamos que $\mathscr{L}(U) = \mu$, $\mathscr{L}(V) = \nu$ y $\pi \in \Pi(\mu, \nu)$ entonces

$$\begin{aligned} d_p(U, V) &= \text{ínf} \left\{ \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi \right)^{1/p} : \pi \in \Pi(\mu, \nu) \right\} \\ &= \text{ínf} \left\{ \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\Pi_1(x, y) - \Pi_2(x, y)\|^p d\pi \right)^{1/p} : \pi \in \Pi(\mu, \nu) \right\}. \end{aligned}$$

Esto es, bajo cualquier probabilidad $\pi \in \Pi(\mu, \nu)$ las proyecciones Π_1 y Π_2 verifican que

$$\mathscr{L}(\Pi_1) = \mu = \mathscr{L}(U) \quad \text{y} \quad \mathscr{L}(\Pi_2) = \nu = \mathscr{L}(V).$$

Entonces, si $a \neq 0$ también se tiene que

$$\mathscr{L}(a\Pi_1) = \mathscr{L}(aU) \quad \text{y} \quad \mathscr{L}(a\Pi_2) = \mathscr{L}(aV).$$

De esta manera, podemos escribir

$$\begin{aligned} d_p(aU, aV) &= \text{ínf}\left\{ \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|a\Pi_1 - a\Pi_2\|^p d\pi \right)^{1/p} : \pi \in \Pi(\mu, \nu) \right\} \\ &= \text{ínf}\left\{ \left(|a|^p \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\Pi_1 - \Pi_2\|^p d\pi \right)^{1/p} : \pi \in \Pi(\mu, \nu) \right\} \\ &= |a| d_p(U, V). \end{aligned}$$

(2) Supongamos que $\mathscr{L}(U) = \mu$, $\mathscr{L}(V) = \nu$ y $\pi \in \Pi(\mu, \nu)$ entonces, razonando como en el caso anterior, bajo π se tiene que

$$\mathscr{L}(\Pi_1) = \mu = \mathscr{L}(U) \quad \text{y} \quad \mathscr{L}(\Pi_2) = \nu = \mathscr{L}(V)$$

luego

$$\mathscr{L}(L\Pi_1) = \mathscr{L}(LU) \quad \text{y} \quad \mathscr{L}(L\Pi_2) = \mathscr{L}(LV).$$

En consecuencia, se tiene que

$$\begin{aligned} d_p(LU, LV) &= \text{ínf}\left\{ \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|L\Pi_1 - L\Pi_2\|^p d\pi \right)^{1/p} : \pi \in \Pi(\mu, \nu) \right\} \\ &\leq \text{ínf}\left\{ \left(\|L\| \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\Pi_1 - \Pi_2\|^p d\pi \right)^{1/p} : \pi \in \Pi(\mu, \nu) \right\} \\ &= \|L\| d_p(U, V). \end{aligned}$$

□

Proposición 2.22. *Sean $\{U_j\}_{j=1}^m$ variables aleatorias independientes con leyes de probabilidad en $\Pi_p(\mathbb{R}^d)$ y sean, análogamente, $\{V_j\}_{j=1}^m$ variables aleatorias independientes con leyes de probabilidad en $\Pi_p(\mathbb{R}^d)$. Entonces se tiene que*

$$d_p\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j \right) \leq \sum_{j=1}^m d_p(U_j, V_j).$$

Demostración. Podemos suponer sin pérdida de generalidad que los pares (U_j, V_j) son independientes entre sí y que además verifican

$$E(\|U_j - V_j\|^p)^{1/p} = d_p(U_j, V_j).$$

Ahora bien,

$$d_p\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right) \leq E\left(\|\sum_{j=1}^m U_j - \sum_{j=1}^m V_j\|^p\right)^{1/p} = E\left(\|\sum_{j=1}^m (U_j - V_j)\|^p\right)^{1/p}$$
$$\overset{(*)}{\leq} \sum_{j=1}^m E(\|U_j - V_j\|^p)^{1/p} = \sum_{j=1}^m d_p(U_j, V_j),$$

donde en la desigualdad marcada por $(*)$ se ha aplicado la clásica desigualdad de Minkowski. □

2.2.3. La 2-distancia de Wasserstein

El caso particular en el que $p = 2$, esto es, el caso de la 2-distancia de Wasserstein es el más importante y existen una gran cantidad de resultados sobre el mismo. Veamos algunas primeras propiedades, siguiendo nuevamente el trabajo de Bickel et al [2].

Una de las ventajas de trabajar con la 2-distancia de Wasserstein es que la norma $\|\cdot\|^2$ proviene de un producto interno lo que nos permite hablar del concepto de ortogonalidad, lo cual hace posible mejorar el resultado recogido en la proposición previa.

Proposición 2.23. *Sean $\{U_j\}_{j=1}^m$ variables aleatorias independientes que toman valores en $\mathbb{R}^d$ y sean también $\{V_j\}_{j=1}^m$ variables aleatorias independientes que toman valores en $\mathbb{R}^d$. Supongamos que las leyes de probabilidad de ambas familias de variables pertenecen a $\Pi_2(\mathbb{R}^d)$ y que $E(U_j) = E(V_j)$. Entonces*

$$d_2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right)^2 \leq \sum_{j=1}^m d_2(U_j, V_j)^2.$$

Demostración. Como en la proposición previa suponemos que los pares (U_j, V_j) son independientes y verifican $d_2(U_j, V_j) = E(\|U_j - V_j\|^2)^{1/2}$. Para poder pro-

bar la desigualdad de interés, en primer lugar veamos que

$$E(< U_j - V_j, U_i - V_i >) = \begin{cases} d_2(U_j, V_j)^2 & \text{si} \quad i = j, \\ 0 & \text{si} \quad i \neq j. \end{cases}$$

- Si $i = j$ se tiene que

$$E(< U_j - V_j, U_j - V_j >) = E(\|U_j - V_j\|^2) = d_2(U_j, V_j)^2$$

- Si $i \neq j$ se tiene que

$$\begin{aligned} &E(< U_j - V_j, U_i - V_i >) = \\ &= E(< U_j, U_i > - < U_j, V_i > - < V_j, U_i > + < V_j, V_i >) \\ &= E(< U_j, U_i >) - E(< U_j, V_i >) - E(< V_j, U_i >) + E(< V_j, V_i >) \\ &=< E(U_j), E(U_i) > - < E(U_j), E(V_i) > - < E(V_j), E(U_i) > + \\ &+ < E(V_j), E(V_i) >= (*) \end{aligned}$$

y teniendo en cuenta que $E(V_i) = E(U_i)$ y $E(V_j) = E(U_j)$, se deduce que

$$(*) = 2 < E(U_j), E(U_i) > -2 < E(U_j), E(U_i) >= 0.$$

Usando lo obtenido se deduce

$$d_2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right)^2 \leq E\left(\|\sum_{j=1}^m U_j - \sum_{j=1}^m V_j\|^2\right) =$$

$$= E\left(< \sum_{j=1}^m (U_j - V_j), \sum_{i=1}^m (U_i - V_i) >\right) = \sum_{i,j=1}^m E\left(< (U_j - V_j), (U_i - V_i) >\right)$$

$$= \sum_{j=1}^m E\left(< (U_j - V_j), (U_j - V_j) >\right) = \sum_{j=1}^m E\left(\|U_j - V_j\|^2\right) = \sum_{j=1}^m d_2(U_j, V_j)^2,$$

como queríamos probar. □

Proposición 2.24. *Sean U y V variables aleatorias con valores en $\mathbb{R}^d$, cuyas leyes pertenecen al conjunto $\Pi_2(\mathbb{R}^d)$. Entonces*

$$d_2(U, V)^2 = d_2(U - E(U), V - E(V))^2 + \|E(U) - E(V)\|^2.$$

Demostración. Denotamos $a = E(U)$ y $b = E(V)$ y elegimos U y V de manera que $d_2(U,V)^2 = E(\|U-V\|^2$. Ahora bien, se tiene que

$$E(\|(U-a)-(V-b)\|^2) = E(\|U-V\|^2) - \|a-b\|^2$$

puesto que

$$\begin{aligned} E(\|(U-a)-(V-b)\|^2) &= E(<(U-V)-(a-b),(U-V)-(a-b)>) \\ &= E(\|(U-V)\|^2 + \|a-b\|^2 - 2<U-V,a-b>) \\ &= E(\|(U-V)\|^2) + E(\|a-b\|^2) - 2E(<U-V,a-b>) \\ &= E(\|(U-V)\|^2) + \|a-b\|^2 - 2<E(U-V),E(a-b)> \\ &= d_2(U,V)^2 + \|a-b\|^2 - 2<a-b,a-b> \\ &= d_2(U,V)^2 + \|a-b\|^2 - 2\|a-b\|^2 = d_2(U,V)^2 - \|a-b\|^2, \end{aligned}$$

de lo que se deduce que

$$d_2(U-a,V-b)^2 \leq d_2(U,V)^2 - \|a-b\|^2.$$

Para probar la desigualdad contraria, basta tomar las variables U y V de manera que

$$E(\|(U-a)-(V-b)\|^2) = d_2(U-a,V-b)^2$$

así:

$$\begin{aligned} d_2(U,V)^2 \leq E(\|U-V\|^2) &= E(\|(U-a)-(V-b)\|^2 + \|a-b\|^2) \\ &= d_2(U-a,V-b)^2 + \|a-b\|^2 \\ \Longrightarrow d_2(U,V)^2 - \|a-b\|^2 &\leq d_2(U-a,V-b)^2 \end{aligned}$$

como queríamos. □

Proposición 2.25. *Supongamos $d = 1$, $\|x\| = |x|$ y $p = 2$. Sean $U_1,\ldots,U_n$ variables aleatorias independientes e igualmente distribuidas cuyas leyes pertenecen a $\Pi_2(\mathbb{R})$ y sea U un vector columna $(U_1,\ldots,U_n)$. Suponemos lo mismo para las variables $V_1,\ldots,V_n$ y V, además que $E(U_i) = E(V_i)$. Sea A una matriz $m \times n$ de escalares, entonces AU y AV son vectores aleatorios en $\mathbb{R}^m$ equipado con la norma euclídea m-dimensional. Escribimos d_2^m para la correspondiente d_2-métrica. Entonces*

$$d_2^m(AU,AV)^2 \leq \text{traza}(AA^t)d_2^1(U_i,V_i)^2.$$

Demostración. Suponemos que los pares de variables aleatorias (U_i, V_i) son independientes entre sí y que, además, verifican $d_2(U_i, V_i) = E((U_i - V_i)^2)^{1/2}$.

Teniendo en cuenta que dado un vector columna w de dimensión m su norma al cuadrado es la suma de sus elementos al cuadrado, que es lo mismo que la traza de la matriz $w\, w^t$, se tiene que

$$\begin{aligned} d_2^m(AU, AV)^2 &\leq E(\|AU - AV\|^2) = E(\|A(U - V)\|^2) \\ &= E\left[\text{traza}(A(U - V)(U - V)^t A^t)\right]. \end{aligned} \tag{2.8}$$

En primer lugar, calculamos la expresión de $A(U - V)$, que viene dada por

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{1n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} U_1 - V_1 \\ U_2 - V_2 \\ U_3 - V_3 \\ \vdots \\ U_n - V_n \end{pmatrix} = \left(\sum_{j=1}^{n} a_{ij}(U_j - V_j) \right)_{i=1,\ldots,m}.$$

Así, la expresión $A(U - V)(U - V)^t A^t$ la podemos calcular como sigue

$$\begin{pmatrix} \sum_{j=1}^n a_{1j}(U_j - V_j) \\ \vdots \\ \sum_{j=1}^n a_{mj}(U_j - V_j) \end{pmatrix}^t \begin{pmatrix} \sum_{j=1}^n a_{1j}(U_j - V_j) \\ \vdots \\ \sum_{j=1}^n a_{mj}(U_j - V_j) \end{pmatrix} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij}(U_j - V_j) \right)^2.$$

Estamos interesados en calcular el valor esperado de la expresión inmediatamente anterior, esto es,

$$E\left[\sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij}(U_j - V_j) \right)^2 \right] = \sum_{i=1}^{m} E\left[\left(\sum_{j=1}^{n} a_{ij}(U_j - V_j) \right)^2 \right].$$

Ahora bien, observamos que:

- Si $i = j$ entonces $E\left[(U_i - V_i)^2\right] = d_2^1(U_i, V_i)^2$.
- Si $i \neq j$ entonces

$$\begin{aligned} E\left[(U_i - V_i)(U_j - V_j)\right] &= E(U_i U_j - U_i V_J - V_i U_j + V_i V_j) = \\ &= E(U_i U_j) - E(U_i V_j) - E(V_i U_j) + E(V_i V_j) = \\ &= E(U_i)E(U_j) - E(U_i)E(V_j) - E(V_i)E(U_j) + E(V_i)E(V_j) = \\ &= E(U_i)E(U_j) - E(U_i)E(U_j) - E(U_i)E(U_j) + E(U_i)E(U_j) = 0, \end{aligned}$$

donde se ha hecho uso de las hipótesis de independencia de las familias de variables aleatorias $\{U_i\}_{i=1}^n$ y $\{V_i\}_{i=1}^n$ y de la hipótesis de que $E(U_i) = E(V_i)$. Además, se ha tenido en cuenta que la familia de los pares $\{(U_i, V_i)\}_{i=1}^n$ es independiente, por lo que las variables aleatorias obtenidas al aplicar las proyecciones sobre el primer y segundo elemento del par también son independientes entre ellas, esto es las variables U_i y V_i son independientes y, en consecuencia, las familias $\{U_i\}_{i=1}^n$ y $\{V_i\}_{i=1}^n$ son independientes entre sí también.

Sabiendo esto

$$E\left[\left(\sum_{j=1}^{n} a_{ij}(U_j - V_j)\right)^2\right] = a_{ii}^2 d_2^1(U_i, V_i)^2,$$

lo que implica que

$$\sum_{i=1}^{m} E\left[\left(\sum_{j=1}^{n} a_{ij}(U_j - V_j)\right)^2\right] = \sum_{i=1}^{m} a_{ii}^2 d_2^1(U_i, V_i)^2.$$

Teniendo en cuenta que las variables U_i están igualmente distribuidas y lo mismo pasa para las variables V_j, se tiene que $d_2^1(U_i, V_i)^2 = d_2^1(U_j, V_j)^2$ para todos $i, j \in \{1, \ldots, n\}$, luego

$$\begin{aligned}\sum_{i=1}^{m} E\left[\left(\sum_{j=1}^{n} a_{ij}(U_j - V_j)\right)^2\right] &= \left(\sum_{i=1}^{m} a_{ii}^2\right) d_2^1(U_i, V_i)^2 = \\ &= \text{traza}(AA^t) d_2^1(U_i, V_i)^2,\end{aligned}$$

como se quería probar. □

Caracterización de soluciones

Para este caso particular, existe una caracterización de la solución basada en la pertenencia a un subgradiente de una función semicontinua inferior y convexa. Vamos a esbozar la prueba de este resultado siguiendo el artículo de Rachev y Rüschendorf [17], pero antes vamos a introducir el problema de dualidad debido a Kantorovich mediante un ejemplo (como hicimos para introducir el problema de Kantorovich, siguiendo el libro más actual de Villani [19]) y algunas nociones sobre análisis convexo que vamos a requerir.

El problema primal de Kantorovich lo presentamos mediante el ejemplo de transportar una masa de arena a un agujero teniendo en cuenta que realizar dicho transporte tenía un coste asociado. Si pensamos en el caso discreto, podríamos plantearlo como el clásico problema de transporte en el que se tienen m fábricas con ciertas ofertas de un producto y n clientes con su respectivas demandas del mismo producto. Así, conocido el coste de transportar una unidad de producto desde cada fábrica hasta cada cliente, el problema primal de Monge-Kantorovich trataba de minimizar el coste del transporte. Esto es, en el problema primal la preocupación central era el coste.

Sin embargo, en el problema dual se da importancia al precio. Imaginemos que una empresa de transporte nos ofrece resolver nuestro porblema de transporte, es decir, nosotros desde cada fábrica vendemos toda nuestra producción a la empresa de transporte y luego ella se encarga de realizar las entregas a los clientes y poner sus precios de venta. Entonces, si suponemos que $\psi(x)$ es el precio por el que vendemos una unidad de producto de la fábrica x y que la empresa de tranporte vende una unidad de producto a precio $\phi(y)$ al cliente y, el dinero ganado por la empresa de transporte es $\phi(y) - \psi(x)$ por cada unidad de producto vendido entre la fábrica x y el cliente y. Desde nuestro punto de vista, la cantidad $\phi(y) - \psi(x)$ es el precio de transporte entre la empresa x y el cliente y. Así, para que los precios de la empresa sean competitivos es sensato pensar que el precio de transporte asociado a la empresa especializada debe ser menor que el precio de transporte que teníamos en el problema primal, esto es $\phi(y) - \psi(x) \leq c(x, y)$ (en caso contrario no saldría rentable contratar a la empresa de transporte ya que sacaríamos más beneficio realizando nosotros las entregas).

Así, cuando nos encargamos del transporte nos centramos en minimizar el coste, mientras que cuando contamos con la empresa de transporte en lo que se fijan ellos es en maximizar los beneficios obtenidos. De estas ideas surge de forma natural el problema dual de Kantorovich.

Definición 2.26 (Problema dual de Kantorovich)**.** Dados dos espacios probabilísticos $(\Omega_1, \sigma_1, \mu)$ y $(\Omega_2, \sigma_2, \nu)$ y siendo el problema primal de Kantorovich

$$\underset{\pi \in \Pi(\mu,\nu)}{\text{Minimizar}} \quad I(\pi) := \int_{\Omega_1 \times \Omega_2} c(x, y) d\pi(x, y),$$

el **problema dual de Kantorovich** se define como sigue

$$\underset{\substack{(\psi,\phi)\in(L_1(\mu),L_1(\nu))\\ \phi(y)-\psi(x)\leq c(x,y)}}{\text{Maximizar}} \int_{\Omega_2}\phi(y)d\nu(y)-\int_{\Omega_1}\psi(x)d\mu(x).$$

Ahora bien, la 2-distancia de Wasserstein entre dos probabilidades $\mu,\nu\in\Pi_2(\mathbb{R}^d)$ sabemos que viene dada por

$$d_2(\mu,\nu)^2=\text{ínf}\left\{\int_{\mathbb{R}\times\mathbb{R}}\|x-y\|^2d\pi(x,y)\ :\ \pi\in\Pi(\mu,\nu)\right\}, \tag{2.9}$$

y como en la versión probabilista se trabaja con variables aleatorias de cuadrado integrable, tenemos que si $\mathscr{L}(X)=\mu$ y $\mathscr{L}(Y)=\nu$ entonces

$$E\|X-Y\|^2=|EX-EY|^2+\text{traza}(\Sigma_\mu)+\text{traza}(\Sigma_\nu)-2\text{traza}(\text{Cov}(X,Y)),$$

donde $\Sigma_\mu=\text{Cov}(X)$ y $\Sigma_\nu=\text{Cov}(Y)$. Fijadas μ y ν, los tres primeros sumandos de la expresión inmediatamente anterior son fijos, luego (2.9) es equivalente a encontrar

$$\sup\left\{\text{traza}(\psi)\ :\ \psi\in\text{Cov}(\mu,\nu)\right\}, \tag{2.10}$$

donde

$$\text{Cov}(\mu,\nu)=\left\{\psi\in\mathbb{R}^d\times\mathbb{R}^d\ :\ \psi=\text{Cov}(X,Y)\ \text{ con }\ \mathscr{L}(X)=\mu\ \text{ y }\ \mathscr{L}(Y)=\nu\right\}.$$

Por último, el problema (2.10) es equivalente a

$$\sup\left\{E<X,Y>:\ \mathscr{L}(X)=\mu\ \text{ y }\ \mathscr{L}(Y)=\nu\right\}. \tag{2.11}$$

Vamos a adaptar el problema dual asociado a este último problema planteado haciendo uso de algunos conceptos de análisis real. Introducimos los conceptos de función semicontinua inferior y función conjugada.

Definición 2.27. Una función $f:\mathbb{R}^d\to\mathbb{R}$ se dice que es semicontinua inferior en el punto $x_0\in\mathbb{R}^d$ si para cada número real $y<f(x_0)$ existe un entorno U de x_0 de manera que $f(x)>y$ para todo $x\in U$. Equivalentemente, f es semicontinua inferior en x_0 si, y sólo si,

$$\liminf_{x\to x_0}f(x)>f(x_0).$$

Así, la función f se dice que es semicontinua inferior si es semicontinua inferior para cada punto de su dominio de definición.

Definición 2.28. Dada una función convexa y semicontinua inferior f en $\mathbb{R}^d$, denotamos por f^* su función conjugada que viene definida por

$$f^*(y) = \sup_{x\in\mathbb{R}^d} \{< x, y > -f(x)\}$$

y denotamos el subdiferencial de f en el punto x como

$$\partial f(x) = \{y \in \mathbb{R}^d \,:\, f(z) \geq f(x) + < z - x, y >,\ z \in \mathbb{R}^d\}.$$

A los elementos de $\partial f(x)$ los denominaremos subgradientes de f en x.

Observación 9. Si f es diferenciable en el punto x, entonces $\partial f(x)$ está formado por un único punto que coincide con $\nabla f(x)$.

Observamos que por definición tenemos que

$$f^*(y) = \sup_{x\in\mathbb{R}^d} \{< x, y > -f(x)\} \geq < x, y > -f(x) \iff f^*(y) + f(x) \geq < x, y >,$$

en consecuencia, para todo $x \in \mathbb{R}^d$ y para todo $y \in \mathbb{R}^d$ se verifica que

$$f^*(y) + f(x) \geq < x, y > \text{ con igualdad si, y sólo si, } y \in \partial f(X).$$

Con las nociones expuestas podemos esbozar la demostración del teorema de caracterización de soluciones.

Teorema 2.29. *Sean $\mu, \nu \in \Pi_2(\mathbb{R}^d)$.*

(a) Existe solución de (2.9), esto es, existen variables aleatorias X e Y con $\mathscr{L}(X) = \mu$ y $\mathscr{L}(Y) = \nu$ tales que $d_2(\mu, \nu) = (E\|X - Y\|^2)^{1/2}$.

(b) Sean X e Y tales que $\mathscr{L}(X) = \mu$ y $\mathscr{L}(Y) = \nu$, entonces (X, Y) es una solución de (2.9) si, y sólo si, $Y \in \partial f(X)$ casi seguro para alguna función semicontinua inferior y convexa f.

Demostración. Que la 2-distancia de Wasserstein entre dos probabilidades $\mu, \nu \in \Pi_2(\mathbb{R}^d)$ alcanza su ínfimo ya lo sabemos, por tanto pasamos a probar el apartado (b).

($\Longleftarrow$) Sean X e Y variables aleatorias y suponemos que $Y \in \partial f(X)$ casi seguro para alguna función semicontinua inferior y convexa f. De esta manera, para otras variables $\tilde{X}$ e $\tilde{Y}$ tales que $\mathscr{L}(\tilde{X}) = \mu$ y $\mathscr{L}(\tilde{Y}) = \nu$, se tiene que

$$E < \tilde{X}, \tilde{Y} > \leq E(f(\tilde{X}) + f^*(\tilde{Y})) = E(f(X) + f^*(Y)) = E < X, Y >,$$

lo que prueba que (X, Y) es solución óptima de (2.9).

($\Longrightarrow$) Para esta implicación necesitamos usar un resultado auxiliar de dualidad expuesto por Kellerer que nos garantiza la siguiente igualdad

$$C(\mu,\nu) := \sup\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} < x, y > d\pi(x,y) \ : \ \pi \in \Pi(\mu,\nu)\right\} =$$
$$= \text{ínf}\left\{\int_{\mathbb{R}^d} \psi(x)d\mu(x) + \int_{\mathbb{R}^d} \phi(y)d\nu(y) \ : \ \begin{array}{l} \psi \in L_1(\mu),\ \phi \in L_1(\nu) \\ < x, y > \leq \psi(x) + \phi(y) \end{array}\right\} =: I(\mu,\nu),$$

para la prueba de este resultado se puede consultar el Teorema 2.6 del artículo [13].

Del apartado (a) sabemos que como $C(\mu,\nu)$ admite solución entonces $I(\mu,\nu)$ también tiene solución y, por lo tanto, existen funciones $\psi \in L_1(\mu)$ y $\phi \in L_1(\nu)$ de manera que

$$I(\mu,\nu) = \int_{\mathbb{R}^d} \psi(x)d\mu(x) + \int_{\mathbb{R}^d} \phi(y)d\nu(y) \quad \text{y} \quad < x, y > \leq \psi(x) + \phi(y).$$

Ahora bien, si definimos $f = \psi^{**}$ tenemos que f es convexa y semicontinua inferior y

$$< x, y > \leq f(x) + f^*(y) \leq \psi(x) + \phi(y)$$

o equivalentemente, el par (f, f^*) es solución de

$$\text{ínf}\left\{\int_{\mathbb{R}^d} \psi(x)d\mu(x) + \int_{\mathbb{R}^d} \phi(y)d\nu(y) \ : \ \begin{array}{l} \psi \in L_1(\mu),\ \phi \in L_1(\nu) \\ < x, y > \leq \psi(x) + \phi(y) \end{array}\right\}.$$

Entonces $< X, Y >= f(X) + f^*(Y)$ casi seguro y, por tanto, $Y \in \partial f(X)$. $\square$

El resultado anterior es muy interesante, pues garantiza que en el caso de existir una solución óptima el par de variables aleatorias correspondientes van a verificar que una pertenece al subgradiente de una función convexa y semicontinua inferior de la otra. Este teorema puede establecerse también haciendo uso de argumentos de monotonía cíclica y es lo siguente que vamos a presentar, sin demostraciones.

Empezamos introduciendo los conceptos de conjunto cíclicamente monótono y de función de transferencia cíclicamente monótona, siguiendo de nuevo el libro de Villani [19].

Definición 2.30. Sea $c : \mathbb{R}^d \times \mathbb{R}^d \to (-\infty, \infty]$ una función. Un **subconjunto** $\Gamma \subset \mathbb{R}^d \times \mathbb{R}^d$ se dice que es ***c*-cíclicamente monótono** si, para cada $N \in \mathbb{N}$ y para cada familia $(x_1, y_1), \ldots, (x_N, y_N)$ de puntos de Γ, se verifica la siguiente desigualdad

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}) \tag{2.12}$$

(con el convenio de que $y_{N+1} = y_1$). Una **función de transferencia** se denomina ***c*-cíclicamente monótona** si está concentrada en un conjunto cíclicamente monótono.

En el caso discreto, en el que tenemos un problema de transporte entre m fábricas y n clientes, este concepto nace de pensar que el coste de transporte obtenido es muy alto y tratar de reducirlo redirigiendo transportes realizados entre una fábrica y un cliente hacia otros clientes más cercanos. Para ello, tomamos una fábrica x_1 que transporta k unidades del producto en cuestión a un cliente y_1, y decidimos que una unidad de producto de ese envío se va a redirigir a otro cliente y_2 que está más cerca de la fábrica x_1. Entonces, en términos de la función de coste c, ganaríamos $c(x_1, y_2) - c(x_1, y_1)$. Como el problema de optimización tenía una solución factible que estamos alterando, ahora resulta que el cliente y_2 tiene un exceso en las unidades de producto solicitado, entonces esa unidad de más que recibe de otra fábrica x_2 debe ser redirigida a otro cliente, digamos y_3. Este proceso se sigue hasta que conseguimos redirigir una unidad de producto desde una fábrica x_N al cliente y_1 en el que hemos empezado los cambios. Así, la nueva solución factible calculada es mejor que la anterior si y sólo si

$$c(x_1, y_2) + c(x_2, y_3) + \ldots + c(x_N, y_1) < c(x_1, y_1) + c(x_2, y_2) + \ldots + c(x_N, y_N).$$

Una función de transferencia c-cíclicamente monótona, es aquella que no puede ser mejorada, esto es, no podemos realizar pertubaciones en ella (refiriéndonos a redirecciones en el transporte) con la finalidad de obtener otra más "económica". Así, si una función de transferencia es óptima, es claro que debe ser también c-cíclicamente monótona. Es bien conocido que el recíproco también se cumple, de hecho, una función es cíclicamente monótona si, y sólo si, esta puede escribirse como el subgradiente de una función semicontinua inferior y convexa.

Teorema 2.31. *Sean $\mu, \nu \in \Pi_2(\mathbb{R}^d)$. Consideramos el problema de Monge-Kantorovich asociado a la función coste $c(x, y) = \|x - y\|^2$. Entonces:*

(i) El vector (X, Y) con $\mathscr{L}((X, Y)) = \pi \in \Pi(\mu, \nu)$ es óptimo si, y sólo si, existe una función φ convexa y semicontinua inferior tal que

$$Y \in \partial\varphi(X), \ \pi\text{-casi seguro.}$$

Además, en estas condiciones, el par (φ, φ^*) es un minimizante del problema dual

$$\text{ínf}\left\{\int_{\mathbb{R}^d} \phi(x)d\mu(x) + \int_{\mathbb{R}^d} \psi(y)d\nu(y) \ : < x, y > \leq \phi(x) + \psi(y)\right\}.$$

(ii) Si μ no da probabilidad a conjuntos de dimensión $d-1$, entonces existe una única probabilidad $\pi \in \Pi(\mu, \nu)$ óptima, que viene dada por

$$d\pi(x, y) = d\mu(x)\delta\left(y = \nabla\varphi(x)\right),$$

donde $\nabla\varphi$ es el único gradiente (μ-casi seguro) de una función que transporta μ en ν, esto es: $\nabla\varphi\#\mu = \nu$.

(iii) Si μ no da probabilidad a conjuntos de dimensión $d-1$, entonces $\nabla\varphi$ es la única solución del problema de Monge

$$\int_{\mathbb{R}^d} \|x - \nabla\varphi(x)\|^2 d\mu(x) = \int_{\mathbb{R}^d} \|x - T(x)\|^2 d\mu(x).$$

(iv) Si μ y ν no dan probabilidad a conjuntos de dimensión $d-1$, entonces para todo x e y se tiene que

$$\nabla\varphi^*\left(\nabla\varphi(x)\right) = x, \quad \mu\text{-casi seguro;}$$

y

$$\nabla\varphi\left(\nabla\varphi^*(y)\right) = y, \quad \nu\text{-casi seguro.}$$

Además, $\nabla\varphi^*$ es el único gradiente, ν-casi seguro, de una función convexa y semicontinua inferior que transporta ν en μ, esto es, $\nabla\varphi^*\#\nu = \mu$. Desde otro punto de vista, $\nabla\varphi^*$ es la solución del problema de Monge de transportar ν en μ considerando el coste cuadrático.

2.3. Cópulas y estructuras de dependencia

Para desarrollar esta sección vamos a seguir como referencia principal el artículo de Cuesta-Albertos et al [11] para los conceptos y resultados relacionados con las cópulas, estructuras de dependencia y el transporte óptimo y, también utilizamos el trabajo de Rüschendorf [16] para los aspectos relacionados con la transformada distribucional y el Teorema de Sklar.

Dado un espacio de probabilidad (Ω, σ, P), sea X una variable aleatoria con función de distribución F y sea V una variable aleatoria con distribución uniforme en $(0,1)$ independiente de X. La **función de distribución modificada** $F(x,\lambda)$ se define como

$$F(x,\lambda) := P(X < x) + \lambda P(X = x).$$

Y se define la **transformada distribucional de X** por

$$U := F(X,V). \tag{2.13}$$

Una representación equivalente de la transformada distribucional de X es la siguiente

$$U = F(X-) + V(F(X) - F(X-)).$$

Para el caso de funciones de distribución continuas es evidente que $F(x,\lambda)$ coincide con $F(x)$ y, en este caso, es conocido que $U = F(X)$ sigue una distribución uniforme en $(0,1)$. En esta línea, el hecho de que la transformada distribucional de una función de distribución F siga una distribución uniforme en $(0,1)$ es siempre cierto, incluso en los casos en los que F no se trate de una función de distribución continua.

Proposición 2.32. *Sea U la transformada distribucional de X definida por* (2.13). *Entonces $U \overset{d}{=} U(0,1)$ y $X = F^{-1}(U)$ casi seguro.*

En el contexto de la 2-distancia de Wasserstein, los emparejamientos óptimos están muy relacionados con la transformada distribucional y con el concepto de variables similarmente ordenadas que exponemos a continuación.

Definición 2.33. Dadas dos variables aleatorias X_1 y X_2 se dice que están similarmente ordenadas si

$$(X_1(\omega) - X_1(\omega'))(X_2(\omega) - X_2(\omega')) \geq 0$$

$P \times P$- casi seguro, y escribiremos $X_1 \overset{s.o.}{\sim} X_2$.

Para desarrollar la teoría de cópulas y de distribuciones con la misma o igual estructura de dependencia vamos a partir del siguiente resultado que se puede encontrar demostrado en [11].

Proposición 2.34. *Sean X_1 y X_2 variables de cuadrado integrable con funciones de distribución F_1 y F_2, respectivamente. Entonces:*

(a) Son equivalentes:

(a.i) (X_1, X_2) es un emparejamiento óptimo para la 2-distancia de Wasserstein.

(a.ii) $F_{(X_1,X_2)}(x, y) = \min\{F_1(x), F_2(y)\}$ para todo x, y.

(a.iii) Existe una variable aleatoria U, uniforme en $(0, 1)$, tal que para funciones crecientes ϕ_1 y ϕ_2 se tiene

$$X_1 = \phi_1(U) \quad \text{y} \quad X_2 = \phi_2(U), \quad P\text{-casi seguro.}$$

(a.iv) $X_1 \overset{s.o.}{\sim} X_2$.

(b) Las funciones ϕ_1 y ϕ_2 en (a) son esencialmente únicas, $\phi_i = F_i^{-1}$ casi seguro respecto de la medida de Lebesgue.

(c) Los pares $(X_1, F_2^{-1} \circ F_1(X_1, V_1))$ y $(F_1^{-1} \circ F_2(X_2, V_2), X_2)$ son óptimos para la 2-distancia de Wasserstein.

(d) Si P_1 es no atómica y (X_1, Y_1) es un emparejamiento óptimo para la 2-distancia de Wasserstein entre P_1 y P_2, entonces

$$Y_1 = F_2^{-1} \circ F_1(X_1), \quad \text{casi seguro.}$$

(e) Si $Y_1 = \phi_1(X_1)$ con ϕ_1 creciente, entonces (X_1, Y_1) es un emparejamiento óptimo para la 2-distancia de Wasserstein.

Este resultado no es del todo novedad, pues ya sabemos que si (X, Y) es un emparejamiento óptimo para la distancia de Wasserstein, del desarrollo hecho en el apartado 2.2.1, su función de distribución era $H_+(X, Y) = \min\{F_1(x), F_2(y)\}$; que corresponde a la equivalencia $(a.i) \Leftrightarrow (a.ii)$. Esto es, la distribución dada en $(a.ii)$ es la asociada a la pareja de funciones cuantiles asociadas a las funciones de distribución marginales.

Los resultados en los que vamos a centrar nuestro interés son los emparejamientos relativos a transformaciones monótonas componente a componente, pues el caso de la 2-distancia de Wasserstein es uno de ellos. Observamos que para la función de coste cuadrática $c_2(x,y) = \|x-y\|^2$ se verifica, trabajando con $x = (x_1, \ldots, x_d)$ e $y = (y_1, \ldots, y_d)$, que:

$$c_2(x,y) = \sum_{i=1}^{d} (x_i - y_i)^2 = \sum_{i=1}^{d} c_2(x_i, y_i).$$

La optimalidad de este tipo de transformaciones se recoge en el siguiente resultado que se enuncia y prueba para la función de coste cuadrática, pero que es cierto en el caso más general de funciones de coste de la forma $c(x,y) = \sum_{i=1}^{d} c_i(x_i, y_i)$ donde $c_i(x_i, y_i) = \phi_i(x_i - y_i)$ con ϕ_i estrictamente convexa.

Proposición 2.35. *Consideramos la función de coste cuadrático. Sean $\mu, \nu \in \Pi_2(\mathbb{R}^d)$ con distribuciones marginales respectivas μ_i, ν_i, $1 \leq i \leq d$ y sean X e Y variables aleatorias con $\mathscr{L}(X) = \mu$ y $\mathscr{L}(Y) = \nu$, entonces:*

(a) $d_2(\mu,\nu)^2 \geq \sum_{i=1}^{d} d_2(\mu_i, \nu_i)^2$.

(b) $X_i \overset{s.o.}{\sim} Y_i$, $1 \leq i \leq d$ *si, y sólo si,* (X,Y) *es un emparejamiento óptimo para la 2-distancia de Wasserstein y* $d_2(\mu,\nu)^2 = \sum_{i=1}^{d} d_2(\mu_i,\nu_i)^2$

Demostración. Para probar el apartado (a) basta observar que

$$\begin{aligned} d_2(\mu,\nu)^2 &= \text{ínf}\left\{ \int \sum_{i=1}^{d} |x_i - y_i|^2 d\pi(x_i, y_i) \ : \ \pi \in \Pi(\mu,\nu) \right\} \\ &\geq \sum_{i=1}^{d} \text{ínf}\left\{ \int |x_i - y_i|^2 d\pi(x_i, y_i) \ : \ \pi \in \Pi(\mu,\nu) \right\} \\ &= \sum_{i=1}^{d} \text{ínf}\left\{ \int |x_i - y_i|^2 d\pi_i(x_i, y_i) \ : \ \pi_i \in \Pi(\mu_i,\nu_i) \right\} = \sum_{i=1}^{d} d_2(\mu_i,\nu_i)^2. \end{aligned}$$

Por otro lado, se da la igualdad en (a) si, y sólo si, existe un emparejamiento óptimo (X,Y) para la la 2-distancia de Wasserstein, de manera que $E(|X_i - Y_i|^2) = d_2(\mu_i,\nu_i)^2$ para todo $1 \leq i \leq d$. Una condición suficiente es que $X_i \overset{s.o.}{\sim} Y_i$ para todo $1 \leq i \leq d$. □

Del resultado previo y de la Proposición 2.34 se deduce el siguiente corolario.

Corolario 2.36. *Consideramos la función de coste cuadrático. Sean $\mu, \nu \in \Pi(\mathbb{R}^d)$ con distribuciones marginales μ_i, ν_i, donde μ_i es no atómica, y funciones de distribución marginales F_i y G_i, $1 \leq i \leq d$. Sean X e Y vectores aleatorios con $\mathscr{L}(X) = \mu$ y $\mathscr{L}(Y) = \nu$, entonces son equivalentes:*

(1) (X, Y) es un emparejamiento óptimo para la 2-distancia de Wasserstein y $d_2(\mu, \nu)^2 = \sum_{i=1}^d d_2(\mu_i, \nu_i)^2$.

(2) $Y_i = G_i^{-1} \circ F_i(X_i)$ casi seguro $1 \leq i \leq d$.

Demostración.

(1) $\Longrightarrow$ (2) Con las hipótesis de (1) estamos en condiciones de aplicar el apartado (b) de la Proposición 2.35, del que se deduce que $X_i \overset{s.o.}{\sim} Y_i$ para todo $1 \leq i \leq d$. Y con lo obtenido del apartado (a) de la Proposición 2.34 se tiene que (X_i, Y_i) es un emparejamiento óptimo para $d_2(\mu_i, \nu_i)$ para todo $1 \leq i \leq d$. Nuevamente del apartado (d) de la Proposición 2.34 se tiene que $Y_i = G_i^{-1} \circ F_i(X_i)$ casi seguro para todo $1 \leq i \leq d$.

(2) $\Longrightarrow$ (1) Como las funciones $G_i^{-1} \circ F_i$ son crecientes del apartado (e) de la Proposición 2.34 se deduce que (X_i, Y_i) es un emparejamiento óptimo para la 2-distancia de Wasserstein para todo $1 \leq i \leq d$. Esto sabemos que es equivalente a que $X_i \overset{s.o.}{\sim} Y_i$ para todo $1 \leq i \leq d$ en virtud del apartado (a) de la Proposición 2.34. Por último, aplicando la proposición previa tenemos que (X, Y) es un emparejamiento óptimo para la función de coste $\|x - y\|^2$ y $d_2(\mu, \nu)^2 = \sum_{i=1}^d d_2(\mu_i, \nu_i)^2$. □

A continuación introducimos el concepto de cópula y tratamos el problema de la existencia y construcción de la misma vía el Teorema de Sklar.

Definición 2.37. Se dice que una aplicación $C : [0,1]^d \to [0,1]$ es una **cópula** si es la función de distribución de un vector aleatorio d-dimensional en el cubo unidad con funciones de distribución marginales uniformes.

Es muy conocido el resultado de existencia de cópulas para cualquier función de distribución d-dimendional, el cual presentamos en el siguiente teorema y demostramos haciendo uso de la transformada distribucional.

Teorema 2.38 (Teorema de Sklar)**.** *Sea F una función de disfribución d-dimensional con funciones de distribución marginales $F_1, \ldots, F_d$. Entonces*

existe una cópula C tal que

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{2.14}$$

Demostración. Sea $X = (X_1, \ldots, X_d)$ un vector aleatorio en un espacio de probabilidad (Ω, σ, P) con función de distribución F y consideremos V independiente de X y con distribución uniforme en $(0, 1)$. Trabajando con las transformadas distribucionales $U_i := F_i(X_i, V)$, $1 \leq i \leq d$, tenemos que $U_i \stackrel{d}{=} U(0, 1)$ y $X_i = F_i^{-1}(U_i)$ casi seguro para $1 \leq i \leq d$, en virtud de la Proposición 2.32.

Ahora bien, definiendo C como la función de distribución del vector $U = (U_1, \ldots, U_n)$ tenemos que

$$\begin{aligned} F(x) &= P(X \leq x) = P(X_i \leq x_i, 1 \leq i \leq d) = P(F^{-1}(U_i) \leq x_i, 1 \leq i \leq d) \\ &= P(U_i \leq F(x_i), 1 \leq i \leq d) = C(F_1(x_1), \ldots, F_d(x_d)), \end{aligned}$$

esto es, C es una cópula de F como queríamos probar. □

Definición 2.39. Denominaremos a cualquier función C como en (2.14) una **función de dependencia** de F.

Por la construcción realizada en la demostración del Teorema de Sklar, es evidente que si las funciones de distribución marginales F_i son continuas, $1 \leq i \leq d$, entonces la función C está determinada de forma única.

Otra forma de realizar esta construcción sería tomar una familia $\{V_i\}_{i=1}^d$ de variables aleatorias, independientes e igualmente distribuidas con distribución uniforme en $(0, 1)$ que sean independientes de $\{X_i\}_{i=1}^d$. De esta manera tomando $U_i = F_i(X_i, V_i)$ tendríamos que $U_i \stackrel{d}{=} U(0, 1)$ y $X_i = F_i^{-1}(U_i)$ y, como antes, definiríamos una cópula de F como

$$C^*(F_1(x_1), \ldots, F_d(x_d)) = F(x_1, \ldots, x_d),$$

o equivalentemente

$$C^*(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots U_n \leq u_d).$$

En este caso, C^* depende de la elección de la familia $\{V_i\}_{i=1}^n$, pero si las funciones de distribución marginales son continuas entonces coincide con la función

$$D(u_1, \ldots, u_d) = P(F_1(X_1) \leq u_1, \ldots, F_d(X_d) \leq u_d).$$

Definición 2.40. Se dice que dos probabilidades μ, ν tienen la **misma estructura de dependencia** si existe una función de dependencia común para μ y ν. Esto es, si para la misma eleción de la familia $\{V_i\}_{i=1}^d$ se tiene que

$$C_X^* = C_Y^*, \quad \text{para algún par de vectores } \mathscr{L}(X) = \mu, \mathscr{L}(Y) = \nu.$$

Continuamos con el último resultado de esta sección, que relaciona las afirmaciones vistas en proposiciones previas sobre las transformciones componente a componente con las distribuciones con la misma estructura de dependencia.

Teorema 2.41. *Sean $\mu, \nu \in \Pi_2(\mathbb{R}^d)$ con distribuciones marginales μ_i, ν_i, $1 \leq i \leq d$. Consideramos X un vector aleatorio tal que $\mathscr{L}(X) = \mu$, entonces son equivalentes:*

(a) $d_2(\mu, \nu)^2 = \sum_{i=1}^d d_2(\mu_i, \nu_i)^2$.

(b) *μ y ν tienen la misma estructura de dependencia.*

(c) *$(X, \overline{Y})$ es un emparejamiento óptimo para $d_2(\mu, \nu)$, donde $\overline{Y}_i = G_i^{-1} \circ F_i(X_i, V_i)$ siendo $\{V_i\}_{i=1}^d$ variables aleatorias independientes, igualmente distribuidas e independientes de X, $V_i \overset{d}{=} U(0,1)$.*

(d) *Existe Y con $\mathscr{L}(Y) = \nu$ tal que $X_i \overset{s.o.}{\sim} Y_i$, $1 \leq i \leq d$.*

Demostración. Para realizar la demostración vamos a probar las siguientes implicaciones $(a) \iff (d)$ y $(b) \Rightarrow (c) \Rightarrow (d) \Rightarrow (b)$.

$(a) \iff (d)$ Es consecuencia directa del apartado (b) de la Proposición 2.35.

$(c) \implies (d)$ Si $(X, \overline{Y})$ es un emparejamiento óptimo para $d_2(\mu, \nu)$, entonces $\mathscr{L}(\overline{Y}) = \nu$ donde el vector $\overline{Y}$ está definido por $\overline{Y_i} = G_i^{-1} \circ F_i(X_i, V_i)$ siendo $\{V_i\}_{i=1}^d$ variables aleatorias independientes, igualmente distribuidas e independientes de $\{X_i\}_{i=1}^d$, $V_i \overset{d}{=} U(0,1)$. Luego ya tenemos una variable aleatoria con ley de distribución ν, sólo falta probar que, componente a componente, está similarmente ordenada a X.
Como $\overline{Y_i} = G_i^{-1} \circ F_i(X_i, V_i)$ para todo $1 \leq i \leq d$, del Corolario 2.36 se sigue que $d_2(\mu, \nu)^2 = \sum_{i=1}^d d_2(\mu_i, \nu_i)^2$ y de la proposición 2.35 apartado (b) se deduce que $X_i \overset{s.o.}{\sim} Y_i$, $1 \leq i \leq d$.

$(b) \implies (c)$ Si μ y ν tienen la misma estructura de dependencia y F_i, G_i son su respectivas funciones de distribución marginales, entonces existen vectores

aleatorios X e Y con $\mathscr{L}(X) = \mu$ y $\mathscr{L}(Y) = \nu$ de manera que los vectores $U_i = F_i(X_i, V_i)$ y $\tilde{U}_i = G_i(Y_i, V_i)$ verifican

$$U = (U_1, \ldots, U_d) \overset{d}{=} \tilde{U} = (\tilde{U}_1, \ldots, \tilde{U}_d)$$

y, en consecuencia,

$$Y = (G_1^{-1}(\tilde{U}_1), \ldots, G_d^{-1}(\tilde{U}_d)) \overset{d}{=} (G_1^{-1}(U_1), \ldots, G_d^{-1}(U_d)) = \overline{Y} \implies \mathscr{L}(\overline{Y}) = \nu.$$

Además, en virtud de los apartados (c) y (a) de la Proposición 2.34, se tiene que:

(1) Para todo $1 \leq i \leq d$ el vector

$$(X_i, G_i^{-1} \circ F_i(X_i, V_i)) = (X_i, \overline{Y}_i)$$

es un emparejamiento óptimo para $d_2(\mu_i, \nu_i)$.

(2) Como $(X_i, \overline{Y}_i)$ es un emparejamiento óptimo para $d_2(\mu_i, \nu_i)$ entonces

$$X_i \overset{s.o.}{\sim} \overline{Y}_i, \quad 1 \leq i \leq d.$$

Ahora bién, de la Proposición 2.35 se sigue que $(X, \overline{Y})$ es un emparejamiento óptimo para $d_2(\mu, \nu)$.

$(d) \implies (b)$ Si existen X e Y tales que $\mathscr{L}(X) = \mu$, $\mathscr{L}(Y) = \nu$ y $X_i \overset{s.o.}{\sim} Y_i$, de los apartados (b) y (a) de la Proposición 2.34 se deduce que existen variables $U_1, \ldots, U_d$ uniformes en $(0, 1)$ tales que

$$X_i = F_i^{-1}(U_i) \quad \text{y} \quad Y_i = G_i^{-1}(U_i), \quad 1 \leq i \leq d.$$

Consideramos $C(x_1, \ldots, x_d) = P(U_1 \leq x_1, \ldots, U_n \leq x_d)$, entonces

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$$

y

$$G(x_1, \ldots, x_d) = C(G_1(x_1), \ldots, G_d(x_d)),$$

lo que prueba que μ y ν tienen la misma estructura de dependencia. $\square$

Como en la Proposición 2.35 y su correspondiente corolario, este resultado también es cierto para funciones de coste de la forma $c(x, y) = \sum_{i=1}^{d} c_i(x_i, y_i)$ donde $c_i(x_i, y_i) = \phi_i(x_i - y_i)$ con ϕ_i estrictamente convexa.

Lo más destacado de todo lo expuesto es que al buscar emparejamientos óptimos para la 2-distancia de Wasserstein entre distribuciones con la misma estructura de dependencia, es suficiente trabajar componente a componente. Este resultado tendrá mayor relevancia en el próximo capítulo.

2.4. Factorización polar

En las secciones previas hemos visto que los subgradientes de funciones convexas son los grandes protagonistas en la teoría del transporte óptimo para la función de coste cuadrática. En este caso particular, la factorización polar establece que si se busca transportar una medida de probabilidad a otra mediante una aplicación de transporte óptimo, entonces esta aplicación puede expresarse como la composición de una transformación monótona y una rotación. Es decir, el transporte óptimo se puede descomponer en una parte que reorganiza la masa de manera monótona y una parte que rota la masa.

Los resultados de factorización polar abstracta suelen asociarse a Brenier, ya que dedicó parte de su investigación a formular un teorema de factorización polar en un contexto abstracto que conseguía generalizar otros resultados bien conocidos como la descomposición polar de matrices o el teorema de descomposición de campos de vectores de Helmholtz. En esta línea, siguiendo los artículos [5] y [6] de Brenier, empezamos introduciendo los conceptos necesarios para poder exponer el Teorema de factorización polar en el caso particular y después ilustramos su relación con el transporte óptimo. Las dos nociones que aparecen en el resultado de interés son las aplicaciones que conservan medidas y las reorganizaciones o reordenaciones (en inglés, rearrangements).

Definición 2.42. Una **aplicación que conserva medidas** de un espacio de probabilidad (Ω_1, μ) en otro espacio de probabilidad (Ω_2, ν) es una aplicación $s : \Omega_1 \to \Omega_2$ tal que $s^{-1}(A)$ es un conjunto μ-medible de Ω_1 y $\mu(s^{-1}(A)) = \nu(A)$, para todo conjunto A ν-medible de Ω_2. O equivalentemente, $f \circ s$ es una función integrable respecto de μ y $\int_{\Omega_1} f \circ s d\mu = \int_{\Omega_2} f d\nu$, para cada función

f integrable respecto de ν.

Observación 10. Esta definición observamos que es totalmete análoga a la de aplicación que transporta μ en ν, esto es, que la aplicación $s : \Omega_1 \to \Omega_2$ conserve medidas de (Ω_1, μ) en (Ω_2, ν) es equivalente a $\nu = s\#\mu$.

Definición 2.43. Si $d \in \mathbb{N}$, llamamos **reordenación** de $u \in L^2(\Omega_1, \mu; \mathbb{R}^d)$ en (Ω_2, ν), a cualquier función $v \in L^2(\Omega_2, \nu; \mathbb{R}^d)$ que verifica

$$\int_{\Omega_1} f(u(x))d\mu(x) = \int_{\Omega_2} f(v(y))d\nu(y)$$

para cada función f continua en $\mathbb{R}^d$ tal que $|f(z)| \leq C(1 + \|z\|^2)$.

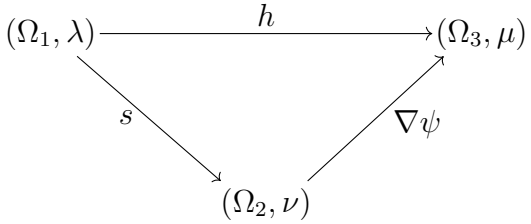
Observación 11. En la terminología probabilística, suponiendo que μ y ν son probabilidades, la definición previa es equivalente a decir que u y v son igualmente distribuidas, esto es, las medidas que engendran en $\mathbb{R}^d$ son iguales.

Como se recoge en el tercer capítulo del libro de Villani [19], el Teorema de factorización polar introducido por Brenier, para el caso particular en el que se trabaja con $p = 2$, puede exponerse de la siguiente manera:

Teorema 2.44. *Sean Ω_1 y Ω_2 subconjuntos de $\mathbb{R}^d$ y consideramos medidas de probabilidad $\lambda \in \Pi(\Omega_1)$ y $\nu \in \Pi_2(\Omega_2)$. Sea $h : \Omega_1 \to \Omega_3 \subset \mathbb{R}^d$ una función de $L^2(\Omega_1, \lambda)$ y consideramos $\mu = h\#\lambda$. Suponiendo que μ y ν no dan masa a conjuntos de dimensión $d-1$, entonces existe un único par $(\nabla\psi, s)$ tal que*

- *$\psi : \Omega_2 \to \Omega_3$ es convexa,*
- *$s : \Omega_1 \to \Omega_2$ es tal que $s\#\lambda = \nu$,*
- *$h = \nabla\psi \circ s$.*

En particular, el diagrama



es conmutativo. Además, s es la única proyección de h en $S(\Omega_1, \Omega_2)$, el conjunto de aplicaciones σ que verifican $\sigma\#\lambda = \nu$.

Este teorema resulta ser prácticamente equivalente al Teorema 2.31, en el que se recogían diferentes caracterizaciones de la solución para la 2-distancia de Wasserstein.

Observación 12. En el contexto anterior, a la función h se le puede asociar una probabilidad α de manera que

$$\int_{\mathbb{R}^d} f(y)d\alpha(y) = \int_{\Omega_2} f(h(x))d\nu(x)$$

para toda función f continua en $\mathbb{R}^d$ y de soporte compacto.

Un problema de interés sería tratar el caso en el que se trabaja con una sucesión de medidas de probabilidad $\{\alpha_n\}_{n=1}^{\infty}$ que converge débilmente hacia otra probabilidad α, en vez de trabajar únicamente con la medida de probabilidad α que se le ha asociado a la función h.

Aunque no aparezca recogido en el resultado anterior, en [6] (Teorema 1.3) Brenier demostró que, al trabajar con la sucesión de funciones convexas $\{\psi_n\}_{n=1}^{\infty}$ de $L^2(\Omega_2, \nu)$ asociada a la sucesión de funciones $\{h_n\}_{n=1}^{\infty}$ de $L^2(\Omega_1, \lambda)$ en el sentido del teorema previo, si la sucesión de probabilidades $\{\alpha_n\}_{n=1}^{\infty}$ que se le asocia a la sucesión $\{h_n\}_{n=1}^{\infty}$ verifica que $\alpha_n \to \alpha$ débilmente, entonces se tiene la convergencia

$$\nabla\psi_n \to \nabla\psi \quad \text{en} \quad L^2(\Omega, \nu).$$

En la misma línea, en 1997 Cuesta-Albertos et al demostraron en [10] (Teorema 3.4) que la convergencia $\nabla\psi_n \to \nabla\psi$ es, de hecho, ν-casi seguro. Enunciamos a continuación el resultado.

Teorema 2.45. *Sean $\{\mu_n\}_{n=1}^{\infty}$, $\{\nu_n\}_{n=1}^{\infty}$, μ y ν medidas en $\Pi_2(\mathbb{R}^d)$ tales que μ es absolutamente continua respecto la medida de Lebesgue, $\mu_n \to \mu$ débilmente y $\nu_n \to \nu$ débilmente. Supongamos que H_n (resp. H) son transportes óptimos entre μ_n y ν_n (resp. entre μ y ν) para todo $n \in \mathbb{N}$. Entonces, si $\{X_n\}_{n=1}^{\infty}$ es una sucesión de vectores aleatorios que converge casi seguro a X con $\mathscr{L}(X) = \mu$ y tales que $\mathscr{L}(X_n) = \mu_n$ para todo $n \in \mathbb{N}$, entonces se tiene que*

$$H_n(X_n) \to H(X), \quad \nu\text{-casi seguro}.$$

Por último, en esta misma línea pueden consultarse el Teorema 2.8 y su observación posterior del artículo de Del Barrio y Loubes [12] publicado en 2019.

Capítulo 3

Analogías y peculiaridades

En esta sección vamos a analizar ciertas analogías entre las demostraciones presentadas para los resultados relativos a la Ley de Convergencia de Tipos estudiados en el primer capítulo del trabajo y el problema del transporte óptimo expuesto en el segundo capítulo.

En el Teorema 2.44 ya se intuían las analogías con el problema del transporte óptimo y los emparejamientos al aparecer aplicaciones que conservan la medida y subgradientes de funciones convexas. Para mayor énfasis, observamos que el hecho de que la aplicación s sea una proyección de h en $S(\Omega_1, \Omega_2)$ nos lleva a buscar s como la aplicación que minimiza

$$\int_{\Omega_1} \|h(\omega) - \sigma(\omega)\|^2 \, d\lambda(\omega)$$

entre todas las aplicaciones $\sigma \in S(\Omega_1, \Omega_2)$. Si consideramos la medida $\pi = (h \times \sigma)\#\lambda$ entonces el problema se convierte en

$$\text{mín} \left\{ \int_{\Omega_3 \times \Omega_2} \|x - y\|^2 \, d\pi(x, y) \ : \ \pi = (h \times \sigma)\#\lambda, \ \sigma\#\lambda = \nu \right\}.$$

En términos probabilísticos, π es la distribución conjunta de (h, σ), viéndolo como una pareja de variables aleatorias en Ω_1. Observamos que $\pi \in \Pi(h\#\lambda, \sigma\#\lambda) = \Pi(\mu, \nu)$ donde $\mu = h\#\lambda$.

Si buscamos s tal que $s\#\lambda = \nu$ y $s = \nabla\varphi \circ h$, para alguna función φ convexa, entonces $\pi = (h \times s)\#\lambda$ estaría concentrado en el grafo de la función $\nabla\varphi$ y

por lo visto en 2.31 sería solución del problema más general

$$\text{mín}\left\{\int_{\Omega_3\times\Omega_2}\|x-y\|^2\,d\pi(x,y)\ :\ \pi\in\Pi(\mu,\nu)\right\}$$

que es exactamente el problema de minimización de Kantorovich entre μ y ν.

En el caso más general, la factorización polar introducida por Brenier unifica algunos resultados ya conocidos como pueden ser la factorización polar de matrices o el teorema de descomposición de campos de vectores de Helmholtz.

La prueba para el caso de la descomposición de campos de vectores de Helmholtz puede consultarse en cualquiera de los dos artículos de Brenier [5] o [6], nosotros nos vamos a centrar en el caso de la descomposición polar de matrices que es más conocido y nos va a resultar realmente útil.

Escogiendo $(\Omega_1,\lambda)=(\overline{\Omega},\beta)=(\overline{\Omega},|\cdot|)$ donde Ω es una bola centrada en el origen y h es la aplicación lineal $h(x)=Ax$ para cierta matriz real A de dimensión $d\times d$. En este caso la condición de no degeneración se traduce en que la matriz A sea no singular. Así, en estas condiciones la clásica factorización polar $A=RU$ donde U es una matriz ortogonal y R es una matriz definida positiva y simétrica corresponde a la factorización $h=\nabla\psi\circ s$ donde $\psi(x)=\frac{1}{2}Rx\cdot x$ es convexa y $s(x)=Ux$ conserva la medida en $(\Omega,|\cdot|)$.

3.1. Caso de aplicaciones lineales

La demostración de la Ley de Convergencia de Tipos en $\mathbb{R}$ la desarrollamos vía el Teorema de Representación de Skorohod y, hasta el momento, para el caso multivariado hemos visto los resultados del artículo de Billingsley [3] que recogemos en el primer capítulo del trabajo.

En este capítulo volveremos a abordar el problema de la Ley de Convergencia de Tipos utilizando representaciones de Skorohod, ahora en el caso multivariante. Obtendremos nuevas demostraciones de los Teoremas 1.11, 1.12 y 1.13 en el caso de aplicaciones lineales, teniendo en cuenta la descomposición polar de matrices y el Teorema 2.45. Esto es, queremos tratar el caso multivariado como realizamos el caso unidimensional: vía la Representación de Skorohod, recurriendo además a los recursos que nos proporciona el transporte óptimo.

- Para el Teorema 1.11 partimos de una sucesión de vectores aleatorios $\{X_n\}_{n=1}^{\infty}$ y una sucesión de aplicaciones lineales $\{\alpha_n\}_{n=1}^{\infty}$ tales que $X_n \xrightarrow{d} X$ y $\alpha_n X_n \xrightarrow{d} X$ donde X es un vector aleatorio no degenerado en ningún subespacio. ¿Qué conclusiones podemos sacar?

  De la representación de Skorohod se tiene que existen vectores aleatorios X^* y X_n^* para todo $n \in \mathbb{N}$ de manera que $\mathscr{L}(X_n^*) = \mathscr{L}(X_n)$, $\mathscr{L}(X^*) = \mathscr{L}(X)$ y $X_n^* \xrightarrow{c.s.} X^*$. Así, $\alpha_n X_n^* \overset{d}{=} \alpha_n X_n \xrightarrow{d} X$.

  Ahora bien, como cada aplicación lineal α_n viene definida por una matriz M_n, vamos a considerar su descomposición polar:

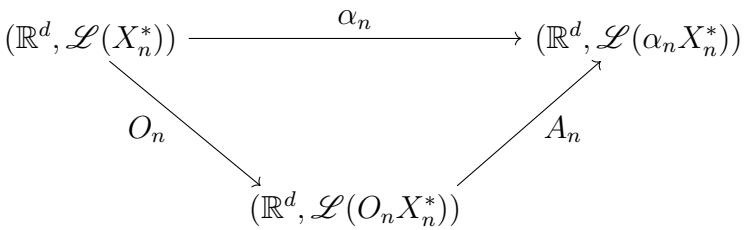
$$M_n = A_n O_n \quad \text{donde} \quad \begin{cases} A_n : \text{simétrica y definida positiva,} \\ O_n : \text{ortogonal.} \end{cases}$$

  Como la sucesión $\{O_n\}_{n=1}^{\infty}$ está contenida en $\mathcal{O}$, el grupo de matrices ortogonales que se trata de un grupo compacto, sabemos que cada subsucesión convergente de la misma lo hace hacia un elemento de $\mathcal{O}$. Consideramos $\{O_{n_k}\}_{k=1}^{\infty}$ con $O_{n_k} \to O_0 \in \mathcal{O}$.

  Así, se sigue que

$$\left.\begin{array}{l} X_{n_k}^* \xrightarrow{c.s.} X^* \\ O_{n_k} \to O_0 \end{array}\right\} \quad \Longrightarrow \quad O_{n_k} X_{n_k}^* \xrightarrow{c.s.} O_0 X^*,$$

  además atendiendo al esquema de Brenier:



  donde A_n es un transporte óptimo entre $\mathscr{L}(O_n X_n^*)$ y $\mathscr{L}(\alpha_n X_n^*)$.

  Ahora bien:

$$\begin{array}{ccc} \mathscr{L}(O_{n_k}X^*_{n_k}) & \xrightarrow{\quad w. \quad} & \mathscr{L}(O_0X^*) \\ \Big\downarrow A_{n_k} & & \Big\downarrow T \\ \mathscr{L}(\alpha_{n_k}X^*_{n_k}) & \xrightarrow{\quad w. \quad} & \mathscr{L}(X^*) \end{array}$$

donde T es un transporte óptimo entre $\mathscr{L}(O_0X^*)$ y $\mathscr{L}(X^*)$.

Como $O_{n_k}X^*_{n_k} \xrightarrow{c.s.} O_0X^*$ del Teorema 2.45 se sigue que

$$A_{n_k}O_{n_k}X^*_{n_k} \xrightarrow{c.s.} T(O_0X^*) \stackrel{d}{=} X^* \quad \Longrightarrow \quad A_{n_k}O_{n_k}X^*_{n_k} \xrightarrow{d} X^*. \quad (3.1)$$

Como $O_{n_k} \to O_0$, $X^*_{n_k} \xrightarrow{c.s.} X^*$ y X^* es no degenerado (3.1) implica que la sucesión $\{A_{n_k}\}_{k=1}^{\infty}$ es de adherencia compacta, luego $A_{n_{k_m}} \to A_0$ siendo A_0 una matriz simétrica y definida positiva.

De esta manera:

$$\left.\begin{array}{l} A_{n_{k_m}}O_{n_{k_m}}X^*_{n_{k_m}} \xrightarrow{d} X^* \\ A_{n_{k_m}}O_{n_{k_m}}X^*_{n_{k_m}} \xrightarrow{d} A_0O_0X^* \end{array}\right\} \quad \Longrightarrow \quad A_0O_0X^* \stackrel{d}{=} X^*$$

y como $\mathscr{L}(X^*) = \mathscr{L}(X)$ se tiene que $A_0O_0 \in \mathscr{G}_X$.

Para conseguir resultados análogos a los obtenidos en el Teorema 1.11 nos gustaría demostrar que $A_0 = I$ y que $O_0 \in \mathscr{G}_X$.

De hecho, si $O_0 \in \mathscr{G}_X$ entonces es fácil ver que $A_0 = I$ y viceversa, como comprobamos a continuación.

- Si $A_0 = I$, entonces $X \stackrel{d}{=} A_0O_0X = O_0X$ y, por tanto, $O_0 \in \mathscr{G}_X$.
- Si $O_0 \in \mathscr{G}_X$, entonces $X \stackrel{d}{=} A_0O_0X \stackrel{d}{=} A_0X$.

  En el caso real, tenemos $X \stackrel{d}{=} aX$ con $a > 0$. Considerando F^{-1} la función cuantil asociada a $\mathscr{L}(X)$ y G^{-1} la función cunatil asociada a $\mathscr{L}(aX)$, sabemos que $G^{-1}(t) = aF^{-1}(t)$ y que existe t_0 tal que $F^{-1}(t_0) \neq 0$ por ser X una variable aleatoria no degenerada. En consecuencia $aF^{-1}(t_0) = F^{-1}(t_0)$, de donde se deduce que $a = 1$.

  En el caso multivariado, como A es una matriz definida positiva y simétrica, sabemos que diagonaliza ortogonalmente. Por lo tanto,

existe una matriz ortogonal P de manera que $P^{-1}A_0P = D$, siendo D la matriz diagonal que contiene los autovalores de A_0 que vamos a denotar por $\lambda_1, \ldots, \lambda_d$ y que sabemos que son estrictamente mayores que cero. De esta manera, observamos que

$$A_0X \stackrel{d}{=} X \iff P^{-1}DPX \stackrel{d}{=} X \iff DPX \stackrel{d}{=} PX$$

y, denotando $Y = PX$, tenemos una variable aleatoria no degenerada tal que $DY \stackrel{d}{=} Y$. Ahora bien

$$DY \stackrel{d}{=} Y \iff (\lambda_1 Y_1, \ldots, \lambda_d Y_d) \stackrel{d}{=} (Y_1, \ldots, Y_d)$$

y, en consecuencia, $\lambda_i Y_i \stackrel{d}{=} Y_i$ para todo $i = 1, \ldots, d$ con $\lambda_i > 0$. Del caso real se deduce que necesariamente $\lambda_i = 1$ para todo $i = 1, \ldots, d$, lo que demuestra que $A_0 = I$.

Esta parece ser la única solución posible, pero por ahora es simplemente una conjetura, que seguiremos analizando. Recordemos que el trabajo de Billingsley recurre al Teorema 1.14 para caracterizar el grupo de transformaciones lineales que dejan invariante la ley de X en términos de un grupo de transformaciones ortogonales, dejando abierta su caracterización. Nuestra forma de proceder utiliza las descomposiciones polares de las transformaciones implicadas dando lugar a una descomposición polar de la transformación límite, en su parte positiva y una transformación ortogonal, conjeturando que es la última la que permite conseguir, en su caso, distintas transformaciones que dejan invariante la distribución límite.

Observación 13. Este resultado sería totalmente análago al Teorema 1.11, pues se escribirían las aplicaciones lineales como producto de unas transformaciones que convergen a la identidad y otras cuyos puntos de acumulación mantienen la distribución del vector aleatorio límite.

Si recordamos en la demostración debida a Billingsley, se escribían las transformaciones afines como producto de unas transformaciones δ_n y γ_n, donde la sucesión $\{\delta_n\}_{n=1}^{\infty}$ convergía a la identidad y la sucesión $\{\gamma_n\}_{n=1}^{\infty}$ verificaba que conservaba la distribución del vector aleatorio límite. Pero, estas transformaciones γ_n se construían eligiendo un punto de acumulación de la sucesión de transformaciones afines adecuado, que dejaba fija la distribución del vector límite.

Sabiendo que, en la notación de Billingsley, las transformaciones δ_n convergen a la identidad, elegir puntos de acumulación de la sucesión $\{\alpha_n\}_{n=1}^{\infty}$ sería equivalente a, en la notación utilizada en este capítulo, considerar puntos de acumulación de la sucesión $\{O_n\}_{n=1}^{\infty}$.

- Para el Teorema 1.13 partimos de una sucesión de vectores aleatorios $\{X_n\}_{n=1}^{\infty}$ y una sucesión de aplicaciones lineales $\{\alpha_n\}_{n=1}^{\infty}$ tales que $X_n \xrightarrow{d} X$ y $\alpha_n X_n \xrightarrow{d} Y$ donde X e Y son vectores aleatorios no degenerados en ningún subespacio. ¿Qué conclusiones podemos sacar?

  De la representación de Skorohod se tiene que existen vectores aleatorios X^* y X_n^* para todo $n \in \mathbb{N}$ de manera que $\mathscr{L}(X_n^*) = \mathscr{L}(X_n)$, $\mathscr{L}(X^*) = \mathscr{L}(X)$ y $X_n^* \xrightarrow{c.s.} X^*$. Así, $\alpha_n X_n^* \overset{d}{=} \alpha_n X_n \xrightarrow{d} Y$.

  Ahora bien, como cada aplicación lineal α_n viene definida por una matriz M_n, vamos a considerar su descomposición polar:

$$M_n = A_n O_n \quad \text{donde} \quad \begin{cases} A_n : \text{simétrica y definida positiva,} \\ O_n : \text{ortogonal.} \end{cases}$$

  Como la sucesión $\{O_n\}_{n=1}^{\infty}$ está contenida en $\mathcal{O}$, que es un grupo compacto, sabemos que cada subsucesión convergente de la misma lo hace hacia un elemento de $\mathcal{O}$. Así, tomamos $\{O_{n_k}\}_{k=1}^{\infty}$ de manera que $O_{n_k} \to O \in \mathcal{O}_0$.

  En consecuencia, como $O_{n_k} \to O_0$, $X_{n_k}^* \xrightarrow{c.s.} X^*$ y $A_{n_k} O_{n_k} X_{n_k}^* \xrightarrow{d} Y$ siendo Y una variable no degenerada, se sigue que la sucesión $\{A_{n_k}\}_{k=1}^{\infty}$ tiene adherencia compacta, luego $A_{n_{k_m}} \to A_0$ donde A_0 es una matriz simétrica y definida positiva.

  De esta manera, definiendo la aplicación lineal α_0 como $\alpha_0 x = A_0 O_0 x$ se tiene que:

$$\left.\begin{array}{l} \alpha_{n_{k_m}} X_{n_{k_m}}^* \xrightarrow{c.s} Y \\ \alpha_{n_{k_m}} X_{n_{k_m}}^* \xrightarrow{d} \alpha_0 X^* \end{array}\right\} \quad \Longrightarrow \quad Y \overset{d}{=} \alpha_0 X^* \overset{d}{=} \alpha_0 X,$$

  en consecuencia, X e Y son vectores aleatorios del mismo tipo. Si atendemos a las hipótesis de partida y a la conclusión deducida, esto es lo que se recoge en el Teorema 1.12.

Una vez obtenido lo anterior, denotando $\beta_n = \alpha_0^{-1}\alpha_n$, podemos escribir nuestro problema como una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ y una sucesión de aplicaciones lineales $\{\beta_n\}_{n=1}^{\infty}$ tales que $X_n \xrightarrow{d} X$ y $\beta_n X_n \xrightarrow{d} X$ siendo X un vector aleatorio no degenerado en ningún subespacio. Y basta aplicar el resultado inmediatamente anterior relativo al Teorema 1.11.

Observación 14. Si escribimos las descomposiciones polares de las aplicaciones β_n como $\beta_n x = B_n Q_n x$, siendo Q_n la matriz ortogonal y B_n la matriz simétrica y definida positiva, si realmente se verificase que $B_n \to I$ y que el conjunto de puntos de acumulación de $\{Q_n\}_{n=1}^{\infty}$ está contenido en $\mathscr{G}_X$, como conjeturamos que debería suceder, tendríamos, de la definición de β_n, que

$$\alpha_n x = \alpha_0 B_n Q_n x$$

donde α_0 es una aplicación lineal tal que $\alpha_0 X \overset{d}{=} Y$. Esto coincidiría con la descomposición que presenta Billingsley volviendo a tener en cuenta que las transformaciones γ_n en su caso dejaban invariante la distribución del vector límite y, en nuestro caso, es el conjunto de puntos de acumulación de $\{Q_n\}_{n=1}^{\infty}$ el que estaría contenido en $\mathscr{G}_X$.

Observación 15. En la demostración anterior, cabe destacar que si se tomase otra subsucesión $\{O_{n_m}\}_{m=1}^{\infty}$ de manera que $O_{n_m} \to O_0^*$, razonando de manera totalmente análoga, como $O_{n_m} \to O_0^*$, $X_{n_m}^* \xrightarrow{c.s.} X^*$ y $A_{n_m} O_{n_m} X_{n_m}^* \xrightarrow{d} Y$ siendo Y una variable no degenerada, se sigue que la sucesión $\{A_{n_m}\}_{k=1}^{\infty}$ tiene adherencia compacta, luego $A_{n_{m_l}} \to A_0^*$ donde A_0^* es una matriz simétrica y definida positiva. En este caso, definiendo la aplicación lineal α_0^* como $\alpha_0^* x = A_0^* O_0^* x$ se deduce que $Y \overset{d}{=} \alpha_0^* X$. Luego necesariamente debe verificarse $\alpha_0 X \overset{d}{=} \alpha_0^* X$.

Nuevamente, considerando $\eta_n = (\alpha_0^*)^{-1}\alpha_n$ y tomando su descomposición polar $\eta_n x = B_n^* Q_n^* x$, donde Q_n^* es la matriz ortogonal y B_n^* es la matriz definida positiva y simétrica, si tuviésemos que $B_n^* \to I$ y que el conjunto de puntos de acumulación de $\{Q_n^*\}_{n=1}^{\infty}$ está contenido en $\mathscr{G}_X$, por definición de η_n se tendría que

$$\alpha_n x = \alpha_0^* B_n^* Q_n^* x.$$

Es decir, se seguiría verificando la descomposición del Teorema 1.13 sólo que sería una descomposición diferente de la anterior. En el Teorema 1.13 no se

asegura la unicidad, simplemente se afirma la existencia de una aplicación lineal α de manera que $X \overset{d}{=} \alpha Y$ y $\alpha_n = \alpha\delta_n\gamma_n$ con $\delta_n \to \delta$ y $\gamma_n \in \mathscr{G}_X$.

3.2. Caso de estructuras de dependencia

Se dice que dos medidas de probabilidad P y Q tienen la **misma estructura de dependencia en alguna base**, si para vectores aleatorios X e Y con $\mathscr{L}(X) = P$ y $\mathscr{L}(Y) = Q$ existe un cambio de base, esto es una matriz ortogonal O, tal que las leyes de OX y OY tienen la misma estructura de dependencia.

Además, también sabemos que si OX y OY tienen la misma estructura de dependencia existe T transporte óptimo entre las leyes $\mathscr{L}(OX)$ y $\mathscr{L}(OY)$ tal que

$$T(x_1, x_2, \ldots, x_d) = (T^1(x_1), T^2(x_2), \ldots, T^d(x_d))$$

donde las aplicaciones T^i son crecientes para $i = 1, 2, \ldots, d$.

Ahora vamos a tratar de sacar conclusiones análogas a los Teoremas 1.11, 1.12 y 1.13 pero, en vez de trabajar con una sucesión de vectores aleatorios y con su sucesión transformada por una sucesión de aplicaciones lineales, vamos a realizar el estudio con dos sucesiones de vectores aleatorios que componente a componente sus leyes tienen la misma estructura de dependencia en alguna base.

- Consideramos $\{X_n\}_{n=1}^{\infty}$ e $\{Y_n\}_{n=1}^{\infty}$ sucesiones de vectores aleatorios tales que $L(X_n)$ y $L(Y_n)$ tienen la misma estructura de dependencia en alguna base para todo $n \in \mathbb{N}$, $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{d} X$ siendo X un vector aleatorio no degenerado. ¿Qué conclusiones podemos sacar?

  En primer lugar, de la representación de Skorohod se tiene que existen vectores aleatorios X^* y X_n^* para todo $n \in \mathbb{N}$ de manera que $\mathscr{L}(X_n^*) = \mathscr{L}(X_n)$, $\mathscr{L}(X^*) = \mathscr{L}(X)$ y $X_n^* \xrightarrow{c.s.} X^*$. Ahora, sea O_n la matriz ortogonal correspondiente al cambio de base de manera que las leyes de $O_nX_n^*$ y O_nY_n tienen la misma estructura de dependencia.

  Como $\mathscr{L}(O_nX_n^*)$ y $\mathscr{L}(O_nY_n)$ tienen la misma estructura de dependencia entonces existe T_n transporte óptimo entre las leyes $\mathscr{L}(O_nX_n^*)$ y $\mathscr{L}(O_nY_n)$ tal que

  $$T_n(x_1, x_2, \ldots, x_d) = (T_n^1(x_1), T_n^2(x_2), \ldots, T_n^d(x_d))$$

donde las aplicaciones T_n^i son crecientes para $i = 1, 2, \ldots, d$.

Ahora bien, la sucesión $\{O_n\}_{n=1}^{\infty}$ está formada por matrices ortogonales que pertenecen a $\mathcal{O}$, grupo compacto, luego cada subsucesión convergente de la misma lo hace hacia un elemento de $\mathcal{O}$. Tomamos $\{O_{n_k}\}_{k=1}^{\infty}$ de manera que $O_{n_k} \to O \in \mathcal{O}$, así

$$\begin{array}{ccc} \mathscr{L}(O_{n_k}X_{n_k}^*) & \xrightarrow{\qquad w. \qquad} & \mathscr{L}(OX^*) \\ \Big\downarrow T_{n_k} & & \Big\downarrow T \\ \mathscr{L}(O_{n_k}Y_{n_k}) & \xrightarrow{\qquad w. \qquad} & \mathscr{L}(OX) \end{array}$$

siendo T un transporte óptimo entre $\mathscr{L}(OX^*)$ y $\mathscr{L}(OX)$.

Como $O_{n_k} \to O$ y $X_{n_k}^* \xrightarrow{c.s.} OX^*$ se tiene que $O_{n_k}X_{n_k}^* \xrightarrow{c.s.} OX^*$ y en virtud del Teorema 2.45 se deduce que

$$T_{n_k}(O_{n_k}X_{n_k}^*) \xrightarrow{c.s.} T(OX^*) \stackrel{d}{=} OX \stackrel{d}{=} OX^*.$$

Denotando $Y^* = OX^*$ tenemos $T(Y^*) \stackrel{d}{=} Y^*$, luego de la unicidad casi seguro del transporte óptimo se tiene que $T = Id$, $\mathscr{L}(Y^*)$-c.s.

Tomando otra subsucesión $\{O_{n_m}\}_{m=1}^{\infty}$ de manera que $O_{n_m} \to O_1 \neq O$, realizando un razonamiento análogo al anterior se sigue que el nuevo transporte óptimo límite T_1 también coincide con la identidad casi seguro . Por lo tanto, $T_n \xrightarrow{c.s.} I$.

Para poder desarrollar la siguiente parte, cuando las dos sucesiones con las que tratamos presentan límites distintos, vamos a necesitar dos resultados que se recogen en el artículo [9] de Cuesta-Albertos et al y que vamos a comentar a continuación.

La formulación general para un transporte óptimo H entre probabilidades P y Q con la misma estructura de dependencia en la base $\{e_1, \ldots, e_d\}$ es

$$H(x) = H\left(\sum_{i=1}^{d} x_i e_i\right) = \sum_{i=1}^{d} f_i(x_i) e_i,$$

para ciertas funciones crecientes $\{f_1 \ldots, f_d\}$. En el artículo mencionado se prueba que esta representación que depende de la base es esencialmente única

y que los transportes óptimos correspondientes están intrínsecamente relacionados con descomposiciones ortogonales del espacio. En lo que sigue vamos a enunciar y comentar estos resultados sin incluir las demostraciones de los mismos que pueden consultarse en el apéndice del propio artículo [9].

En primer lugar, recordamos que si H es un transporte óptimo, entonces $H + a$ también es un transporte óptimo para todo $a \in \mathbb{R}^d$. Luego se puede suponer, sin pérdida de generalidad que las funciones f_i verifican $f_i(0) = 0$ para todo $i = 1, \ldots, d$. Más concretamente, consideremos $H : \mathbb{R}^d \to \mathbb{R}^d$ una aplicación que puede ser expresada como

$$H\left(\sum_{i=1}^{d} x_i e_i\right) = \sum_{i=1}^{d} f_i(x_i) e_i$$

y

$$H\left(\sum_{i=1}^{d} \hat{x}_i \hat{e}_i\right) = \sum_{i=1}^{d} g_i(\hat{x}_i) \hat{e}_i$$

donde $\{e_1, \ldots, e_d\}$ y $\{\hat{e}_1, \ldots, \hat{e}_d\}$ son bases ortonormales y $f_1, \ldots, f_d$ son funciones reales y crecientes verificando $f_i(0) = 0$. Además, para evitar trivialidades vamos a suponer que que no existen dos pares de vectores e_i, $\hat{e}_j$ linealmente dependientes, esto es $e_i \neq \pm \hat{e}_j$ para todo i, j.

Proposición 3.1. *En las condiciones anteriores, se tiene que H es una aplicación lineal que puede ser escrita como $H\left(\sum_{i=1}^{d} x_i e_i\right) = \sum_{i=1}^{d} (\beta_i x_i) e_i$, donde los autovalores β_i son positivos con multiplicidades $m_i \geq 2$, para todo $i = 1, \ldots, d$.*

Esta proposición viene a decir que todo transporte óptimo entre medidas de probabilidad que tienen misma estructura de dependencia en alguna base determina de forma única una descomposición ortogonal del espacio, hecho que se recoge en el siguiente corolario.

Corolario 3.2. *Existe una única descomposición de $\mathbb{R}^d$ como suma directa de subespacios ortogonales, $\mathbb{R}^d = V_1 \oplus \ldots \oplus V_k$ tales que:*

(*i*) *Cada vector en un subespacio V_i genera una dirección invariante para H.*

(*ii*) *Si $dim(V_i) > 1$ para algún i, existe $\beta \geq 0$ tal que $H(v) = \beta v$ para todo $v \in V_i$.*

(*iii*) *Si* $i \neq j$ *y* $H|_{V_i}$ *y* $H|_{V_j}$ *son lineales, entonces los autovalores asociados son diferentes.*

(*iv*) *Si* $v = v_1 + \ldots + v_k$ *con* $v_i \in V_i$, *entonces* $H(v) = H(v_1) + \ldots + H(v_k)$.

Observación 16. El hecho de que se consideren trivialidades los casos en los que existen índices $i, j \in \{1, \ldots, d\}$ de manera que $e_i = \pm\hat{e}_j$, se debe a que para $x \in \mathbb{R}$ se tiene:

- si $e_i = \hat{e}_j$, entonces
$$\begin{aligned} T(xe_i) = T(x\hat{e}_j) &\iff f_i(x)e_i = g_j(x)\hat{e}_j = g_j(x) \\ &\iff f_i(x) = g_j(x); \end{aligned}$$

- si $e_i = -\hat{e}_j$, entonces
$$\begin{aligned} T(xe_i) = T(-x\hat{e}_j) &\iff f_i(x)e_i = g_j(-x)\hat{e}_j = -g_j(-x)e_i \iff \\ &\iff f_i(x) = -g_j(-x), \end{aligned}$$

Esto es, en ambos casos, se deduce una relación entre las aplicaciones f_i y g_j, que va a ser determinante, en lo que sigue, para comprobar que todo funciona bien componente a componente.

Con los resultados expuestos pasamos a analizar el caso en el que dos sucesiones de vectores aleatorios que presentan la misma estructura de dependencia en alguna base tienen límites distintos.

- Si ahora consideramos sucesiones $\{X_n\}_{n=1}^{\infty}$ e $\{Y_n\}_{n=1}^{\infty}$ de vectores aleatorios tales que $\mathscr{L}(X_n)$ y $\mathscr{L}(Y_n)$ tienen la misma estructura de dependencia en alguna base para todo $n \in \mathbb{N}$, y vectores aleatorios X e Y verificando $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{d} Y$ siendo Y no degenerado en ningún subespacio. ¿Qué conclusiones podemos sacar?

  Nuevamente, de la representación de Skorohod se tiene que existen vectores aleatorios X^* y X_n^* para todo $n \in \mathbb{N}$ de manera que $\mathscr{L}(X_n^*) = \mathscr{L}(X_n)$, $\mathscr{L}(X^*) = \mathscr{L}(X)$ y $X_n^* \xrightarrow{c.s.} X^*$. Ahora, sea O_n la matriz ortogonal correspondiente al cambio de base de manera que $O_nX_n^*$ y O_nY_n tienen la misma estructura de dependencia.

  Como $O_nX_n^*$ y O_nY_n tienen la misma estructura de dependencia entonces existe $T_n : \mathscr{L}(O_nX_n^*) \to \mathscr{L}(O_nY_n)$ transporte óptimo tal que
$$T_n(x_1, x_2, \ldots, x_d) = (T_n^1(x_1), T_n^2(x_2), \ldots, T_n^d(x_d))$$

donde las aplicaciones T_n^i son crecientes para $i = 1, 2, \ldots, d$.

Ahora bien, como la sucesión $\{O_n\}_{n=1}^{\infty}$ está formada por elmentos de $\mathcal{O}$, grupo compacto, toda subsucesión convergente de la misma lo hace hacia un elemento de $\mathcal{O}$. Tomamos una subsucesión $\{O_{n_k}\}_{k=1}^{\infty}$ de manera que $O_{n_k} \to O \in \mathcal{O}$, así

$$\begin{array}{ccc} \mathscr{L}(O_{n_k}X^*_{n_k}) & \xrightarrow{\qquad w. \qquad} & \mathscr{L}(OX^*) \\ \Big\downarrow T_{n_k} & & \Big\downarrow T \\ \mathscr{L}(O_{n_k}Y^*_{n_k}) & \xrightarrow{\qquad w. \qquad} & \mathscr{L}(OY) \end{array}$$

siendo T un transporte óptimo entre $\mathscr{L}(OX^*)$ y $\mathscr{L}(OY)$.

Como $O_{n_k} \to O$ y $X^*_{n_k} \xrightarrow{c.s.} OX^*$ se tiene que $O_{n_k}X^*_{n_k} \xrightarrow{c.s.} OX^*$ y, en virtud del Teorema 2.45, se deduce que

$$T_{n_k}(O_{n_k}X^*_{n_k}) \xrightarrow{c.s.} T(OX^*) \stackrel{d}{=} OY.$$

Si las aplicaciones T_{n_k} convergen entonces sus componentes $T^i_{n_k}$ también lo hacen, esto es

$$\begin{array}{ccccc} T_{n_k}(x_1, \ldots, x_d) & = & \big(T^1_{n_k}(x_1), \ldots, T^d_{n_k}(x_d) \big) & & \\ \Big\downarrow k \to \infty & & \Big\downarrow k \to \infty & & \Big\downarrow k \to \infty \\ T(x_1, \ldots, x_d) & = & \big(T^1(x_1), \ldots, T^d(x_d) \big) & . & \end{array}$$

Luego $\mathscr{L}(OX^*)$ y $\mathscr{L}(OY)$ tienen la misma estructura de dependencia y, en consecuencia, $\mathscr{L}(X)$ y $\mathscr{L}(Y)$ tienen la misma estructura de dependencia en alguna base. Esta primera parte sería el equivalente al Teorema 1.12 pero en el caso que estamos tratando de estructuras de dependencia.

Ahora bien, ¿qué ocurre si tomamos dos subsucesiones diferentes de $\{O_n\}_{n=1}^{\infty}$ con límites distintos?

Supongamos que $O_{n_k} \to O_1 \in \mathcal{O}$ y $O_{n_m} \to O_2 \in \mathcal{O}$ con $O_1 \neq O_2$. A priori, con el desarrollo anterior llegaríamos a dos transportes óptimos

T_1 y T_2 entre las leyes $\mathscr{L}(O_1X^*)$ y $\mathscr{L}(O_1Y)$, y $\mathscr{L}(O_2X^*)$ y $\mathscr{L}(O_2Y)$ respectivamente. Sin embargo, las leyes iniciales no varían ($\mathscr{L}(X)$ y $\mathscr{L}(Y)$) entonces tenemos un transporte óptimo T de manera que:

→ Como $\mathscr{L}(X^*)$ y $\mathscr{L}(Y)$ tienen la misma estructura de dependencia en la base $B = \{e_1, \ldots, e_d\}$ (correspondiente a la matriz O_1) entonces

$$T\left(\sum_{i=1}^{d} x_i e_i\right) = \sum_{i=1}^{d} f_i(x_i) e_i$$

donde las funciones f_i son crecientes para todo $i = 1, \ldots, d$.

→ Como $\mathscr{L}(X^*)$ y $\mathscr{L}(Y)$ tienen la misma estructura de dependencia en la base $\hat{B} = \{\hat{e}_1, \ldots, \hat{e}_d\}$ (correspondiente a la matriz O_2) entonces

$$T\left(\sum_{i=1}^{d} \hat{x}_i \hat{e}_i\right) = \sum_{i=1}^{d} g_i(\hat{x}_i) \hat{e}_i$$

donde las funciones g_i son crecientes para todo $i = 1, \ldots, d$.

Observamos que

$$\sum_{i=1}^{d} f_i(0) e_i = T\left(\sum_{i=1}^{d} 0 e_i\right) = T(\mathbf{0}) = T\left(\sum_{i=1}^{d} 0 \hat{e}_i\right) = \sum_{i=1}^{d} g_i(0) \hat{e}_i,$$

luego

$$\tilde{T}\left(\sum_{i=1}^{d} x_i e_i\right) = \sum_{i=1}^{d} (f_i(x_i) - f_i(0)) e_i = \sum_{i=1}^{d} \tilde{f}_i(x_i) e_i$$

y

$$\tilde{T}\left(\sum_{i=1}^{d} \hat{x}_i \hat{e}_i\right) = \sum_{i=1}^{d} (g_i(\hat{x}_i) - g_i(0)) \hat{e}_i = \sum_{i=1}^{d} \tilde{g}_i(\hat{x}_i) \hat{e}_i$$

son el mismo transporte óptimo expresado en bases distintas, pero verificando que $\tilde{f}_i(0) = 0$ para todo índice i (y, de hecho, también se tiene que $\tilde{g}_i(0) = 0$).

En estas condiciones, considerando los conjuntos

$$I = \{i \in \{1, \ldots, d\} \,:\, e_i = \pm \hat{e}_j, \text{ para algún } j \in \{1, \ldots, d\}\}$$

y

$$J = \{j \in \{1, \dots, d\} \, : \, \hat{e}_j = \pm e_i, \text{ para algún } i \in \{1, \dots, d\}\},$$

la Proposición 3.1 nos asegura que el transporte óptimo $\tilde{T}$ se puede escribir como

$$\tilde{T}\left(\sum_{k=1}^{d} x_k e_k\right) = \sum_{k \notin I} \beta_k x_k e_k + \sum_{k \in I} \tilde{f}_k(x_k) e_k,$$

donde los autovalores β_k son positivos y tienen multiplicidades $m_k > 1$.

En consecuencia, el transporte óptimo T viene dado por

$$\begin{aligned} T\left(\sum_{i=1}^{d} x_i e_i\right) &= \sum_{i \in I} \tilde{f}_i(x_i) e_i + \sum_{i \notin I} \beta_i x_i e_i + \sum_{i=1}^{d} f_i(0) e_i \\ &= \sum_{i \in I} f_i(x_i) e_i + \sum_{i \notin I} (\beta_i x_i + f_i(0)) e_i. \end{aligned}$$

Es decir, el transporte óptimo T es, salvo para los vectores indexados en I, una aplicación afín componente a componente.

Observación 17. Cabe destacar que, como el transporte óptimo $\tilde{T}$ expresado en la base $\hat{B} = \{\hat{e}_1, \dots, \hat{e}_d\}$ también verifica que las aplicaciones crecientes $\tilde{g}_i$ cumplen $\tilde{g}_i(0) = 0$ para todo índice i, podemos aplicar el resultado de la Proposición 3.1 a esta representación de $\tilde{T}$ obteniendo que

$$\tilde{T}\left(\sum_{k=1}^{d} \hat{x}_k \hat{e}_k\right) = \sum_{k \notin I} \hat{\beta}_k \hat{x}_k \hat{e}_k + \sum_{k \in I} \tilde{g}_k(x_k) \hat{e}_k,$$

donde los autovalores $\hat{\beta}_k$ son positivos y tienen multiplicidades $\hat{m}_k > 1$.

El Corolario 3.2 nos asegura que existe una única descomposición de $\mathbb{R}^d$ como suma directa de subespacios ortogonales verificando las condiciones (i), (ii), (iii) y (iv), pero en este caso, aplicado al transporte óptimo T.

A continuación, vamos a considerar las siguientes descomposiciones de $\mathbb{R}^d$ como suma directa de subespacios ortogonales:

(1) Para $i \in I$ nos quedamos con los subespacios $< e_i >$. Por otro lado, los vectores no indexados por el conjunto I los vamos a agrupar en subespacios $V_1, \ldots, V_k$ de manera que V_i está formado por los vectores $e^i_1, \ldots, e^i_{n_i}$ que tienen mismo autovalor asociado. Entonces, tenemos la descomposición ortogonal

$$\mathbb{R}^d = \left(\bigoplus_{i \in I} < e_i > \right) \oplus V_1 \oplus \ldots \oplus V_k$$

que trivialmente verifica las condiciones (i), (ii), (iii) y (iv).

(2) Descomposición análoga para la base $\{\hat{e}_1, \ldots, \hat{e}_d\}$, entonces

$$\mathbb{R}^d = \left(\bigoplus_{j \in J} < \hat{e}_j > \right) \oplus W_1 \oplus \ldots \oplus W_m.$$

En primer lugar, ya sabemos que si $i \in I$ existe $j \in J$ de manera que $e_i = \pm \hat{e}_j$ y, en consecuencia, $< e_i >=< \hat{e}_j >$. Luego los subespacios de dimensión 1 no presentan ningún problema.

De la unicidad de la descomposición de $\mathbb{R}^d$ asegurada por el Corolario 3.2 se deduce que, necesariamente, $k = m$ y que para todo $i \in \{1, \ldots, k\}$ existe $j \in \{1, \ldots, m = k\}$ de manera que $V_i = W_j$ de lo que se deduciría $\beta_i = \hat{\beta}_j$.

En esencia, lo anterior viene a decir que no importa la base elegida dentro de aquellas que verifican que las leyes de los vectores X e Y presenten la misma estructura de dependencia en la misma. Para ilustrarlo vamos a reordenar las bases $B = \{e_1, \ldots, e_d\}$ y $\hat{B} = \{\hat{e}_1 \ldots \hat{e}_d\}$ de la manera siguiente:

- En primer lugar vamos a colocar los vectores correspondientes a subespacios de dimensión 1, de manera que $e_1 = \pm \hat{e}_1$, $e_2 = \pm \hat{e}_2$, …, $e_{|I|} = \pm \hat{e}_{|I|}$.
- A continuación, en la reordenación de la base B, colocamos los vectores $e^1_1, \ldots, e^1_{n_1}$ que generan el subespacio V_1. En la reordenación de la base $\hat{B}$ colocaríamos los vectores $\hat{e}^j_1, \ldots, \hat{e}^j_{m_j}$ siendo j el índice que verifica $V_1 = W_j$ (luego necesariamente $n_1 = m_j$).
- Después, en la reordenación de la base B, colocamos los vectores $e^2_1, \ldots,$ $e^2_{n_2}$ que generan el subespacio V_2. En la reordenación de la base $\hat{B}$

colocaríamos los vectores $\hat{e}^j_1, \ldots, \hat{e}^j_{m_j}$ siendo j el índice que verifica $V_2 = W_j$ (luego necesariamente $n_2 = m_j$).

- E iteramos este mismo proceso hasta incluir los vectores que generan el subespacio V_k a la reordenación de la base B.

Así, escribiendo $|I| = h$ y $l = h + \sum_{i=1}^{n-1} n_i = h + \sum_{i=1}^{n-1} dim(V_i)$, lo que tenemos es:

- En la reordenación de la base B, denotando $\boldsymbol{x} = (x_1, \ldots, x_d)_B$ se tiene:

$$T_1(\boldsymbol{x}) = (f_1(x_1), \ldots, f_h(x_h), \beta_1 x_{h+1} - f_{h+1}(0), \ldots, \beta_1 x_{h+n_1} - f_{h+n_1}(0), \\ \ldots, \beta_k x_{l+1} - f_{l+1}(0), \ldots, \beta_k x_d - f_d(0))_B.$$

- En la reordenación de la base $\hat{B}$, denotando $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_d)_{\hat{B}}$ se tiene:

$$T_2(\hat{\boldsymbol{x}}) = (g_1(x_1), \ldots, g_h(x_h), \beta_1 \hat{x}_{h+1} - g_{h+1}(0), \ldots, \beta_1 \hat{x}_{h+n_1} - g_{h+n_1}(0), \\ \ldots, \beta_k \hat{x}_{l+1} - g_{l+1}(0), \ldots, \beta_k \hat{x}_d - g_d(0))_{\hat{B}}.$$

Esto da sentido a decir que nos da igual trabajar con T_1 (es decir, T expresada en la base B) o T_2 (T expresada en la base $\hat{B}$), pues dado $x \in \mathbb{R}^d$ podemos escribir $x = x_1 e_1 + \ldots + x_d e_d$ y $x = \hat{x}_1 \hat{e}_1 + \ldots + \hat{x}_d \hat{e}_d$, y observamos teniendo en cuenta la unicidad de la descomposición de $\mathbb{R}^d$ como suma directa de subespacios ortogonales:

- Si $e_i = \hat{e}_j$ entonces $x_i = \hat{x}_i$ y en consecuencia

$$f_i(x_i) e_i = g_i(x_i) \hat{e}_i,$$

  o equivalentemente,

$$(0, \ldots, f_i(x_i), \ldots 0)_B = (0, \ldots, g_i(\hat{x}_i), \ldots, 0)_{\hat{B}}\,.$$

- Si $e_i = -\hat{e}_j$ entonces $x_i = -\hat{x}_i$ y en consecuencia

$$f_i(x_i) e_i = -f_i(-\hat{x}_i) \hat{e}_i = g_i(\hat{x}_i) \hat{e}_i,$$

  o equivalentemente,

$$(0, \ldots, f_i(x_i), \ldots 0)_B = (0, \ldots, g_i(\hat{x}_i), \ldots, 0)_{\hat{B}}\,.$$

- Para los coeficientes de las combinaciones afines correspondientes a un subespacio V_i se tiene que $x_1^i e_1^i + \ldots + x_{n_i}^i e_{n_i}^i = \hat{x}_1^i \hat{e}_1^i + \ldots + \hat{x}_{n_i}^i \hat{e}_{n_i}^i$ y al tratarse las restricciones de T_1 y T_2 respecto de V_i de aplicaciones lineales con el mismo autovalor β_i asociado, se tiene que

$$T_1\left(x_1^i e_1^i + \ldots + x_{n_i}^i e_{n_i}^i\right) = T_2\left(\hat{x}_1^i \hat{e}_1^i + \ldots + \hat{x}_{n_i}^i \hat{e}_{n_i}^i\right),$$

o equivalentemente,

$$\begin{gathered}\left(0, \ldots, \beta_i x_1^i - c_1^i, \ldots, \beta_i x_{n_i}^i - c_{n_i}^i, \ldots, 0\right)_B \\ \| \\ \left(0, \ldots, \beta_i \hat{x}_1^i - d_1^i, \ldots, \beta_i \hat{x}_{n_i}^i - d_{n_i}^i, \ldots, 0\right)_{\hat{B}},\end{gathered}$$

siendo las constantes c_j^i y d_j^i los valores adecuados de los vectores $(f_1(0), \ldots, f_d(0))_B$ y $(g_1(0), \ldots, g_d(0))_{\hat{B}}$.

En conclusión, el transporte óptimo T expresado en las bases B y $\tilde{B}$ consiste, salvo en las componentes correspondientes a subespacios unidimensionales, en aplicaciones afines en las que la parte lineal es común en cada subespacio de dimensión mayor que uno y lo que varía es la constante que se resta en cada componente.

Bibliografía

[1] Ash, R. B. *Real analysis and probability.* Academic Press, 1972.

[2] Bickel, P. J., and Freedman, D. A. Some asymptotic theory for the bootstrap. *The Annals of Statistics 9*, 6 (1981), 1196–1217.

[3] Billingsley, P. Convergence of Types in k-Space. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 5* (1966), 175–179.

[4] Billingsley, P. *Probability and measure.* John Wiley & Sons, 2017.

[5] Brenier, Y. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math. 305* (1987), 805–808.

[6] Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics 44*, 4 (1991), 375–417.

[7] Cambanis, S., Simons, G., and Stout, W. Inequalities for $\mathrm{E}k(X, Y)$ when the marginals are fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 36*, 4 (1976), 285–294.

[8] Cuesta, J. A., and Matran, C. A short proof of the law of convergence of types. *Sankhyā: The Indian Journal of Statistics, Series A* (1990), 259–260.

[9] Cuesta-Albertos, J., Matrán, C., and Tuero-Diaz, A. Optimal maps for the L^2-Wasserstein distance. *Preprint* (1996).

[10] Cuesta-Albertos, J., Matrán, C., and Tuero-Diaz, A. Optimal transportation plans and convergence in distribution. *Journal of Multivariate Analysis 60*, 1 (1997), 72–83.

[11] Cuesta-Albertos, J., Ruschendorf, L., and Tuero-Díaz, A. Optimal coupling of multivariate distributions and stochastic processes. *Journal of Multivariate Analysis 46*, 2 (1993), 335–361.

[12] del Barrio, E., and Loubes, J.-M. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability 47*, 2 (2019), 926 – 951.

[13] Kellerer, H. G. Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 67* (1984), 399–432.

[14] Loève, M. *Probability Theory I*. Springer, 1977.

[15] Major, P. On the invariance principle for sums of independent identically distributed random variables. *Journal of Multivariate Analysis 8*, 4 (1978), 487–517.

[16] Rüschendorf, L. On the distributional transform, sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference 139*, 11 (2009), 3921–3927.

[17] Rüschendorf, L., and Rachev, S. T. A characterization of random variables with minimum L^2-distance. *Journal of Multivariate Analysis 32*, 1 (1990), 48–54.

[18] Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY 55*, 58-63 (2015), 94.

[19] Villani, C. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

[20] Villani, C. *Optimal transport: old and new*. Springer, 2009.